Associated Learning Architecture Equipped with Parallel Impedance Sensing Strategy to Enhance Cross-modal Object Perception

Zhibin Li, Yuxuan Zhang, Weirong Dong, Jing Yang, Jiansong Feng, Chengbi Zhang, Xiaolong Chen, and Taihong Wang.

***Abstract*—Human beings can infer the shape and material characteristics of grasping objects based on multisensory information, which is still a technical challenge for modern robots. The cross-modal object perception mechanism holds promise to assist robots in effectively executing various operations or interactive tasks in complex applications, particularly in harsh visual scenes. Here, we present an associated learning architecture equipped with parallel impedance sensing strategy, which enhances the perception of captured objects by integrating visual data with somatosensory data from Frequency Division Multiplexing (FDM) parallel impedance and finger bending angles of the robotic hand. We design a Cross-modal Generative Adversarial Network (CGAN) within this architecture to achieve cross-modal feature learning for two types of sensory data, mimicking the psychological cognition of human senses. Additionally, the dynamic attention fusion mechanism is employed for feature transfer and fusion learning, enabling the network to adaptively adjust weights based on input cross-modal features, resulting in dynamic feature fusion. The architecture has undergone training and testing with ten categories of objects, successfully achieving cross-modal feature learning and fusion recognition of the two sensory data. Under low-quality image conditions, the recognition accuracy of attention fusion reaches up to 94.0%, significantly surpassing the accuracy of vision alone. This highlights the potential of our architecture to enhance robots to accurately perceive the outside world by integrating visual and somatosensory data, especially in challenging visual environments.**

***Index Terms*—Parallel impedance, Cross-modal perception, Multimodal fusion, Robotic perception, Associated learning**

I. Introduction

HUMAN perception of objects in the real world depends on the highly intertwined fusion of somatosensory and visual information in the neural network of the brain, and mental imagery is established between different sensory information [1]. Relying on the somatosensory feedback on the mechanoreceptors, people can infer the shape and material characteristics of grasping objects, so as to explore, learn and adapt to the world [2], [3]. To this end, novel somatosensory sensors in Internet of Things (IoT) scenarios require the ability to perceive material and morphological properties and integrate cross-modal learning, which facilitates robots with more accurate perception and interaction with surroundings [4]–[7].

Somatosensory sensors have been developed with a variety of mechanisms to facilitate cross-modal fusion of robotic systems, including piezoresistive array [8], capacitive pressure grid [9], piezoelectric sensor [10], magnetic microelectromechanical sensor (m-MEMS) [11], optoelectronic strain sensor [12], and triboelectric nanogenerator (TENG) [13]. However, such sensors must arrange dense matrices to obtain the contour and stiffness of the object, so that collecting a large number of sensor data poses a major challenge to system robustness. Moreover, they are limited to perceiving primarily mechanical pressure, thus lacking the capability to sense material characteristics.

Electrical impedance technology has been proposed to be used in electrochemistry, material science and biology, from material characterization of objects to detection of biological tissues [14]–[16]. It can detect valuable information about electrical transmission and charge separation at interface in voltammetry, which helps to determine the inherent properties of materials. To this end, the emerging impedance sensor has been extensively explored recently in the field of object recognition because of its low-cost, fast response, non-invasive.

Neto et al. [17] and Pietro et al. [18] have proposed the electrical impedance was used for the identification of the maturation degree of fruits based on variation of bulk resistance dependence with maturation of fruits. Vela et al. [19] have developed electrical impedance in IoT to identify the physiological state of biological tissues, organs,and fluids. Zhou et al. [20] have presented a 3D-printed impedance flow cytometer array for biological cell classification. Liu et al. [21] have studied the detection and imaging of objects by non-visual environmental impedance. Gong et al. [22] have adopted conductive fabric to measure the impedance of contact objects to identify daily necessities with a certain impedance range. Hence, the material dependence of impedance data on somatosensory perception cannot be ignored.

Cross-modal perception aims to establish multimodal association models of information to improve the overall accuracy of decision-making results [23]. Compared to unimodal approaches, it has the potential to provide more robust and accurate predictions by utilizing complementary information

This work was financially supported by Guangdong Major Talent Project (2019CX01X014, 2019QN01C177) and Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation. ('Climbing Program' Special Funds, pdjh2023c11001). (Corresponding author: Taihong Wang, wangth@sustech.edu.cn)

Zhibin Li (12231066@mail.sustech.edu.cn), Yuxuan Zhang (12012508@mail.sustech.edu.cn), Weirong Dong, Jing Yang, Jiansong Feng, Chengbi Zhang, Xiaolong Chen and Taihong Wang are with the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, Shenzhen 518077, China.

