Citation for published version (APA):
Cao, G., Jiang, J., Bollegala, D., Li, M., & Luo, S. (in press). Multimodal Zero-Shot Learning for Tactile Texture Recognition. *Robotics and Autonomous Systems*.

Multimodal Zero-Shot Learning for Tactile Texture Recognition

Guanqun Cao[a], Jiaqi Jiang[b], Danushka Bollegala[a], Min Li[c] and Shan Luo[b,*]

[a] *Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, United Kingdom*
[b] *Department of Engineering, King's College London, London, WC2R 2LS, United Kingdom*
[c] *School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, 710049, China*

ARTICLE INFO

Keywords:
Tactile sensing
Zero-Shot Learning
Material recognition
Multi-modal perception
Textile sorting
Robot perception

ABSTRACT

Tactile sensing plays an irreplaceable role in robotic material recognition. It enables robots to distinguish material properties such as their local geometry and textures, especially for materials like textiles. However, most tactile recognition methods can only classify known materials that have been touched and trained with tactile data, yet cannot classify unknown materials that are not trained with tactile data. To solve this problem, we propose a tactile Zero-Shot Learning framework to recognise materials when they are touched for the first time, using their visual and semantic information, without requiring tactile training samples. The biggest challenge in tactile Zero-Shot Learning is to recognise disjoint classes between training and test materials, i.e., the test materials that are not among the training ones. To bridge this gap, the visual modality, providing tactile cues from sight, and semantic attributes, giving high-level characteristics, are combined together and act as a link to expose the model to these disjoint classes. Specifically, a generative model is learnt to synthesise tactile features according to corresponding visual images and semantic embeddings, and then a classifier can be trained using the synthesised tactile features for zero-shot recognition. Extensive experiments demonstrate that our proposed multimodal generative model can achieve a high recognition accuracy of 83.06% in classifying materials that were not touched before. The robotic experiment demo and the FabricVST dataset are available at *https://sites.google.com/view/multimodalzsl*

1. Introduction

The material properties of the object's surface, such as roughness, texture, and hardness, are key information for robots to interact with their surroundings. As such, recognising the surface materials, which allows robots to be aware of the object categories and properties, is fundamental to many manipulation tasks, such as grasping [1], labware handling [2], and material sorting for recycling [3].

The most widely used methods for material recognition are based on vision as it provides shapes, colours, and appearances to perceive properties of different materials [4]. Vision-based methods, however, are subject to lighting and occlusion [5, 6]. Moreover, due to the enormous range of appearances that a single material might exhibit, it is difficult to establish distinguishable representations from vision alone. To address this problem, the semantic attributes, e.g., "knitted" and "fibrous" to describe the wool, have been introduced as complimentary information to assist visual material recognition [7, 8].

Unlike vision and semantic attributes, tactile sensing can measure the micro-structures of the object's surface even if the appearance and shape are changed, which allows the robot to recognise different materials effectively [9, 10]. Furthermore, many exclusive physical properties that cannot be obtained in other sensory modalities, such as friction and compressibility, can be measured by tactile sensing through rich physical interaction, giving a good understanding of different materials.

In recent studies on material recognition by tactile sensing, a large amount of tactile data need to be collected first, and then a projection function between the collected tactile data and material classes is learnt with optimisation algorithms or machine learning methods [11, 12]. By using the learnt projection function, robots can predict the classes of the contacted material. However, there are two main issues limiting the application of such methods: 1) The material to be recognised must be known and included in the classes of the training dataset, which is hard to be met due to the continuous development of new materials. 2) The collection of a large amount of tactile data is costly as the delicate tactile sensors are easily damaged after numerous physical contacts and data collection could be time-consuming.

As a result, the lack of tactile samples for training and the absence of annotations pose challenges to recognise the materials never touched by robots before. Hence, Zero-Shot Learning (ZSL) for tactile recognition is desired. It aims to identify the unknown (untouched) materials using tactile sensing upon the first contact, for which there are no training tactile samples, by applying the knowledge learnt from tactile data of known (touched) materials. This capability can be acquired by obtaining a shared subspace (e.g., with the visual domain) to transfer the knowledge learnt from touched materials to untouched materials.

A good example of tactile ZSL is the recognition of daily fabrics, which provide a variety of appearances, physical properties, and tactile feelings. For humans, it is an easy task to recognise a new material by the sense of touch based on our prior knowledge and descriptions. For example, if we are given a description of silk "*the silk material is usually very smooth, soft and cool*", we can recognise a fabric made of silk even when it is the first time for us to touch a piece

[*] This work was funded in part by the EPSRC project "ViTac: Visual-Tactile Synergy for Handling Flexible Materials" (EP/T033517/2).

[*] Corresponding author
✉ shan.luo@kcl.ac.uk (S. Luo)
ORCID(s): 0000-0002-3462-2567 (G. Cao)

of silk. This capability to recognise new materials that were never touched before is due to that our brain is able to transfer the knowledge gained in one sensory modality to another, i.e., cross-modal transfer [13]. Similarly, humans can also *imagine* a tactile feeling when *observing* the materials [14].

Inspired by the above, we propose a multimodal Zero-Shot Learning approach for tactile textures recognition that employs both visual information and semantic attributes from fabrics to synthesise corresponding tactile features. As shown in Fig. 1, firstly, a generative model is trained to synthesise tactile features with touched materials. After training, the generative model is used to synthesise tactile features of untouched materials using corresponding visual information and semantic attributes. Then, a classifier is trained on the synthesised tactile features of untouched materials. Finally, the robot can recognise these untouched materials by tactile sensing. Our proposed tactile ZSL method addresses a practical problem in material recognition using robots: even if tactile data is not available to train the robot for new materials, they can still be recognised and sorted when being touched for the first time.

The contributions of this paper can be summarised as follows:

1. We propose a multimodal ZSL framework to recognise materials that have not been touched before;
2. We develop a generative model to synthesise tactile features from visual images and semantic attributes to achieve a synergistic effect in tactile ZSL, which is the first of its kind;
3. We collect a new dataset, named as *FabricVST*, from 50 pieces of fabrics to train the model, including visual images, semantic attributes and tactile data, and validate our proposed method on the untouched materials.

The remainder of this paper is structured as follows: Section 2 reviews the related works; Section 3 introduces the problem formulation of tactile ZSL; Section 4 details the multimodal framework for ZSL of tactile recognition; Section 5 introduces the experimental setup; Section 6 shows the experimental results; Section 7 discusses several aspects of tactile ZSL; Finally, Section 8 summarises the paper and gives conclusions.

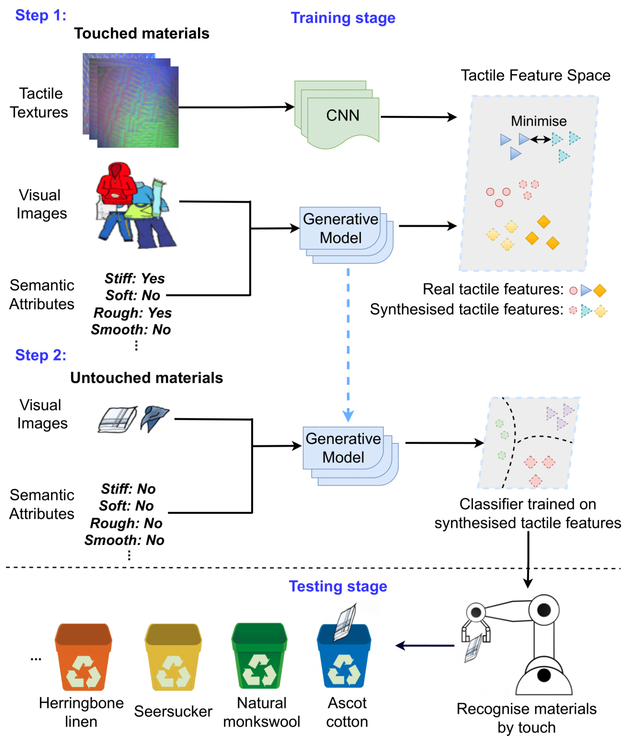


Figure 1: Tactile Zero-Shot Learning for material recognition. Training stage: Step 1: Given visual images, semantic attributes and tactile textures of known (touched) materials, a generative model is trained to synthesise tactile features by minimising the distance between the distributions of real tactile features and synthesised tactile features. Step 2: A classifier is trained by the synthesised tactile features of unknown (untouched) materials. Testing stage: By using the classifier, the robot is able to recognise and sort unknown materials by tactile sensing.

2. Related Works

In this section, we will first review works on material recognition with tactile textures, followed by discussions of Zero-Shot Learning for visual recognition and tactile recognition, respectively.

2.1. Material Recognition with Tactile Textures

Tactile textures are crucial in understanding the properties of materials since they convey important information of local geometry and micro-structures of the object's surface. With the development of tactile sensing, various tactile sensors based on different sensing technologies, such as microphones [15, 16], strain gauges [17], MEMS [18, 19], capacitive [20, 21], and piezoresistive [3], have been implemented in the tactile texture recognition. Recently, thanks to their high resolution and low cost, camera-based optical tactile sensors, such as the GelSight sensor [22] and the TacTip sensor [23], have been implemented in the material perception. In [24], the GelSight is applied in an autonomous tactile exploration that enables the robots to perceive material properties. In [10], a spatio-temporal attention mechanism is proposed to emphasise the salient features to recognise textures present in the tactile image sequences collected by a GelSight sensor. In [25], a joint latent space of vision and touch sensing for sharing features is learnt to improve the cloth material recognition. However, these methods are limited to classifying known materials and cannot recognise unknown materials that are not included in the training classes. In [26], a generative model is proposed to generate textures to be rendered on a haptic display from visual images, which can be touched by human participants. In this work, we also utilise a generative model to synthesise tactile features from other modalities including visual images. However, in contrast to [26] for haptic rendering to human users, our method is designed for a robot to identify novel materials through tactile sensing.