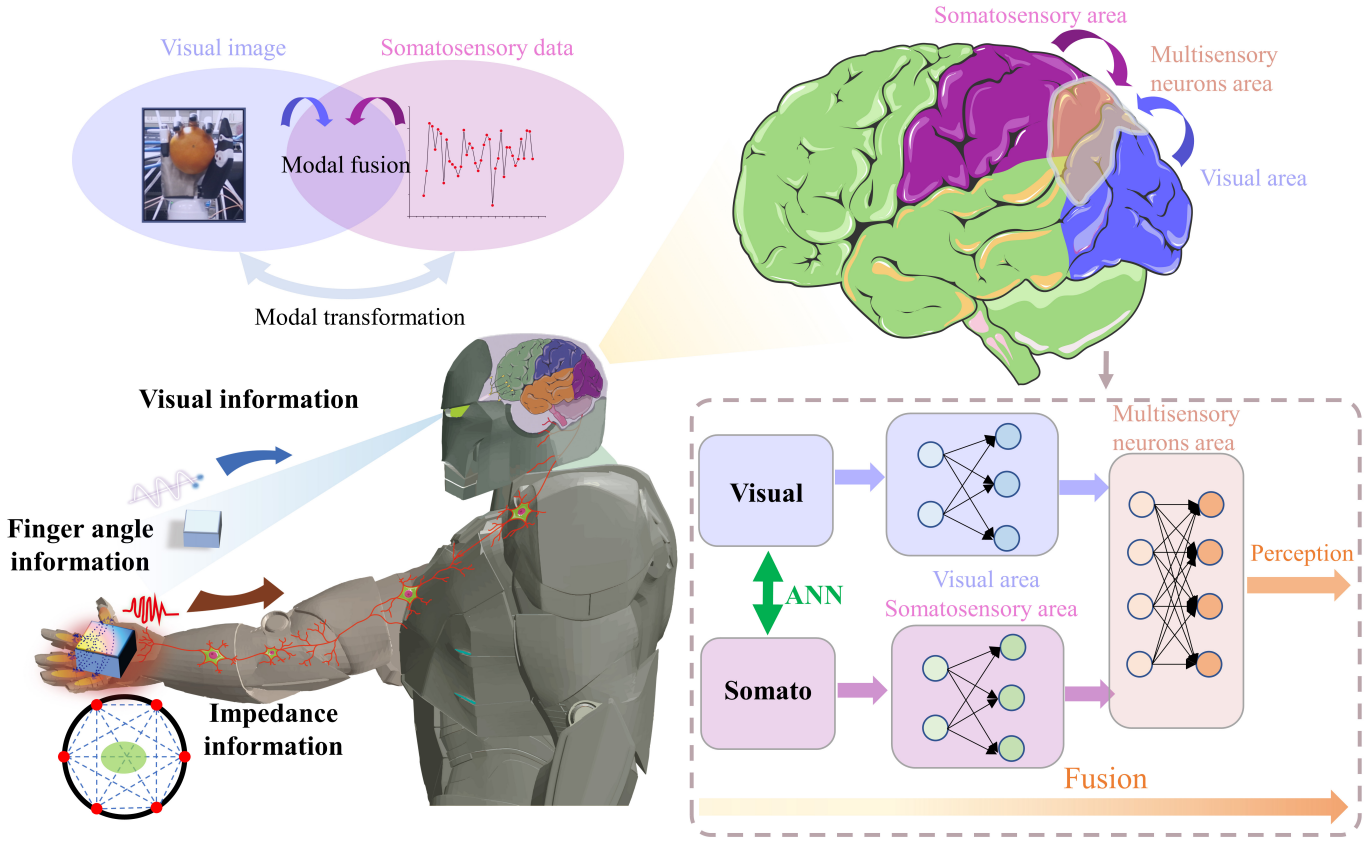


Fig. 1. Associated learning architecture equipped with parallel impedance sensing strategy. This framework integrates somatosensory and visual data through associative learning with artificial neural networks (ANN) to mimic the early interaction of different modalities (visual and somatosensory) in the multisensory neuron regions of the brain [6].

from different modalities. Current research focuses on various fusion strategies, including early fusion, late fusion, and hybrid fusion [24]. However, most of these multimodal methods are designed in a task-specific manner, which limits their performance on other tasks. Especially when the manual labeling of modality tags is costly or not involved, extracting and fusing multimodal features is challenging. Therefore, it is crucial to develop an adaptive cross-modal fusion strategy to help encoders capture shared semantic features across modalities.

In this article, we present an associated learning architecture equipped with parallel impedance sensing strategy (Fig. 1), integrating somatosensory and visual data through associated learning with the artificial neural network (ANN). To capture the object's material information, we use the lock-in amplifiers to build a fast impedance measurement system under multifrequency synchronous excitation. This system collects electrical impedance between the fingers when the manipulator grabs the object. Additionally, the linear motor on the manipulator provides feedback on the finger bending angle data, sensing the object's shape information. These two data serve as somatosensory information when the robot grasps the object. During object perception, the operator uses a camera to visualize and grasp it with manipulator, obtaining both visual and somatosensory information.

As shown in Fig. 1, through associated learning, ANN establishes cross-modal feature learning network and attention fusion learning network. Our core contributions are as follows.

(1) We propose to integrate parallel impedance as part of the robot's somatosensory information, effectively assisting vision in cross-modal object perception.

(2) To ensure real-time object impedance sensing, we design the impedance measurement unit based on Frequency Division Multiplexing (FDM) for manipulator object sensing, avoiding switching time loss associated with Time Division Multiplexing (TDM).

(3) The cross-modal learning ANN is designed with a Cross-modal Generative Adversarial Network (CGAN), achieving unsupervised cross-modal shared feature learning between somatosensory and visual information.

(4) By transferring the feature network of CGAN and employing a neural network based on the somatosensory-visual (SV) attention fusion hierarchical structure, the weights of cross-modal features are controlled, resulting in improved accuracy of object recognition, particularly under adverse visual conditions.

II. Somatosensory and Visual Sensing IoT System

We designed an SV sensing IoT system to collect material and morphological characteristics of the captured object as active tactile somatosensory information for the robot, along with RGB images for visual information. The IoT system consists of three main parts: impedance measuring unit, grasping

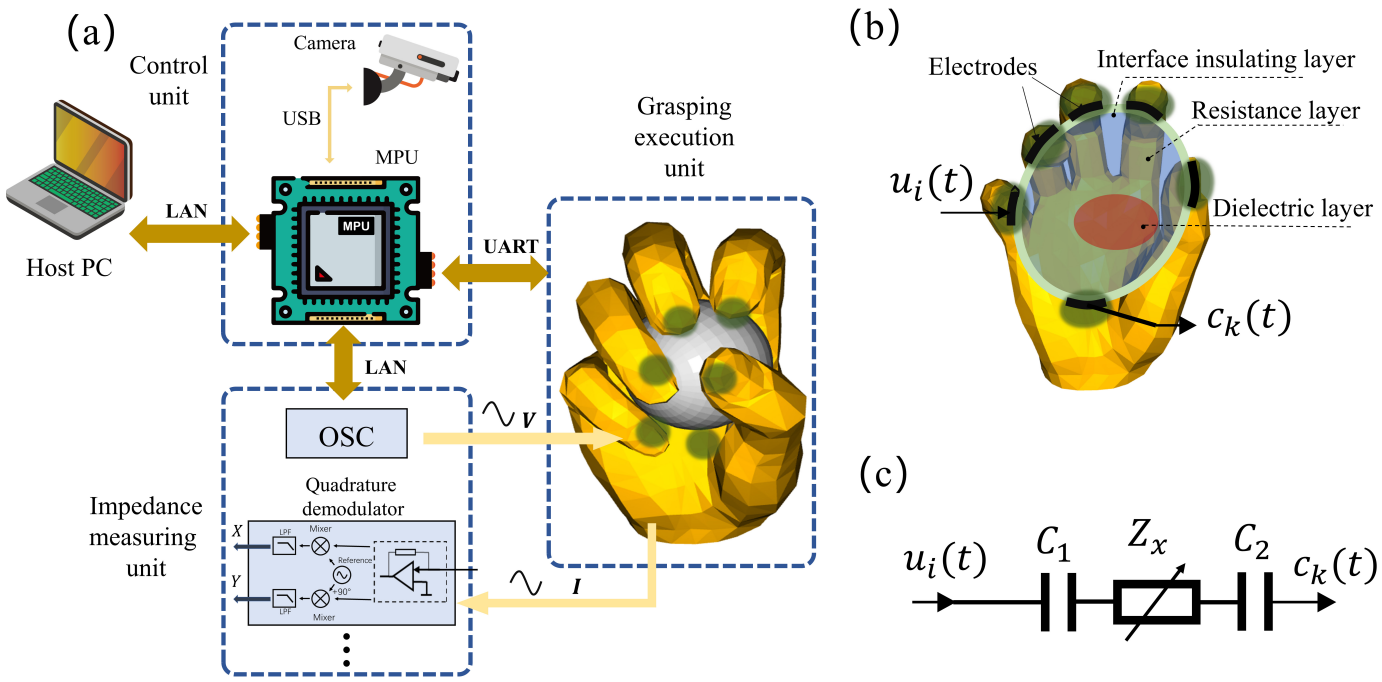


Fig. 2. Hardware system design equipped with parallel impedance sensing strategy. (a) Composition of somatosensory and visual sensing IoT system. (b) Measurement principle of parallel impedance for somatosensory perception. (c) Equivalent circuit of electrode pair.

execution unit, and control unit. The impedance measuring unit generates excitation signals and performs signal demodulation. The grasping execution unit includes a commercial dexterous manipulator (RH56DFX) used for object grasping and feedback of joint bending angle data. The control unit is responsible for manipulator grasping control and data processing. (Fig. 2(a))

A. Measurement Principle of Impedance

Electrical impedance technology has been proposed to be applied to the image reconstruction of electrical impedance distribution in the measured area [14], [21], [25]. This method can reconstruct the impedance distribution image of the object by injecting extremely small safe excitation current or voltage into the interior through the array electrode placed on the surface of the object, and measuring the voltage or current on the body surface, which would be used as an important reference for object recognition.