2.2. Zero-Shot Learning for Visual Recognition

An increasing interest has been shown in recognising unseen objects based on visual images without any training examples, i.e., visual Zero-Shot Learning. The use of semantic embeddings learnt from semantic attributes is common to bridge the gap between the seen classes and the unseen classes. A projection function can be learnt, for instance, from visual to semantic space [27, 28], from semantic to visual space [29, 30], or a shared latent space [31, 32], to connect vision and semantic information for the recognition. However, as the classes of the seen data and unseen data can be unrelated, the data distributions may be different. If the projection function that is learnt from seen data is applied to the unseen data directly, it may generate unknown bias, known as the domain shift problem [33]. Another popular method is based on the generative model [34, 35], where the visual features of unseen classes are synthesised using semantic information, and synthesised visual features are used to train a classifier to recognise unseen data, which alleviates the domain shift problem significantly. However, these studies focus on the visual ZSL problem and there have been no works on tactile ZSL based on generative models that use visual images and semantic attributes together to synthesise tactile features.

2.3. Zero-Shot Learning for Tactile Recognition

Due to the difficulty of the tactile data collection and annotation, a great demand exists for tactile ZSL. However, compared with the visual ZSL, the ZSL problem for tactile recognition has been much less investigated. In [36], visual images are used as the auxiliary information to connect the touched objects and untouched objects with dictionary learning. In [37], the semantic attributes are predicted using tactile data with Direct Attributes Prediction (DAP), and the corresponding categories of untouched materials can be determined by the predicted attributes. In [38], semantic attributes are learnt from both visual data and tactile data using DAP. In [39], a generative model is developed to synthesise tactile features with semantic attribute inputs for tactile ZSL. Although many studies have demonstrated the superiority of the generative model-based method in visual ZSL [40, 41] and the visual modality is able to provide an objective measurement for the target object, there has been no work using the generative model-based method conditioned on visual images in tactile ZSL. Moreover, to the best of the authors' knowledge, the multimodal generative framework, conditioned on both visual images and semantic attributes, has not been attempted before for tactile ZSL. In this work, we employ the GelSight sensor [22], a camera-based tactile sensor, and explore how to use the GelSight sensor for tactile ZSL. Specifically, we propose a multimodal generative model that integrates visual images and semantic attributes to synthesise features of GelSight images of untouched objects for zero-shot tactile recognition, for the first time.

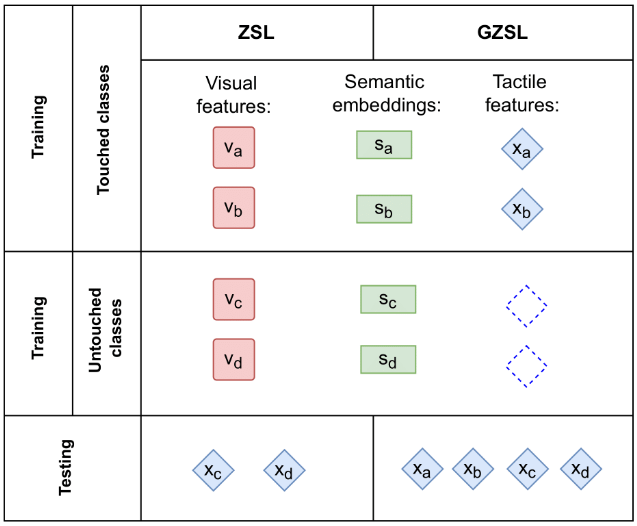


Figure 2: The different configurations between conventional ZSL and GZSL. In the training, visual features, semantic embeddings and tactile features are available for touched classes, whereas only visual features and semantic embeddings are available for untouched classes; in the testing stage, in conventional ZSL, only the tactile features from untouched classes will be tested. In GZSL, the tactile features from both touched classes and untouched classes will be tested.

3. Problem Formulation

The tactile ZSL aims to use tactile sensing to recognise new materials that have no tactile samples during the training process, when they are touched for the first time. Specifically, tactile ZSL can be divided into two stages: the training stage and the inference (testing) stage. The materials to be recognised can also be split into two sets: the touched classes that are touched by the robot during training and the untouched classes that are not touched in training and will be used in the test. To satisfy the zero-shot assumption that there are no tactile training samples from untouched classes, the tactile ZSL requires training a model only on the touched classes, and classifying the tactile data of untouched classes.

An essential component of the tactile ZSL is to use auxiliary information, which can provide additional characteristics of the target object (such as semantic attributes), to bridge the gap between touched and untouched classes. In our multimodal ZSL framework, we use visual images and semantic attributes from both touched materials and untouched materials as auxiliary information and recognise the tactile data of untouched classes based on the knowledge learnt from the touched classes.

Let $x \in \mathbb{R}^{d_x}$, $v \in \mathbb{R}^{d_v}$, $s \in \mathbb{R}^{d_s}$ represent the tactile features, the visual features, and the semantic embeddings learnt from tactile data X, visual images V and semantic attributes S, respectively. $Y = \left\{y^1, y^2, \dots, y^n\right\}$ denotes the corresponding label set. We use subscripts t and u to represent the touched materials and the untouched materials, respectively. The data of touched set can be denoted as $D_t = \left\{(x_t, v_t, s_t, y_t)\right\}$. The data of untouched set can be denoted as

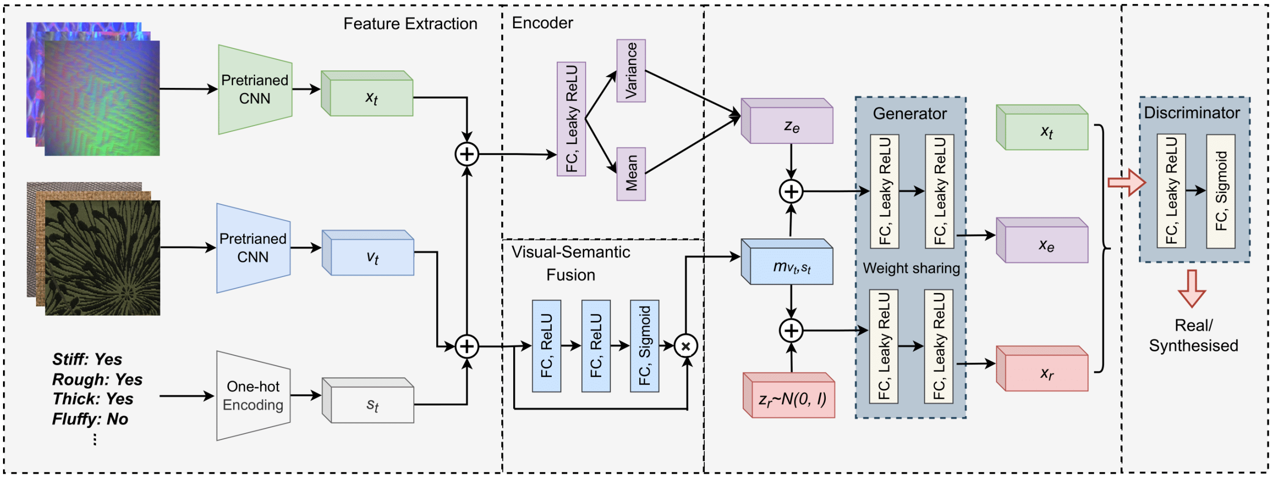


Figure 3: Illustration of the proposed VS2T-ZSL generative model. Our generative framework is composed of five key components: (1) a feature extraction module; (2) an encoder module, denoted as E; (3) a module visual-semantic fusion, labelled VS-F; (4) a generator module, G; and (5) a discriminator module, D. In our model, E encodes the input features to a continuous latent space while VS-F highlights salient features across the visual domain and the semantic domain. The generator module, G, is used to reconstruct tactile features, whereas D discriminates synthesised tactile features from real features. After training, the generator G is used to synthesise the tactile features of untouched materials based on corresponding visual images and semantic attributes.

$D_u = \left\{\left(x_u, v_u, s_u, y_u\right)\right\}$. In tactile ZSL, the touched classes and the untouched classes are disjoint, which means that $Y_t \cap Y_u = \emptyset$.

Following [42], we formalise our tactile ZSL in two different settings: one in conventional ZSL and the other in Generalised Zero-Shot Learning (GZSL). As an example shown in Fig. 2, there are two touched materials and two untouched materials, and we use the subscripts *a, b, c, d* to represent them respectively. In the training stage of conventional ZSL, visual features, semantic embeddings and tactile features are available for materials *a* and *b*, while only visual features and semantic embeddings are available for materials *c* and *d*. The tactile features of materials *c* and *d* are not available in the training stage. In GZSL, the data availability is the same with the ZSL setting in the training stage. The main difference between the conventional ZSL and GZSL comes from the testing stage: In conventional ZSL, tactile features from materials *c* and *d* will be tested. In GZSL, the tactile features from all materials *a*, *b*, *c* and *d* will be tested.

To summarise, given D_t and D_u, in conventional ZSL, a classifier $f_{ZSL} : X_u \rightarrow Y_u$ will be learnt to recognise the tactile data of untouched materials, and GZSL requires learning a classifier $f_{GZSL} : X \rightarrow Y_t \cup Y_u$ to classify the tactile data from both touched and untouched materials. Note that in both cases X_u is not available during training, and is only used in the test. The reasons why we employ two settings for ZSL are as follows: When provided with prior knowledge about which material is unknown, we focus solely on recognising materials that have never been touched before, representing a ZSL setting. However, in real-world applications, unknown materials might mix with previously recognised materials, exemplifying a typical GZSL setting.

4. Methodologies

Our proposed multimodal tactile ZSL framework that learns tactile features from visual images and semantic attributes, i.e., VS2T-ZSL, consists of two main components: a generative model to synthesise tactile features of untouched materials using auxiliary information, and a recognition model that is trained with synthesised features to recognise the untouched materials through corresponding real surface tactile textures.

As illustrated in Fig. 3, our proposed generative model is a combination of Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) [43]. VAE consists of an encoder and a decoder where an explicit distribution can be obtained by encoding the data to a multi-dimensional Gaussian distribution. However, VAE often generates blurry tactile results due to the limitations of reconstruction loss in the decoder [44]. GAN usually contains a generator and a discriminator. In contrast to VAE, GAN learns an implicit distribution where the discriminator is applied to evaluate the quality of the synthesised tactile results from the generator to get sharp and clear features [45]. However, modal collapse [46] may occur during the training of GAN where the generator of network produces only specific outputs regardless of the inputs. To this end, we combine VAE and GAN in our tactile ZSL framework to perform an efficient training and clear generated results.

After training our joint generative model, the generator G is used to synthesise the tactile features of untouched materials from corresponding visual images and semantic attributes. In conventional ZSL, a classifier CLS_u will be trained on the synthesised untouched tactile features to recognise the real untouched tactile features. In GZSL setting, apart from CLS_u, we train another classifier CLS_t with the touched set to recognise the touched tactile data. More details will be discussed in Subsection 4.2.

4.1. Tactile Feature Generative Model

In this section, our tactile feature generative model will be introduced. As illustrated in Fig. 3, our generative model consists of 1) a feature extraction module that extracts the high dimensional features from visual images, tactile textures, and semantic attributes; 2) an encoder E that maps the tactile features and auxiliary information to a continuous latent space; 3) a visual-semantic fusion module $VS-F$ that emphasises the salient features across the visual domain and the semantic domain for the generation process; 4) a generator (decoder) G that samples from the latent space to reconstruct tactile features conditioned on fused features; 5) and a discriminator D that discriminates if the input is a synthesised tactile feature or a real tactile feature (which is extracted from the tactile texture).

Feature extraction module. Instead of synthesising the tactile textures directly, we focus on generating high dimensional tactile features for the tactile recognition task. Firstly, we extract the features from visual images V_t, tactile textures X_t, and semantic attributes S_t to represent the material from different domains. Specifically, two pretrained ResNet50 models [47] are fine-tuned using the visual images and the tactile textures from touched materials respectively, and visual features v_t and tactile features x_t are extracted from the last pooling layer of the fine-tuned models. To obtain semantic representations, we use the one-hot encoding [48] with semantic attributes to get semantic embeddings s_t.

Encoder module. Similar to VAE, the encoder E encodes input features and outputs the mean vector and the variance vector of latent space. We minimise the Kullback–Leibler (KL) divergence between the output latent distribution $q\left(z \mid x_t, v_t, s_t; \phi_E\right)$ and the standard normal distribution $p(z)$ to ensure a continuous latent space for the generation process:

$$\mathcal{L}_{KL} = KL\left(q\left(z \mid x_t, v_t, s_t; \phi_E\right) \| p(z)\right), \tag{1}$$

where ϕ_E is the parameters of the encoder, and $KL(\cdot)$ represents the KL divergence. Compared with a single GAN, the encoder in our joint model makes a continuous space for generation, which enables a better generalisation ability to the untouched materials while synthesising their tactile features.

Visual-semantic fusion module. As the generator has multimodal input of heterogeneous sources for the generation of tactile features, a simple concatenation of visual features and semantic features is not sufficient in practice. Inspired by [49], we design a visual-semantic fusion function to emphasise the task-relevant features across different modalities. The fused features can be given by:

$$m_{v_t,s_t} = f \otimes \left(\sigma\left(W_3 \delta\left(W_2 \delta\left(W_1 f\right)\right)\right)\right), \tag{2}$$

where $f = [v_t, s_t]$, and $[.,.]$ represents the concatenation operation. W_1, W_2, W_3 are learnable matrices, which are implemented by Fully Connected (FC) layers. δ and σ represent a ReLU and a Sigmoid activation function, respectively, and $\otimes$ denotes element-wise product. Compared with a simple concatenation, our visual-semantic fusion module is able to highlight salient features and suppress redundant features across visual and semantic modalities, by assigning different weights to the feature vector.

Generator module. The generator G tries to reconstruct the tactile features using latent vectors with fused features. For the generator G, we minimise ℓ_2 reconstruction loss and pairwise feature matching loss [43] for feature reconstruction:

$$\mathcal{L}_{rec} = \mathbb{E}_{x_t,x_e} \left\|x_t - x_e\right\|_2^2 + \mathbb{E}_{x_t,x_e} \left\|f_D(x_t) - f_D(x_e)\right\|_2^2, \tag{3}$$

where $x_e = G(z_e, m_{v_t,s_t}) \in \mathbb{R}^{d_x}$ represents the synthesised tactile features, and $z_e \sim q\left(z \mid x_t, v_t, s_t; \phi_E\right)$ denotes the latent vectors sampled from the latent distribution. f_D denotes the outputs of the last hidden layer from the discriminator.

Discriminator module. The discriminator D is used to identify synthesised tactile features from real tactile features. Concretely, the discriminator is learnt by minimising the loss:

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}_{x_t}[\log D(x_t)] - \mathbb{E}_{x_e}[\log(1 - D(x_e))] \\ &- \mathbb{E}_{x_r}[\log(1 - D(x_r))],\end{aligned} \tag{4}$$

where $x_r = G(z_r, m_{v_t,s_t}) \in \mathbb{R}^{d_x}$ denotes the synthesised tactile features using random noise $z_r \sim \mathcal{N}(0, I)$ and fused features m_{v_t,s_t}. At the same time, the generator G tries to fool the discriminator, such as by minimising $\mathcal{L}'_{GD} = -\mathbb{E}_{\boldsymbol{x}_e}[\log D(x_e)] - \mathbb{E}_{\boldsymbol{x}_r}[\log D(x_r)]$ from a normal GAN's objective [45]. Through the competition between the discriminator and the generator, the generator is encouraged to synthesise more clear and realistic tactile features.

However, during the training of a GAN, it is found that the real tactile features and the synthesised tactile features are distant between each other, which means that the discriminator can always classify real and synthesised features correctly, i.e., $D(x_t) \rightarrow 1$, $D(x_e) \rightarrow 0$ and $D(x_r) \rightarrow 0$, particularly in the beginning of training. As a result, it is undesirable for the generator to fool the discriminator and a gradient vanishing problem may occur because of $\partial\mathcal{L}'_{GD}/\partial D\left(\boldsymbol{x}_e\right) \rightarrow -\infty$ and $\partial\mathcal{L}'_{GD}/\partial D\left(\boldsymbol{x}_r\right) \rightarrow -\infty$. To address this issue, in addition to $\mathcal{L}_{rec}$, we optimise the generator by minimising the mean feature matching loss [43] between real tactile features and synthesised features:

$$\begin{aligned}\mathcal{L}_{GD} = &\left\|\mathbb{E}_{x_t}[f_D(x_t)] - \mathbb{E}_{x_e}[f_D(x_e)]\right\|_2^2 \\ &+ \left\|\mathbb{E}_{x_t}[f_D(x_t)] - \mathbb{E}_{x_r}[f_D(x_r)]\right\|_2^2.\end{aligned} \tag{5}$$

The centre of synthesised tactile features and the centre of real tactile features should be as close as possible to meet this objective. It solves the gradient vanishing problem when the synthesised feature and the real feature do not overlap with each other, which allows a more stable training and a faster convergence speed, thus assisting the zero-shot tactile textures recognition. To summarise, the overall generator loss can be given as:

$$\mathcal{L}_G = \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{GD}, \tag{6}$$

where λ_1 and λ_2 are hyperparameters weighting the losses of the generator, and are set to $1, 20$ respectively in our experiments through a grid search with a validation set.

For the training process of our proposed generative model, we optimise the encoder E, the generator G and the discriminator D iteratively, with the parameters in feature extraction module frozen. During training, the visual-semantic fusion module is considered as a component of the generator and updated with the generator as a whole. An Adam optimiser [50] is applied to optimise the model and the learning rates are set to 1e−4, 1e−4, 1e−5 for E, G and D, respectively. The training pipeline of our generative model is described in Algorithm 1.

4.2. Tactile Zero-Shot Recognition

After training the generative model using touched materials, we can use the generator to synthesise the tactile features of untouched materials, with semantic features and visual features. The synthesised tactile features can be represented as $x_{syn} = G(z_r, m_{v_u,s_u})$, where $z_r \sim \mathcal{N}(0, I)$.

In the conventional ZSL, a Softmax classifier CLS_u is trained using the synthesised tactile features of untouched materials. The classifier minimises the following loss:

$$L_{cls_u} = -\frac{1}{|\mathcal{T}_{syn}|} \sum_{(x_{syn}, y_u) \in \mathcal{T}_{syn}} \log\left(p\left(y_u \mid x_{syn}; \phi_u\right)\right) \tag{7}$$

where ϕ_u is the parameters of the classifier CLS_u, and $\mathcal{T}_{syn} = \left\{(x_{syn}, y_u)\right\}$. Then, we can use the learnt classifier to classify the tactile features of untouched materials. The label of the test data can be predicted by:

$$\hat{y} = \arg\max_{y \in Y_u} p\left(y \mid x_u; \phi_u\right). \tag{8}$$

Algorithm 1 Training pipeline of our proposed method

```
Input: The tactile features; the visual features; the semantic embeddings; the number of iterations $K$
Output: The parameters of encoder $\phi_E$; the parameters of generator $\phi_G$; the parameters of discriminator $\phi_D$.
for i = 1 to K do
    Sample $\left\{x_t, v_t, s_t, y_t\right\}$ from the touched set;
    Sample $z_e \sim q\left(z \mid x_t, v_t, s_t; \phi_E\right)$;
    Synthesise tactile features $x_e = G(z_e, m_{v_t,s_t})$;
    Sample a batch of random noise $z_r \sim \mathcal{N}(0, I)$;
    Synthesise tactile features $x_r = G(z_r, m_{v_t,s_t})$;
    $\phi_E \leftarrow \phi_E - \nabla_{\phi_E}\left(\mathcal{L}_{KL} + \mathcal{L}_{rec}\right)$
    $\phi_G \leftarrow \phi_G - \nabla_{\phi_G} L_G$ (The visual-semantic fusion module is considered as a part of the generator)
    $\phi_D \leftarrow \phi_D - \nabla_{\phi_D} \mathcal{L}_D$
end for
```

In GZSL, the key is to understand if the input is from touched classes or untouched classes. Considering the mathematical simplicity and tractability, we apply a Gaussian distribution to model the data and measure the distribution of probability density of touched tactile features [28]. An input sample is categorised as touched if it is located in a high-density region, and as untouched if it is not. Concretely, we use the touched set $\left\{x_t^1, x_t^2, \ldots, x_t^n\right\}$ to determine the parameters $\phi_{gau} = (\mu, \sigma^2)$ of the distribution by maximum likelihood estimation:

$$\hat{\phi}_{gau} = \arg\max_{\phi_{gau}} \log \prod_{i=1}^{n} p\left(x_t^i; \phi_{gau}\right), \tag{9}$$

then if $\log p(x; \phi_{gau})$ is greater than a selected threshold β, x is from the touched classes, otherwise, it is from the untouched classes.