Fig. 2(b) illustrates the principle of six electrodes on the hand for object impedance measurement. In this work, a simplified hand grasping object model is adopted, which includes the contact interface between skin and object and the impedance distribution inside the object [25]. The structure of impedance sensing area, which mainly includes three layers: interface insulating layer, resistance layer and dielectric layer. Six electrodes are installed on the five fingers and palm of the manipulator. When the manipulator grabs the object, six electrodes would be connected to the surface of the target object. For any two electrodes, the interface insulating layer forms two coupling capacitors (C_1, C_2), and the inner resistance layer and dielectric layer can be equivalent to impedance Z_x. Therefore, the equivalent circuit of any electrode pair can be simplified as two coupling capacitors in series with the impedance Z_x (Fig. 2(c)).

The total impedance Z of an electrode pair is:

$$Z = Z_x + \frac{1}{j\omega C_n} = Z_x - j\frac{1}{2\pi f C_n} \tag{1}$$

While ω and f are respectively the angular frequency and frequency of the excitation signal. C_n is the series equivalent sum of two coupling capacitors C_1 and C_2 on the interface. Z_x is the self impedance of the target object.

Impedance measurement is divided into four-terminal method and two-terminal method [26]. In this work, we use AC voltage as the excitation source of the two-terminal method. The AC voltage source $u_i(t)$ is applied to the corresponding each electrode at different frequencies as frequency coding. At the same time, the electrode demodulates the mixed current signal $c_k(t)$, which reflects the impedance of the sensing area between the electrode pairs.

B. Parallel Impedance Measuring Unit

Conventional multi-channel impedance measurement typically relies on Time Division Multiplexing (TDM), where the impedance between two electrodes is sequentially measured in each cycle [27], [28]. Compared to mainstream switch-based circuits based on TDM, our designed impedance measurement unit based on frequency-division multiplexing (FDM) avoids

the switching time loss of multiplexers, ensuring the real-time impedance detection of objects [29], [30]. Excitation and measurement are carried out simultaneously on the electrode, resulting in the superposition of signals in the acquisition channel. To address this, quadrature demodulation based on lock-in amplifier is used to extract overlapping signals in FDM.

The electrodes are distributed on the five fingers and palm of the manipulator, which are composed of conductive fabrics. The voltage $u_i(t)$ with the one-to-one frequency is applied to the electrode i, and the expression is as follows:

$$u_i(t) = A_0 \exp\left(j\left(2\pi f_i t + \varphi_0\right)\right) \tag{2}$$

Where A_0 is the voltage amplitude (100 mV), φ_0 is the fixed initial phase, and t is the time. The frequency in our example is $(f_i = 90 + 10 \times i)$ kHz, the electrode number i, and $i = 1, 2, ..., 6$. The vector representation of the excitation voltage $V(t)$ is:

$$V(t) = \begin{bmatrix} u_1(t) & u_2(t) & \dots & u_6(t) \end{bmatrix} \tag{3}$$

Under the parallel action of excitation voltage $V(t)$. The superposition of all frequency currents $c_k(t)$ is measured at the current measuring electrode k.

$$c_k(t) = \sum_{i=1}^{6} B_{k,i} \exp\left(j\left(2\pi f_i t + \varphi'_{k,i}\right)\right) \tag{4}$$

where $B_{k,i}$ and $\varphi'_{k,i}$ represent the amplitude and phase of the signal with a frequency of f_i at the current measurement electrode k, $k = 1, 2, ..., 6$.

$$I(t) = \begin{bmatrix} c_1(t) & c_2(t) & \dots & c_6(t) \end{bmatrix} \tag{5}$$