Apart from the classifier CLS_u, we train another Softmax classifier CLS_t with touched set to recognise the data from touched materials. Accordingly, the test tactile features can be fed into different classifiers for recognition. The label of the test data can be predicted by:

$$\hat{y} = \begin{cases} \arg\max_{y \in Y_t} \quad p\left(y \mid x; \phi_t\right) & \text{if } \log p\left(x; \phi_{gau}\right) > \beta \\ \arg\max_{y \in Y_u} \quad p\left(y \mid x; \phi_u\right) & \text{otherwise.} \end{cases} \tag{10}$$

4.3. Network Implementation

Table 1 demonstrates the network structures of different components. The encoder E consists of three Fully Connected (FC) layers, where the second and the third layers are both connected to the first layer to produce the mean vector and the variance vector respectively, and each layer is followed by a LeakyReLu activation function. The visual-semantic fusion module includes three FC layers where the first two layers are each followed by a ReLU activation function, and the last layer is followed by a Sigmoid activation function. The generator G is a network with two FC layers, and each layer is followed by a LeakyReLU activation function. For the discriminator D, we use a network with two FC layers followed by a LeakyReLU and a Sigmoid activation functions respectively for binary classification. The networks of CLS_t and CLS_u are both implemented by three FC layers, where first two layers are each followed by a LeakyReLU activation function, and the last layer is followed by a Softmax activation function. The hyperparameters, such as the number of nodes, are tuned by using the validation set so as to achieve a balance between the complexity of the model and the ZSL performance.

All of the networks in our framework are implemented by the Keras using TensorFlow backend with Python. The scikit-learn is also implemented to calculate the parameters of the Gaussian distribution. Our models are trained on a PC with an AMD Ryzen 7 3700X 8-Core Processor, an Nvidia 2080Ti graphic card and 16 GB RAM.

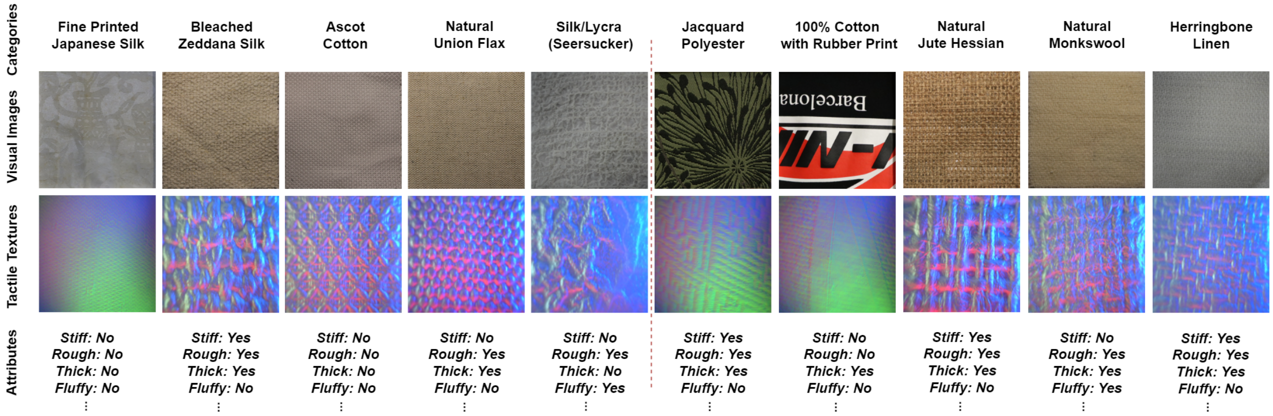

Categories
Visual Images
Tactile Textures
Attributes
Fine Printed Japanese Silk
Bleached Zeddana Silk
Ascot Cotton
Natural Union Flax
Silk/Lycra (Seersucker)
Jacquard Polyester
100% Cotton with Rubber Print
Natural Jute Hessian
Natural Monkswool
Herringbone Linen
Stiff: No Rough: No Thick: No Fluffy: No ⋮
Stiff: Yes Rough: Yes Thick: Yes Fluffy: No ⋮
Stiff: No Rough: No Thick: No Fluffy: No ⋮
Stiff: No Rough: Yes Thick: No Fluffy: No ⋮
Stiff: No Rough: Yes Thick: No Fluffy: Yes ⋮
Stiff: Yes Rough: Yes Thick: Yes Fluffy: No ⋮
Stiff: No Rough: No Thick: Yes Fluffy: No ⋮
Stiff: Yes Rough: Yes Thick: Yes Fluffy: No ⋮
Stiff: No Rough: Yes Thick: Yes Fluffy: Yes ⋮
Stiff: Yes Rough: Yes Thick: Yes Fluffy: No ⋮

Figure 4: Examples in our FabricVST dataset. The first row: categories of different fabrics. The second row: visual images recorded by a digital camera. The third row: tactile textures collected from fabrics by a GelSight sensor. The fourth row: semantic attributes measured by human observations. Left five columns: samples of five example training classes from the training set (we have 40 different fabrics in the training set in total). Right five columns: samples from the test classes (we have 5 unknown fabrics for the test in total).

Table 1
Network implementation in our framework.

Network	Layers	Parameters	
		Number of layers	Number of Nodes
E	FC	3	$2048-2048-2048$
$VS-F$	FC	3	$512-256-2072$
G	FC	2	$2048-2048$
D	FC	2	$512-1$
CLS_t	FC	3	$512-512-40$
CLS_u	FC	3	$512-512-5$

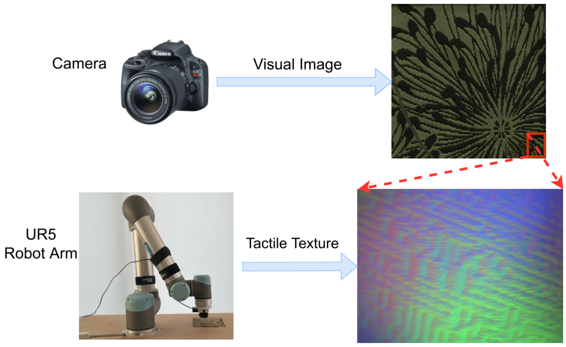


Figure 5: Illustration of data collection. A digital camera is used to take the visual image of the fabric, whereas a GelSight sensor, mounted on the UR5 robot arm, is pressed against the fabric to collect the tactile textures. The pressing locations are recorded, and each tactile texture can be paired with a certain location on the visual image. The example texture is corresponding to the red rectangle region of the visual image.

5. Experimental Setup

Fabric materials. There are 50 different fabrics used in our collected FabricVST dataset. The fabrics include pure materials like cotton, silk, polyester, linen, etc., as well as some mixtures of different materials, such as cotton mixed with wool. Around 30% of the fabrics have different coloured patterns while the rest come in single colours. Most of the fabrics are cloth pieces of $8cm \times 8cm$, whereas some of the fabrics are from daily clothing, and an $8cm \times 8cm$ area from each daily clothing is selected to collect the data. In addition, the fabrics are similar in thickness, ranging from 1mm to 2mm. Specifically, the fabrics are divided randomly with a ratio of 8:1:1, i.e., 40 fabrics are used to collect the training set D_t; 5 fabrics are used to collect the validation set D_{val} to tune the model; and 5 fabrics are used as untouched materials for testing. Some examples are shown in Fig. 4. To scale the model for additional test classes, we can increase the number of neurons in the final layer of the classifier. After this adjustment, the classifier can then be trained with the newly synthesised features.

Table 2
Comparison with other dataset. We compare our collected dataset with other datasets with triple modalities.

Dataset	PHAC-2 [51]	[52]	FabricVST (Ours)
Objects	Household items /Raw materials	Fabrics	Fabrics
Tactile sensor	BioTac	GelSight	GelSight
Number of objects	60	118	50
Number of attributes	24	4	24
Number of trials per object	10	25	225

Data acquisition for training. To train and tune the model with multimodal data, a new dataset, named FabricVST, has been collected, including visual images, tactile textures and semantic attributes from fabrics that are used for training and validation. The test fabrics are only visually inspected and semantically annotated without collecting tactile data. To collect visual data, we use a digital camera Canon 4000D to record the target areas from a fixed distance of $30cm$. The fabric is laid flat on a horizontal plane, while the image plane is parallel with the fabric as well. As shown in the visual images of Fig. 4, we crop the main body of the fabric, and remove the extra background. The resolution after cropping is 1000×1000.

As shown in Fig. 5, a GelSight sensor is applied to record tactile textures by pressing against the fabrics. To collect the data automatically, a Universal Robots UR5 equipped with the GelSight sensor [22] is applied to collect data using Robot Operating System (ROS). The GelSight sensor is typically made up of an internal camera, an elastomer layer on the surface, as well as LEDs with R, G, B colours to provide illumination. When the sensor contacts an object, the camera captures the RGB images of the elastomer's deformation. The sensor has around $1.5cm \times 1.1cm$ perception field with a resolution of 640×480. The surface properties of the object in contact with the sensor's elastomer, such as surface roughness and texture, can be extracted by analysing the tactile images generated by the sensor. The elastomer layer is designed to be soft and sensitive to capture the indentation caused by the interacting object, but the elastomer may be delicate and susceptible to wear and tear, especially with frequent use.