Therefore, the impedance amplitude $|Z_{k,i}|$ between the two electrodes k and i can be expressed as

$$|Z_{k,i}| = \frac{A_0}{B_{k,i}} \tag{6}$$

$$\mathbf{R} = \begin{bmatrix} |Z_{1,1}| & |Z_{1,2}| & \cdots & |Z_{1,6}| \\ |Z_{2,1}| & |Z_{2,2}| & \cdots & |Z_{2,6}| \\ \vdots & \vdots & \ddots & \vdots \\ |Z_{6,1}| & |Z_{6,2}| & \cdots & |Z_{6,6}| \end{bmatrix} \tag{7}$$

where, $\mathbf{R} \in \mathbb{R}^{6\times6}$, A_0 as a known voltage excitation amplitude (100 mV). We need to extract $B_{k,i}$ from the measured mixed current $c_k(t)$ to characterize the object impedance between electrodes. We use phase sensitive detection based on lock-in amplifier to separate the signal of the desired frequency from all other frequency components. The design of impedance measurement system based on FDM is shown in Fig. 3(a). We chose the lock-in amplifier MFLI of Zurich instrument for test, which can produce high-frequency voltage excitation and current demodulation [31]. Six MFLI share a clock oscillator to achieve clock synchronization between the excitation voltage source and the demodulation reference signal. $R_{internal}$ indicates that the internal resistance of the circuit is less than 80 Ω and is ignored. The input current $c_k(t)$ is demodulated into in-phase component $X_{k,i}$ and orthogonal component $Y_{k,i}$ by a reference signal with frequency f_i [31].

$$\begin{aligned} X_{k,i} &= LPF(c_k(t)\cos\left(2\pi f_i t\right)) = B_{k,i}\cos\theta_{k,i}, \\ Y_{k,i} &= LPF(-c_k(t)\sin\left(2\pi f_i t\right)) = B_{k,i}\sin\theta_{k,i} \end{aligned} \tag{8}$$

where, LPF represents an IIR filter with a cutoff frequency of 150 Hz and the order of 10. The reference signal can be generated internally by the Lock-in amplifier, with frequency of f_i and phase of 0. $\theta_{k,i}$ is the phase difference between the input signal and the reference signal. According to formula (8), it can be concluded that:

$$B_{k,i} = \sqrt{{X_{k,i}}^2 + {Y_{k,i}}^2} \tag{9}$$

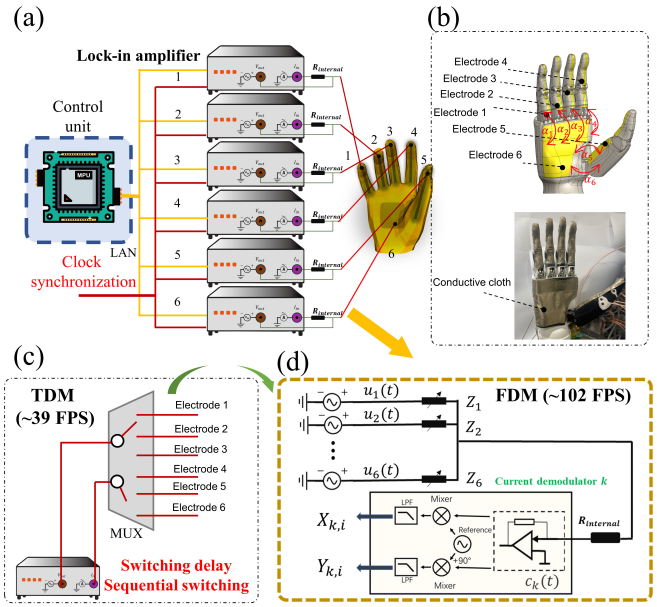


Fig. 3. Impedance measurement system based on FDM. (a) Hardware connection diagram using the parallel impedance sensing strategy. (b) Robotic hand model and electrode distribution diagram. (c) Measurement architecture diagram of the TDM strategy. (d) Measurement architecture diagram of the FDM strategy.

It can be obtained according to formula (6), (7), (9): $|Z_{k,i}|$ can be used to characterize the impedance between electrode k and electrode i. In order to ensure the integrity of the impedance data matrix, the measured line internal resistance data when $k = i$ is only used as filling and not removed. $\mathbf{R}$ as a set of spatial impedance perception.

C. Grasping Execution Unit

As shown on the right side in Fig. 3(b), in order to simulate the grasping action of real human hand, the commercial humanoid five finger dexterous hand is used as the actuator of grasping. The execution unit executes the grasping object command of the control unit and returns the bending angle α_m of the manipulator finger, and m = 1,2,...,6 (shown in Fig. 3(b)). The humanoid five finger dexterous hand adopts the linear motor drive design to grasp with a fixed force. It is responsible for simulating the hand grasping the object

and making the finger electrode contact with the object. The angle information of grasping the object is represented as $\mathbf{A} = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \alpha_6 \end{bmatrix}$.

The manipulator's five fingers and palms are equipped with flexible conductive cloth electrodes. The initial resistance of the conductive cloth is less than 0.5 Ω, and its conductivity remains largely unaffected by bending. This enables the cloth to bend with the fingers, ensuring good contact with the target object.

In conclusion, the grasping execution unit mimics the real hand's grasping action, making contact between the conductive electrode on the hand and the object. Simultaneously, we can obtain the bending angle $\mathbf{A}$ of the grasping finger as part of the hand's somatosensory information.