Different from the visual images taken from a distance, the tactile texture is captured only locally by physical contact. As a result, there exists a scale gap between vision and tactile sensing. To reduce the large scale gap between the visual and tactile sensing, we expect that each tactile texture has a paired visual image at the same location of the fabric. Given the size of the fabrics and the perception area of the GelSight sensor, it is possible for us to record the contact location and match the tactile images with certain regions of the fabric. Particularly, the sensor is directed to initiate contact with the fabrics starting from a corner, moving along the horizontal and vertical directions with a step length of $4.9mm$ and $4.6mm$ respectively, until covering an $8cm \times 8cm$ area. As a result, each tactile texture in our dataset is corresponding to a certain location in a visual image, reducing the scale gap significantly. Finally, each fabric is contacted by the sensor 225 times for tactile data collection. To pair with each tactile texture, each visual image is cropped into 225 parts according to the contact locations as well.

The robot arm is controlled to press against fabrics by about 15N with the GelSight sensor. During the data collection, using a small force during contact might result in blurred textures due to the slight deformation, while excessive force might damage the elastomer. After preliminary experiments, 15N is determined as an appropriate force to collect clear and distinct tactile images.

To collect semantic attributes, each fabric is labelled by humans according to its physical characteristics through visual observation, including *stiff, soft, rough, smooth, thick, thin, cool, warm, fluffy, heavy, delicate, durable, stretchable, absorbent, holey, flat, bumpy, patterned, striped, shiny, hairy, embroidered, jacquard, pigment printed*, to characterise the high-level features of fabrics [53]. Each attribute was given a *True* or *False* value according to their properties. The first six paired attributes are exclusive, and only one attribute can be given *True* from each pair. For example, if *stiff* is given *True* value, attribute *soft* must be *False*. The attributes are measured by 5 researchers who work on tactile sensing, and we compute the final attribute values by majority voting. Then, we use the one-hot encoding to get the semantic representation. For example, if the attributes of one piece of fabric is *{True, False, False, ..., True, True}*, the semantic vector would be [1, 0, 0, ..., 1, 1]. Though our fabric semantic attributes may not encompass all aspects of fabric characteristics and are subject to human bias, they can still give us valid and appropriate information to describe the properties of fabrics.

By comparing our FabricVST dataset against other tactile datasets (as shown in Table 2), there are several advantages of our dataset: (1) compared with the datasets in [51, 52], the scale gap between the visual data and the tactile data is reduced significantly as the visual image is cropped according to the contact location; (2) a UR5 robot arm is applied to collect the data automatically, which is more stable and alleviates human error compared to collecting data manually in [52]; (3) each object is explored by the tactile sensor for 225 times, which results in a larger dataset than existing datasets [51, 52] with triple modalities (i.e., vision, touch and semantic attributes).

The dataset also has some potential limitations. The elastomer of the GelSight sensor is very soft and sensitive to capture the indentation caused by the physical contact. However, it is fragile and prone to damage with frequent use. Even though the sensor can be calibrated after changing the elastomer, the response to the stimulus might vary slightly among tactile sensors due to manufacturing inaccuracies.

Different information embeddings of visual data and tactile data. The visual data are captured from a distance by the digital camera where the global information such as appearance, shapes and colours of materials will be recorded. Different from visual data, tactile data are collected by physical contact between objects and an optical tactile sensor, i.e., the GelSight sensor. When the GelSight sensor interacts with the objects, the elastomer on the sensor will be deformed in response to the contact force, and the surface geometry will be mapped into the deformation, which is captured by the camera inside the sensor. In [22], it is demonstrated that the tactile images from the GelSight have the ability to indicate tactile textures, contact force, surface height, hardness, etc. Though the visual data and tactile data are of the same format (i.e., images), they reflect different properties of the objects.

Tactile zero-shot recognition task. After training the model with the training set D_t and validation set D_{val}, we enable the robot to press against the untouched test materials to collect tactile textures with the GelSight sensor for zero-shot recognition. Concretely, each test material is pressed by the sensor to collect tactile textures 225 times at different locations, with the same tactile data sampling method described in data acquisition, which are then recognised by the trained model. Some robotic experiment demos are shown on our website[1].

[1] https://sites.google.com/view/multimodalzsl

6. Experimental Results and Analysis

In this section, a series of experiments are conducted to evaluate our proposed VS2T-ZSL method for the tactile ZSL problem. The goal of the experiments is three-fold: 1) To learn how different components in our proposed structure improve the ZSL results; 2) To investigate the synergistic effects of multimodal input; 3) To evaluate the effectiveness of our proposed method against other methods in both ZSL and GZSL settings.

6.1. Results of VS2T-ZSL

To investigate how different components work in synthesising tactile features in ZSL, we conduct an ablation study for our proposed VS2T-ZSL structure. Due to the fact that G and D are necessary components to synthesise tactile features, we explore the effect of removing E and VS-F individually, as well as removing them together. To ensure that our results are determined by the proposed method rather than the chosen objects, we repeat the aforementioned process of dividing materials randomly for five times and calculate the average results over all splits to validate the robustness of our proposed method.

As shown in Table 3, the recognition accuracy of untouched materials is given to evaluate the performance in ZSL. Our proposed VS2T-ZSL is able to achieve a higher recognition accuracy of 83.06% to recognise the unknown materials, compared to the results when a certain component is removed. In particular, there is an obvious drop by 4.85% when E is removed, which demonstrates that the continuous latent space generated by the encoder makes a great contribution to synthesising untouched tactile features. The absence of the VS-F produces inferior recognition results, decreased by 3.56%, compared to the result of VS2T-ZSL. This indicates that the visual-semantic fusion module is able to select the salient features across different modalities for the generation task. Moreover, there is a decrease of 8.45% when both E and VS-F are removed, compared to the result of VS2T-ZSL.

Furthermore, the Wasserstein distance is implemented to measure the distance between the distributions of the real tactile features and synthesised features of untouched materials. As shown in Table 3, it is observed that the Wasserstein distance between synthesised tactile features and real tactile features is 1,165 by using the VS2T-ZSL, which are hundreds less than the results from the ablated structures. It means that the distribution of the generated features with the VS2T-ZSL method are closest to the real distribution.

The cosine similarity is also given to evaluate the similarity between each synthesised tactile features and real tactile features, and the average scores are shown in Table 3. It can be seen that the proposed VS2T-ZSL achieves the highest similarity score, i.e., 0.51, among all the structures, which demonstrates that the synthesised features are more similar to the real features by using our proposed structures. It can be concluded that our proposed VS2T-ZSL method, which consists of a feature extraction module, an encoder, a visual-semantic fusion module, a generator and a discriminator, allows the generated features of untouched material to be more realistic, thus providing a better performance in tactile ZSL.

Table 3
Recognition accuracy of untouched materials using various network structures.

Network structure				Average	Warrsserstein	Cosine
E	VS-F	G^*	D^*	Accuracy ↑	Distance ↓	Similarity ↑
		✓	✓	74.61%	1788	0.41
	✓	✓	✓	78.21%	1571	0.43
✓		✓	✓	79.50%	1320	0.47
✓	✓	✓	✓	**83.06%**	**1165**	**0.51**

Table 4
Recognition accuracy of untouched materials using different input modalities.

Input Modality	Average Accuracy ↑	Warrsserstein Distance ↓	Cosine Similarity ↑
S2T-ZSL	55.07%	1458	0.31
V2T-ZSL	72.71%	1364	0.46
VS2T-ZSL	**83.06%**	**1165**	**0.51**

6.2. Multimodal vs. Unimodal Input

To investigate the synergistic effect of multimodal input, we compare our VS2T-ZSL with the methods using unimodal inputs on our FabricVST dataset. Firstly, we only apply semantic attributes to generate the tactile features of untouched materials in ZSL (S2T-ZSL). Secondly, we only use the visual images to synthesise the tactile features (V2T-ZSL).

As shown in Table 4, the recognition accuracy is only 55.07% from S2T-ZSL, while it is 72.71% from V2T-ZSL, which are much lower than the 83.06% recognition accuracy from VS2T-ZSL. The Wasserstein distance and the cosine similarity also show that the generated features with multimodal input are much closer to the real features than with unimodal inputs. The results indicate that multimodal input enables us to produce more realistic tactile features from different perspectives with multiple sources, which is beneficial to zero-shot recognition.

Particularly, it can be observed that the recognition accuracy has an obvious drop, by 17.64% and 27.99% respectively, while using semantic attributes only compared to the results by using visual input only or multimodal input. A possible reason causes this difference is due to the limited number of semantic attributes used in our approach. There are only 24 attributes used to describe fabric characteristics. The semantic attributes cannot encompass all of the fabric's characteristics and may have many overlapping values between different materials. As shown in Fig. 6 (a) (b) (c), we take the data from the first materials split as an example and compute the cosine similarity between the mean values of the input features of each untouched class for multimodal input and unimodal input respectively. It can be seen that the semantic input of each class is more similar to each

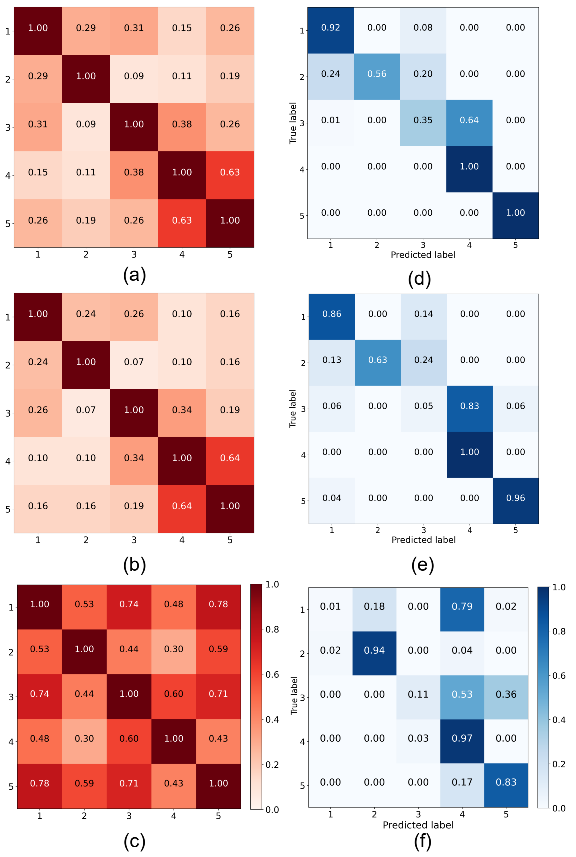


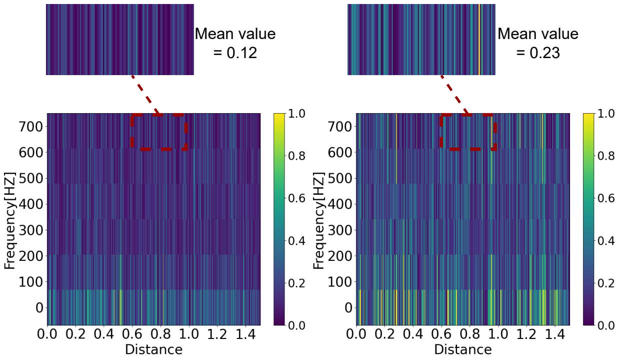
Figure 6: Cosine similarities: (a) similarity of multimodal input between each class, (b) similarity of visual input between each class, (c) similarity of semantic input between each class. **Normalised confusion matrices**: (d) results of the model with multimodal input; (e) results of the model with the visual input; (f) results of the model with the semantic input.