D. Control unit

The control unit utilizes the NVIDIA Jetson Xavier NX as the core microprocessor (MPU) to control the manipulator's grasping action and trigger the measurement system for data acquisition. The MPU communicates with the manipulator through UART, sending hand grasping commands and reading the returned finger angle simultaneously. Once the object is grasped, the MPU controls the phase-locked measurement unit via LAN to initiate impedance measurement. Compared to the sequential channels scanning under the TDM strategy (∼39 FPS, Fig. 3(c)), the FDM parallel impedance measurement we designed can achieve high-speed readout for multichannel impedance (∼102 FPS, Fig. 3(d)). Concurrently, the MPU captures a color image $\mathbf{P}$ from the camera through USB.

The collected data include impedance $\mathbf{R}$, finger angle $\mathbf{A}$ and color image $\mathbf{P}$ when grasped object, and label of object $\mathbf{L}$. Labeling data can be cumbersome, so most unlabeled data will be utilized for feature extraction under modality conversion, while a small portion of labeled data will be used for data fusion recognition.

III. Somatosensory-Visual Associated Learning Architecture

To validate the effectiveness of the proposed method, we utilize the integrated system described above for data collection. In this section, we introduce the datasets and elaborate on the implementation details of the associated learning architecture for visual and somatosensory conversion and fusion.

A. Data Collection and Preprocessing

To ensure the broad representativeness of the data in this study, we referenced the YCB object model dataset [32], which consists of everyday objects of various shapes, sizes, textures, weights, and stiffness. For representative benchmark testing of robot grasping and manipulation, we selected two objects from each category in the YCB model database for data collection.

We choose a set of ten objects (Fig. 4), including tennis, baseball, orange, apple, etc., and collect 350 sets of data for each object. The objects were placed in random postures during data collection, introducing variations that require the model to have higher generalization capability. Using the described hardware system, we synchronously collect visual and somatosensory data after grasping each object. We randomly sample each object 350 times, collecting one frame of data. The dataset comprises 200 unlabeled training data, 100 labeled training data, and 50 labeled testing data. All data are normalized, and Gaussian noise with a mean of 0 and a variance of 0.02 is added to the training set for data augmentation. This approach can enhance robustness against noise and uncertainty, improving performance on unseen data in real-world applications.

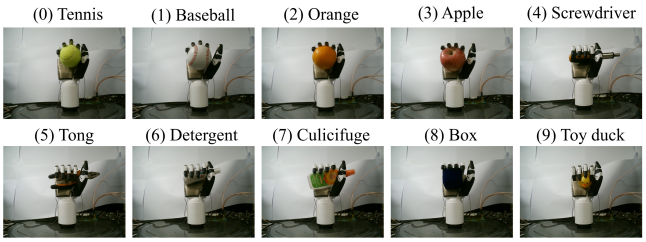


Fig. 4. 10 categories of objects (0-10) in SV dataset.

B. Preliminary Experiment

The perception of visual information is susceptible to the influence of adverse environmental conditions (such as image blurring, inadequate lighting, etc.) [5], [6]. Therefore, we conducted preliminary experiments to assess the susceptibility of individual visual data in object recognition under interference. We employed the widely used Resnet18 network for visual feature extraction. The experiments involved adding images with varying intensities of Gaussian noise interference to evaluate the recognition capability of images severely degraded under non-ideal conditions. As depicted in Fig. 5(a), when the image noise intensity increased to a variance greater than 0.04, the accuracy of image recognition dropped from the original 100% to below 80%. Additionally, we utilized t-Distributed Stochastic Neighbor Embedding (t-SNE) dimensionality reduction technique to visualize the image features [33]. As shown in Fig. 5(b), under the interference of noise with a variance of 0.04, there was a noticeable overlap in the distribution of features across different categories. These results indicate that under adverse interference, a single visual modality struggles to provide valuable clues for object recognition.

C. Somatosensory-Visual Cross-modal Feature Learning

Humans utilize cross-modal learning to connect and associate multimodal information in the high-level cortical areas of the brain, enabling cross-modal recognition and imagination [1]. Building on this cross-modal cognition, we propose a method that leverages somatosensory and visual data for cross-modal feature learning. The somatosensory information perceived by the hand can establish feature relationships with object visual images, achieving the extraction of cross-modal associated features of objects.

Here, we design a Cross-modal Generative Adversarial Network (CGAN) to achieve cross-modal feature learning between somatosensory and visual information, simulating

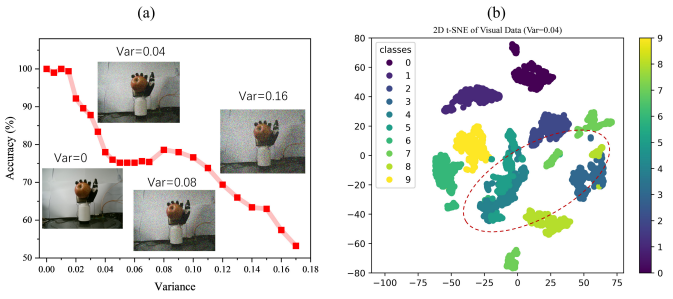


Fig. 5. Preliminary experimental results of object recognition under visual interference. (a) The accuracy of single visual recognition significantly decreases as the intensity of Gaussian noise interference increases. (b) Visualization of features of disturbed visual data using t-SNE dimensionality reduction.

human psychological perception. The feature extraction network based on CGAN is trained using unlabeled data for multimodal features extraction. The generator G takes visual or somatosensory data as conditional input x and produces target data in the other domain with $y = G(x)$. In the training of CGAN, the goal of the discriminator D is to reveal the difference between synthetic results and real data, while the generator G is trained to produce results that can deceive the discriminator D. We take the unlabeled somatosensory data and visual data $\{([\mathbf{R}, \mathbf{A}], \mathbf{P})\}$ in the previous section as our training data $\{(x, y)\}$. In the task of somatosensory $\rightarrow$ visual, x is the somatosensory data and y is the corresponding visual image. Similarly, in the visual $\rightarrow$ somatosensory task, i.e. $(x, y) =$ (visual image, somatosensory data).

WGAN-GP adds gradient penalty on the basis of Wasserstein Generative Adversarial Networks (WGAN) to help achieve optimal learning [34], [35], which can alleviate the problem of unstable training. We extend it to the task of cross-modal learning. The loss function of WGAN-GP used in this article is as follows:

$$\min_G \max_D L(D,G) = -\mathbb{E}_{(x,y)}[D(x,y)] + \mathbb{E}_{(x)}[D(x,G(x))] + \lambda_1 \mathbb{E}_{\hat{y}}\left[(\|\nabla_{\hat{y}}\|_2 - 1)^2\right] \tag{10}$$

Where $-\mathbb{E}_{(x,y)}[D(x,y)] + \mathbb{E}_{(x)}[D(x,G(x))]$ represents the Wasserstein distance between the distribution fitted by the generator and the real data distribution, $\lambda_1 \mathbb{E}_{\hat{y}}\left[(\|\nabla_{\hat{y}}\|_2 - 1)^2\right]$ represents the gradient penalty term, and λ_1 is a weighting constant. Compared with JS divergence and KL divergence, Wasserstein distance can still represent the distance between the two distributions without overlapping parts [34]. Meanwhile, the gradient penalty is proposed to replace the clipping weight, which penalizes the norm of gradient of the critic with respect to its input [35]. WGAN-GP can stably train various Gan architectures without super parameter adjustment, and solves the problems of Gan gradient instability, difficult training and insufficient diversity. We design the visual $\rightarrow$ somatosensory Gan loss $G^{v \rightarrow s}$.

$$G^{v \rightarrow s} = \min_G \max_D L(D,G) + \lambda^{'} \mathcal{L}_1(G) \tag{11}$$

$$\mathcal{L}_1(G) = \mathbb{E}_{(x,y)} \|y - G(x)\|_1 \tag{12}$$

We add a direct regression L_1 loss between the predicted results and the real data. This loss has been shown to help stabilize Gan training [36]. Where $\lambda^{'}$ is set to 2. Then, we designed the somatosensory $\rightarrow$ visual Gan loss $G^{s \rightarrow v}$.

$$G^{s \rightarrow v} = \min_G \max_D L(D,G) + \lambda^{''} \mathcal{L}_1(G) \tag{13}$$

Where $\lambda^{''}$ is set to 1000. The structure of our CGan model is shown in Fig. 6(a). In the s$\rightarrow$v network, we use a convolutional neural network (named S-CNN) as the feature extraction layer for somatosensory signals. The network structure is detailed in Table I. To preserve as much of the relatively low-dimensional somatosensory data's feature information as possible, pooling layers are not used. Transposed convolutional layers are utilized for upsampling to align the somatosensory data with the visual data [37]. In the v$\rightarrow$s network, ResNet18 is employed for visual feature extraction (named V-ResNet18). The residual block, consisting of cascaded convolution layers and a shortcut connection, enables image feature extraction similar to the function of local receptive fields in the biological nervous system [38], [39]. Subsequently, the fully connected network is responsible for generating low dimensional somatosensory information.