Table 5

The recognition accuracy, Wasserstein distance and Cosine similarity of untouched materials using different numbers of semantic attributes (↑ indicating the higher the values are, the better performance or the features are more similar; ↓ indicating that the lower the values are, the features are more similar).

Number of Attributes	Accuracy ↑	Warsserstein Distance ↓	Cosine Similarity ↑
6	24.36%	1231	0.32
12	38.94%	1230	0.32
18	43.35%	1223	0.36
24	**57.16%**	**1194**	**0.39**

other compared to the visual input and multimodal input. Since the input of each class is similar, it will be difficult to generate recognisable tactile features from semantic information, which results in a lower recognition accuracy in ZSL. Compared with semantic attributes, visual input and multimodal input (as can be seen in Fig. 6 (a) (b)) have a smaller similarity between each class, which enables us to synthesise discriminative tactile features for tactile ZSL tasks.



(a)

(b)

Figure 7: The spectrograms of synthesised tactile features. (a) The spectrogram of the synthesised tactile features of the material Jacquard Polyester (shown in Fig. 4 column 6). (b) The spectrogram of synthesised tactile features of the material 100% Cotton with Rubber Print (shown in Fig. 4 column 7). The zoomed-in regions are extracted from the highest frequency, specifically 682.5 Hz in our case. The zoomed-in areas span from 0.6 to 1.0 on the distance axis and the colour displayed corresponds to the intensity of the signals.

Additionally, we investigate the impact of the number of semantic attributes on the performance of tactile ZSL. As shown in Table 5, in the experiments with S2T-ZSL, we test different numbers of attributes, specifically 6, 12, 18, and 24, and these attributes are randomly selected from the pool of 24 attributes. It can be seen that when the number of attributes increases, there is an upward trend in the recognition results. Specifically, there is a 32.8% increase in recognition accuracy when the number of attributes rises from 6 to 24. Furthermore, both Wasserstein distance and the cosine similarity results indicate that the generated features are more similar to the real features when provided with a greater number of semantic attributes.

To visualise the synthesised tactile features of our proposed method, we use the method from [26] to showcase the spectrograms of the synthesised tactile features. Fig. 7(a) is the spectrogram of the synthesised tactile features of the material Jacquard Polyester (shown in Fig. 4 column 6), while Fig. 7 (b) is the spectrogram of synthesised tactile features of the material 100% Cotton with Rubber Print (shown in Fig. 4 column 7). In more detail, we have provided zoomed-in visualisations for two different fabric textures. These zoomed-in views give a compact quantitative representation of the differences in the two textures. Furthermore, we have added the average frequency values for these zoomed-in areas to highlight distinctions in the frequency signals of synthesised features: the smoother material exhibits a higher frequency of signals (100% Cotton with Rubber Print, right in the figure), indicating a smoother surface; in contrast, the rougher material exhibits a lower frequency of signals (Jacquard Polyester, left in the figure), indicating a rougher surface. This illustrates the effectiveness of our generative model in capturing distinct tactile features from the other modalities.

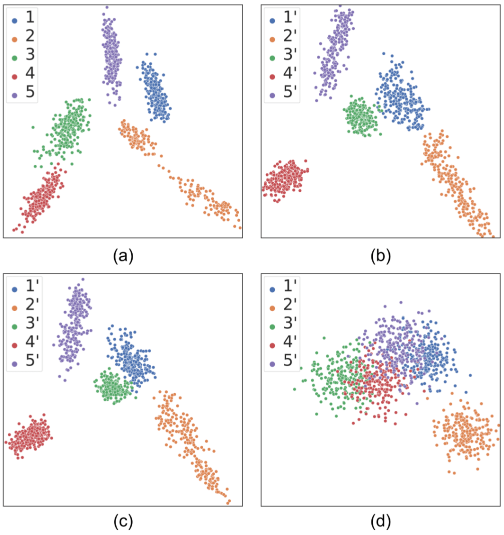


Figure 8: Tactile feature distributions. (a) real tactile features of untouched materials; (b) synthesised tactile features with multimodal input; (c) synthesised tactile features with visual input; (d) synthesised tactile features with semantic input.

To further analyse the synergistic effect of the multimodal input, we illustrate the normalised confusion matrices using different input modalities in Fig. 6 (d) (e) (f). Specifically, we can see that only 5% and 11% tactile features of class 3 (Natural Jute Hessian, see the 8th column in Fig. 4) are classified correctly using the visual input and semantic input, respectively (as shown in Fig. 6 (e) (f)), whereas 35% tactile features are classified correctly with the multimodal input (as shown in Fig. 6 (d)). It indicates that by combining visual and semantic information, we can represent the characteristics of materials from different domains and mitigate biases in a single domain, resulting in a synergistic effect where the result cannot be achieved with a single modality.

What is more, it is worth noting that most tactile data of class 3 (Natural Jute Hessian) are misclassified with both multimodal input and unimodal input. To better understand the reason, we analyse it in terms of both the output distribution and the input similarity. The distributions of real tactile features and synthesised tactile features of untouched materials are shown in Fig. 8, by using Principal Component Analysis (PCA) where we reduce the data dimension to visualise high dimensional tactile features in 2D. If the classifier CLS_u is trained with the synthesised features (e.g., features in Fig. 8 (b)) and is used to predict the categories of real tactile features (in Fig. 8 (a)), the testing tactile features of class 3 (Natural Jute Hessain marked in green) have a higher probability to be misclassified to a closer cluster in synthesised features of class 4 (Natural Monkswool marked in red) due to the bias in the synthesised data. With respect to input similarity, as shown in Fig. 6, class 3 (Natural Jute Hessian) and category that is incorrectly predicted, such as class 4 (Natural Monkswool) have relatively higher input similarity than other classes which could potentially lead to a mix-up for recognition.

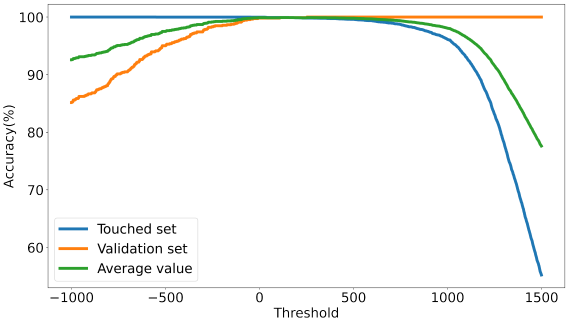


Figure 9: The yellow line represents how much validation data is identified as untouched classes when the threshold is tuned. The blue line represents how much touched data is identified as touched classes when the threshold is tuned. The green line denotes the average value of the two percentages above.

It should be noted that apart from the inaccuracies in the synthetic phase, the classifier is another source that potentially introduces errors in ZSL. To validate this, we use the real tactile features to train the classifier rather than synthetic tactile features, aiming to assess if it can identify the materials correctly. Consequently, the recognition accuracy is 99.6%, falling short of a complete 100%, which means that the classifier also introduces additional errors in the ZSL.

6.3. Comparison with Other Methods

Here, we first compare our method against other generative model-based methods for the tactile ZSL [39, 54]. Particularly, we would like to compare our multimodal generative methods with methods based on semantic input [39] or visual input. Since there is no prior work using generative models conditioned on visual input, the generative model from [54] is implemented as an alternative, in which the visual images are used to generate the tactile textures. In addition to the comparison of generative methods, we also compare our results to those using projection (or mapping) methods, i.e., DAP [37] and VT-FC-ZSL [38]. The DAP employs the touched materials to train a projection model to project tactile data to attribute embedding space, then the model is applied on the untouched materials directly for prediction. The VT-FC-ZSL shares a similar mechanism with DAP, but it uses multimodal input (visual data and tactile data) to train the model and prediction. To make the model suitable for our tactile textures, a fine-tuned ResNet50 is applied to extract the tactile features from the last pooling layer for the methods in [37, 38, 39]. Note that the methods from [37, 38, 54] do not involve the setup for GZSL. We adapt our proposed settings and train a ResNet50 network as CLS_t to simplify the procedure of GZSL.

As detailed in Section 4.2, in GZSL, we fit the real tactile features into a Gaussian distribution first. Then, if $\log p(x; \phi_{gau})$ is greater than a selected threshold β, the input tactile feature x is from touched classes, otherwise, it is from untouched classes. Specifically, we use tactile features from D_t and D_{val} to tune the threshold β. As shown in Fig. 9, taking the data from the first materials split as an example, we select the threshold β that is able to maximise the average

Table 6

Comparison results. We compare our results against other methods in both ZSL and GZSL settings. acc_t and acc_u represent the accuracy of touched and untouched materials, respectively; $H = \frac{2*acc_u*acc_t}{acc_u+acc_t}$ denotes their harmonic mean.

Methods	Generative	Projection based	Unimodal	Multimodal	ZSL	GZSL		
					acc ↑	acc_t ↑	acc_u ↑	H ↑
DAP [37]		✓	✓		43.56%	94.11%	43.47%	59.47%
VT-FC-ZSL [38]		✓		✓	63.20%	94.11%	62.58%	75.17%
Abderrahmane *et al.* [39]	✓		✓		47.20%	83.44%	47.11%	60.22%
Lee *et al.* [54]	✓		✓		51.29%	89.00%	51.20%	65.00%
VS2T-ZSL	✓			✓	**76.36%**	**94.56%**	**76.00%**	**84.27%**

accuracy of classifying $x_t \in D_t$ as touched and $x_{val} \in D_{val}$ as untouched.

As shown in Table 6, the recognition accuracy of our proposed VS2T-ZSL is 76.36%, which is 25.07% and 29.16% higher than the results of other generative methods from [54, 39] in the ZSL setting, respectively. A similar trend can be found in the GZSL for both the touched and untouched classes. Our proposed VS2T-ZSL method has the highest harmonic mean accuracy of 84.27%, which shows the superiority of our VS2T-ZSL over other generative methods.

In the results of projection methods, the use of semantic and visual modalities together in VT-FC-ZSL improves the tactile recognition result largely by 19.64%, compared to the result by using semantic attributes only in DAP in ZSL setting. In GZSL setting, the harmonic mean of the method VT-FC-ZSL is 15.7% higher than the results of DAP. However, VS2T-ZSL performs better than both of these projection methods, which demonstrates the effectiveness of our generative model.

7. Discussion

In this section, we discuss several aspects that affect the results of the tactile ZSL.

7.1. Unimodal vs. multimodal input for Tactile ZSL

The application of the multimodal input allows us to measure the objects from different domains, increasing the dimensions to reflect the properties of different objects. As shown in Fig. 8, the generated features with multimodal input are closer to the real distribution compared to the generated features with unimodal input, and the boundaries of each class are clearer than the others. As a result, the prediction of real tactile features will be more accurate using the model trained on the synthesised tactile features with multimodal input.

However, due to the implementation of multiple modalities, the computational efficiency decreases compared with the model using single modality. For example, 14.36 millions parameters are trained in the generator for VS2T-ZSL method, while 8.44 millions parameters are trained for S2T-ZSL method. Therefore, there is a trade-off between the amount of computation and the accuracy.

7.2. Visual input vs. semantic input for tactile ZSL

Compared to multimodal input, unimodal data is much easier to be collected. Here, we discuss which modality is more effective in tactile ZSL using unimodal input. While visual images provide an objective measurement that includes fine details of fabrics, semantic attributes offer a high-level description. Both of them are able to measure the characteristics of target objects from different perspectives. However, different from visual images that provide objective measurement with high resolution, the semantic attributes are constrained to a limited number of attributes and are subject to human bias. From Table 4, we can observe that the recognition results using visual input are better than those using semantic input. It also explains why the visual-based approaches [38, 54] achieve better results than the baseline approaches [37, 39] that rely only on semantic attributes in Table 6.

Consequently, the visual input is a more objective and accurate auxiliary information to measure the object for tactile ZSL. A possible way to improve the results of semantic input is to use continuous-valued attributes. Compared to binary attributes, continuous-valued attributes can demonstrate the level of properties to improve the recognition results. Moreover, a larger number of attributes are expected to increase the dimension of representation.

7.3. Projection-based vs. generative model-based approaches

As shown in Table 6, we apply two projection methods [37, 38], two generative model-based methods [39, 54], as well as our proposed multimodal generative method for comparison. We can see that the performance of the generative methods is better than the projection methods with the same input. Concretely, the recognition accuracy of our VS2T-ZSL is 13.16% higher than the result from VT-FC-ZSL with multimodal input in ZSL. The accuracy of [39] is 3.64% higher than the result from DAP with unimodal input. It demonstrates the effectiveness of the generative method in tactile ZSL.

For the projection-based method in tactile ZSL, a projection function is learnt between the tactile features and semantic embeddings. However, if the projection function is only learnt from touched classes and is applied to the untouched classes directly, the projected features and semantic embeddings may be kept away due to the bias in touched classes, which is a domain shift problem [55, 56]. In contrast, the generative model-based method tries to reconstruct the tactile features using the available auxiliary information, e.g., tactile cues embedded in visual images or semantic attributes. By reconstructing the tactile features, one constraint is met: tactile cues from the visual/semantic domain have to be preserved in the tactile feature generation. Moreover, due to the fact that the auxiliary information is from the same latent space and is conceptually interlinked, the generator can generate meaningful tactile features for untouched materials, which is able to alleviate the domain shift problem [40, 57].

7.4. Limitations of the proposed approach

The proposed method, while promising, does come with certain potential limitations that may constrain its performance in tactile ZSL.

Firstly, extensive effort is required from annotators to label the semantic attributes of fabric textures for our dataset FabricVST, making the process both resource-intensive and susceptible to human bias. An alternative approach could be to leverage the material compositions typically listed by fabric manufacturers for garments or fabric samples. These compositions could serve as a valuable source of semantic information, such as identifying that a fabric made of silk typically exhibits the semantic attribute of “smooth”. It presents an intriguing avenue for exploration to determine if tactile ZSL Learning can be enhanced by incorporating visual data and material compositions derived from fabric labels.

Secondly, in our approach tactile data is collected using an optical tactile sensor that uses a camera to capture the deformation of a silicone layer above it. This methodology results in tactile images that share the same format as visual images captured by conventional digital cameras. While this provides valuable tactile information, the tactile data collected via this approach may primarily capture tactile information that aligns with the visual appearance of materials. This could differ from the physical signals obtained through other tactile sensors, such as MEMS based tactile sensors [18]. An interesting avenue for future research would be to explore how tactile ZSL performs when utilising different types of tactile sensors, thus enriching our understanding of cross-modal transfer capabilities.

Thirdly, while the incorporation of multimodal input does enhance the performance of tactile ZSL, it also comes at the cost of increased computational resources due to larger data volume. Therefore, a key area for future work could focus on optimising the efficiency of utilising multimodal input, thus addressing the associated computational overhead.

8. Conclusions

In this work, we propose a novel multimodal generative framework to address the tactile ZSL problem. A new dataset, *FabricVST*, including visual images, semantic attributes and tactile data, is collected from different kinds of fabrics. This dataset is larger than any other existing datasets in the field with these triple modalities. A joint generative model, which integrates the VAE and GAN, is proposed to synthesise the tactile features of untouched materials from visual images and semantic attributes. It is the first work that uses multimodal input to generate tactile features for tactile ZSL in robotic perception. In the ablation study, the results demonstrate that our multimodal approach which combines semantic attributes, representing high-level characteristics, and visual images, providing tactile cues from sight, can achieve a synergistic effect, compared to using a single modality. The extensive experimental results show that the proposed method enables a high recognition accuracy of 83.06% in classifying unknown materials using tactile sensing. Our proposed VS2T-ZSL method allows the robots to recognise the materials never touched before.

In the future, we will apply the VS2T-ZSL method to different unstructured scenarios where the tactile information is more robust and cannot be easily changed, even if appearance or colour has been altered, e.g., folded clothes. Additionally, different tactile sensors will be evaluated in our experiments, such as the GelTip sensor [58, 59], the GelFinger sensor [60] or the MEMS based tactile sensors. Compared with the GelSight sensor, there are significant disparities in data types, scales, and modalities between the features obtained from a vision camera and those from MEMS. Although this gap could present potential challenges in achieving cross-modal transfer for tactile zero-shot recognition, we plan to employ these differences to validate our proposed method. We will also test our methods in other tasks, for example, zero-shot tactile learning for grasping and manipulation of objects with auxiliary information of visual and semantic attributes, to generalise our methods on different tasks.

Furthermore, in our proposed framework, we use visual images and semantic attributes to generate the data of inaccessible tactile domain. Likewise, it is possible to switch the input and output domains, such as using visual and tactile data to synthesise semantic attributes, or using tactile and semantic attributes to synthesise visual data using our framework, which is promising to solve the semantic/visual ZSL problem in a multimodal manner in the future.

References

[1] S. Chitta, J. Sturm, M. Piccoli, W. Burgard, Tactile sensing for mobile manipulation, IEEE Trans. Robot. 27 (2011) 558–568.

[2] Y. Zhao, X. Jing, K. Qian, D. F. Gomes, S. Luo, Skill generalization of tubular object manipulation with tactile sensing and sim2real learning, Robotics and Autonomous Systems 160 (2023) 104321.

[3] D. Guo, H. Liu, B. Fang, F. Sun, W. Yang, Visual affordance guided tactile material recognition for waste recycling, IEEE Trans. Autom. Sci. Eng. (2021).

[4] D. Hu, L. Bo, X. Ren, Toward robust material recognition for everyday objects., in: BMVC, volume 2, Citeseer, 2011, p. 6.

[5] Y. LeCun, F. J. Huang, L. Bottou, Learning methods for generic object recognition with invariance to pose and lighting, in: CVPR, volume 2, 2004, pp. II–104.

[6] D. G. Lowe, Object recognition from local scale-invariant features, in: ICCV, volume 2, 1999, pp. 1150–1157.

[7] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, A. Vedaldi, Describing textures in the wild, in: CVPR, 2014, pp. 3606–3613.

[8] W. Zhai, Y. Cao, J. Zhang, Z.-J. Zha, Deep multiple-attribute-perceived network for real-world texture recognition, in: ICCV, 2019, pp. 3613–3622.

[9] S. Luo, J. Bimbo, R. Dahiya, H. Liu, Robotic tactile perception of object properties: A review, Mechatronics 48 (2017) 54–67.

[10] G. Cao, Y. Zhou, D. Bollegala, S. Luo, Spatio-temporal attention model for tactile texture recognition, in: IROS, 2020, pp. 9896–9902.

[11] P. Giguere, G. Dudek, A simple tactile probe for surface identification by mobile robots, IEEE Trans. Robot. 27 (2011) 534–544.

[12] E. Kerr, T. M. McGinnity, S. Coleman, Material recognition using tactile sensing, Expert Syst. Appl. 94 (2018) 94–111.

[13] A. W. Gottfried, S. A. Rose, W. H. Bridger, Cross-modal transfer in human infants, Child development (1977) 118–123.

[14] R. B. Banati, G. Goerres, C. Tjoa, J. P. Aggleton, P. Grasby, The functional anatomy of visual-tactile integration in man: a study using positron emission tomography, Neuropsychologia 38 (2000) 115–124.

[15] S. Luo, L. Zhu, K. Althoefer, H. Liu, Knock-knock: acoustic object recognition by using stacked denoising autoencoders, Neurocomputing 267 (2017) 18–24.

[16] J. Edwards, J. Lawry, J. Rossiter, C. Melhuish, Extracting textural features from tactile sensors, Bioinspir. Biomim. 3 (2008) 035002.

[17] N. Jamali, C. Sammut, Majority voting: Material classification by tactile sensing using surface texture, IEEE Trans. Robot. 27 (2011) 508–521.

[18] F. De Boissieu, C. Godin, B. Guilhamat, D. David, C. Serviere, D. Baudois, Tactile texture recognition with a 3-axial force mems integrated artificial finger., in: RSS, 2009, pp. 49–56.

[19] S.-H. Kim, J. Engel, C. Liu, D. L. Jones, Texture classification using a polymer-based mems tactile sensor, J. Micromech. Microeng. 15 (2005) 912.

[20] T. Taunyazov, H. F. Koh, Y. Wu, C. Cai, H. Soh, Towards effective tactile identification of textures using a hybrid touch approach, in: ICRA, 2019, pp. 4269–4275.

[21] J.-P. Roberge, L. L'Écuyer-Lapierre, J. Kwiatkowski, P. Nadeau, V. Duchaine, Tactile-based object recognition using a grasp-centric exploration, in: CASE, IEEE, 2021, pp. 494–501.

[22] W. Yuan, S. Dong, E. H. Adelson, Gelsight: High-resolution robot tactile sensors for estimating geometry and force, Sensors 17 (2017) 2762.

[23] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, N. F. Lepora, The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies, Soft Robot. 5 (2018) 216–227.

[24] W. Yuan, Y. Mo, S. Wang, E. H. Adelson, Active clothing material perception using tactile sensing and deep learning, in: ICRA, 2018, pp. 4842–4849.

[25] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, R. Fuentes, Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition, in: ICRA, 2018, pp. 2722–2727.

[26] G. Cao, J. Jiang, N. Mao, D. Bollegala, M. Li, S. Luo, Vis2hap: Vision-based haptic rendering by cross-modal generation, arXiv preprint arXiv:2301.06826 (2023).

[27] C. H. Lampert, H. Nickisch, S. Harmeling, Attribute-based classification for zero-shot visual object categorization, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2013) 453–465.

[28] R. Socher, M. Ganjoo, C. D. Manning, A. Y. Ng, Zero-shot learning through cross-modal transfer, in: Adv. Neural Inf. Process. Syst., 2013, pp. 935–943.

[29] D. Das, C. G. Lee, Zero-shot image recognition using relational matching, adaptation and calibration, in: IJCNN, 2019, pp. 1–8.

[30] L. Zhang, T. Xiang, S. Gong, Learning a deep embedding model for zero-shot learning, in: CVPR, 2017, pp. 2021–2030.

[31] Z. Akata, F. Perronnin, Z. Harchaoui, C. Schmid, Label-embedding for attribute-based classification, in: CVPR, 2013, pp. 819–826.

[32] J. Lei Ba, K. Swersky, S. Fidler, et al., Predicting deep zero-shot convolutional neural networks using textual descriptions, in: ICCV, 2015, pp. 4247–4255.

[33] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, Transductive multi-view zero-shot learning, IEEE Trans. Pattern Anal. Mach. Intell. 37 (2015) 2332–2345.

[34] Y. Xian, T. Lorenz, B. Schiele, Z. Akata, Feature generating networks for zero-shot learning, in: CVPR, 2018, pp. 5542–5551.

[35] R. Felix, I. Reid, G. Carneiro, et al., Multi-modal cycle-consistent generalized zero-shot learning, in: ECCV, 2018, pp. 21–37.

[36] H. Liu, F. Sun, B. Fang, D. Guo, Cross-modal zero-shot-learning for tactile object recognition, IEEE Trans. Syst., Man, Cybern.: Syst. 50 (2018) 2466–2474.

[37] Z. Abderrahmane, G. Ganesh, A. Crosnier, A. Cherubini, Haptic zero-shot learning: Recognition of objects never touched before, Robot. Auton. Syst. 105 (2018) 11–25.

[38] Z. Abderrahmane, G. Ganesh, A. Crosnier, A. Cherubini, Visuo-tactile recognition of daily-life objects never seen or touched before, in: ICARCV, 2018, pp. 1765–1770.

[39] Z. Abderrahmane, G. Ganesh, A. Crosnier, A. Cherubini, A deep learning framework for tactile recognition of known as well as novel objects, IEEE Trans. Ind. Inform. 16 (2019) 423–432.

[40] Y. Ye, Y. He, T. Pan, J. Li, H. T. Shen, Alleviating domain shift via discriminative learning for generalized zero-shot learning, IEEE Trans. Multimed. (2021).

[41] J. Li, M. Jing, L. Zhu, Z. Ding, K. Lu, Y. Yang, Learning modality-invariant latent representations for generalized zero-shot learning, in: 28th ACM Int. Conf. Multimedia, 2020, pp. 1348–1356.

[42] W.-L. Chao, S. Changpinyo, B. Gong, F. Sha, An empirical study and analysis of generalized zero-shot learning for object recognition in the wild, in: ECCV, 2016, pp. 52–68.

[43] J. Bao, D. Chen, F. Wen, H. Li, G. Hua, Cvae-gan: fine-grained image generation through asymmetric training, in: ICCV, 2017, pp. 2745–2754.

[44] L. Cai, H. Gao, S. Ji, Multi-stage variational auto-encoders for coarse-to-fine image generation, in: SDM, 2019, pp. 630–638.

[45] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[46] N.-T. Tran, T.-A. Bui, N.-M. Cheung, Dist-gan: An improved gan using distance constraints, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 370–385.

[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: CVPR, 2016, pp. 770–778.

[48] G. Hackeling, Mastering Machine Learning with scikit-learn, Packt Publishing Ltd, 2017.

[49] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: CVPR, 2018, pp. 7132–7141.

[50] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).

[51] V. Chu, I. McMahon, L. Riano, C. G. McDonald, Q. He, J. M. Perez-Tejada, M. Arrigo, T. Darrell, K. J. Kuchenbecker, Robotic learning of haptic adjectives through physical interaction, Robot. Auton. Syst. 63 (2015) 279–292.

[52] W. Yuan, S. Wang, S. Dong, E. Adelson, Connecting look and feel: Associating the visual and tactile properties of physical materials, in: CVPR, 2017, pp. 5580–5588.

[53] P. Venkatraman, Fabric properties and their characteristics, Materials and technology for sportswear and performance apparel (2015) 53–86.

[54] J.-T. Lee, D. Bollegala, S. Luo, "touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception, in: ICRA, 2019, pp. 4276–4282.

[55] E. Kodirov, T. Xiang, Z. Fu, S. Gong, Unsupervised domain adaptation for zero-shot learning, in: ICCV, 2015, pp. 2452–2460.

[56] E. Kodirov, T. Xiang, S. Gong, Semantic autoencoder for zero-shot learning, in: CVPR, 2017, pp. 3174–3183.

[57] W. Wang, Y. Pu, V. K. Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, L. Carin, Zero-shot learning via class-conditioned deep generative models, in: AAAI, 2018.

[58] D. F. Gomes, P. Paoletti, S. Luo, Generation of gelsight tactile images for sim2real learning, IEEE Robot. Autom. Lett. 6 (2021) 4177–4184.

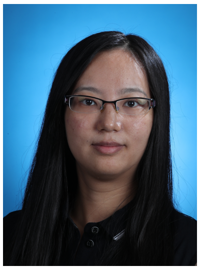
[59] D. F. Gomes, Z. Lin, S. Luo, Blocks world of touch: Exploiting the advantages of all-around finger sensing in robot grasping, Frontiers in Robotics and AI 7 (2020) 541661.

[60] Z. Lin, J. Zhuang, Y. Li, X. Wu, S. Luo, D. F. Gomes, F. Huang, Z. Yang, Gelfinger: A novel visual-tactile sensor with multi-angle tactile image stitching, IEEE Robotics and Automation Letters (2023).

Guanqun Cao received the B.S. degree from Nanjing Audit University, and the M.Sc. degree from the University of Liverpool. He is currently a Ph.D. candidate in the Department of Computer Science, the University of Liverpool. His research interests include tactile perception and multimodal perception.

Jiaqi Jiang received the B.S. and M.S degrees from Beijing Institute of Technology, in 2016 and 2019, respectively. He is currently a Ph.D. candidate in the Department of Engineering, King's College London. He was a Ph.D. candidate in the Department of Computer Science, the University of Liverpool. His research interests include robot grasping and sensory synergy of vision and touch.

Danushka Bollegala is a Professor at the Department of Computer Science, University of Liverpool. He is also an Amazon Scholar. He obtained his PhD in 2009 from the University of Tokyo, Japan. He has received many prestegious awards such as the IEEE Young Author Award, GECCO best paper award and JSAI best journal paper award. He has published over 170 peer-reviewed papers in top international venues in NLP, ML and AI.



Min Li is an Associate Professor at the School of Mechanical Engineering, Xi'an Jiaotong University, China. She received her B.Sc. degree in Mechanical Engineering and her M.Sc. degree in Agricultural Mechanization Engineering from Northwest A&F University, China, in 2007 and 2010, respectively. She was awarded the Ph.D. degree in Robotics at King's College London, UK, in 2014. From 2015 to 2017, she was a Lecturer with Xi'an Jiaotong University. Her research interests include haptic feedback for robots, soft robots, rehabilitation robots.

Shan Luo is a Reader (Associate Professor) at the Department of Engineering, King's College London. Previously, he was a Lecturer at the University of Liverpool, and Research Fellow at Harvard University and University of Leeds. He was also a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT. He received the B.Eng. degree in Automatic Control from China University of Petroleum, Qingdao, China, in 2012. He was awarded the Ph.D. degree in Robotics from King's College London, UK, in 2016. His research interests include tactile sensing, robot learning and robot visual-tactile perception.