TABLE I
THE ARCHITECTURE OF THE S-CNN NETWORK

Layer	Patch size/Stride	Output Size	Activation
Conv1	3×3 /1	$8 \times 6 \times 7$	ReLU
Conv2	3×3 /1	$16 \times 6 \times 7$	ReLU
Conv3	3×3 /1	$32 \times 6 \times 7$	ReLU
Flatten	-	1344	-
Linear	-	512	-

We conducted experiments using Python 3.9 and the PyTorch framework on a system with 4 NVIDIA T4 Tensor Core GPUs. We use the Adam optimizer with a mini-batch size of 20. For the v$\rightarrow$s network, the learning rates for the generator and discriminator are set to 0.0001, and training is run for 100 epochs. For the s$\rightarrow$v network, the learning rates are 0.00005, with training over 500 epochs.

D. Somatosensory-Visual Attention Fusion Learning

Cross-modal fusion learning for object recognition has been proved to significantly improve accuracy, especially under limited image quality conditions [6]. However, the effectiveness of this method is limited by the low-quality sensor data, necessitating the use of more abundant and effective multimodal features for fusion learning. Therefore, we designed a parallel impedance system to collect the material and morphological features of the captured object as active somatosensory information for the robot. Additionally, we need to consider the structure of cross-modal fusion due to the mismatch between data dimension and data density.

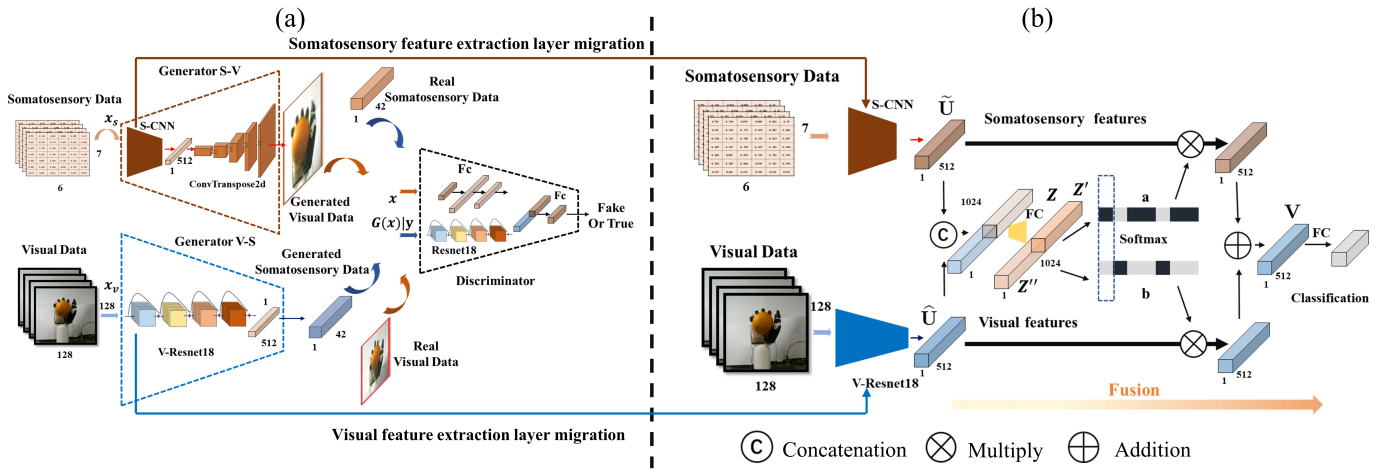


Fig. 6. Diagram of the associated learning architecture (a) The architecture of somatosensory-visual cross-modal learning based on Gan model. (b) The architecture of somatosensory-visual fusion learning based on attention model.

To demonstrate the feasibility and effectiveness of parallel impedance data and grasping angle as somatosensory data for object perception, we design an SV attention fusion network (named SV-AF) for cross-modal associated learning (Fig. 6(b)). The feature layers are trained through the CGAN in Fig. 6(a), which aims to extract visual and somatosensory features through establishing modal correlation. The feature extraction layers, V-Resnet18 and S-CNN, trained under CGAN, are transferred to the SV-AF network for feature extraction. At a higher level, the late fusion network is responsible for associating visual features with low-dimensional somatosensory information.

In the SV-AF network, we design neurons to adaptively adjust their weights based on modality information, controlling the flow of information from different modalities. The feature extraction layers are initially frozen to focus on training the feature fusion layer, and then unfrozen later. By combining and aggregating visual and somatosensory feature information, a globally integrated representation with selective weights is obtained [40].

In order to use gates to control the information flow of different modal feature information transferred to the next layer of neurons. We first fuse the features of the two modes(Fig. 6(b)) via an element-wise summation:

$$\mathbf{U} = \tilde{\mathbf{U}} + \widehat{\mathbf{U}} \tag{14}$$

Visual and somatosensory information is extracted by different feature extraction layers to obtain $\tilde{\mathbf{U}}$ and $\widehat{\mathbf{U}}$ features, and $\tilde{\mathbf{U}} \in \mathbb{R}^{512\times1}$, $\widehat{\mathbf{U}} \in \mathbb{R}^{512\times1}$. A simple FC layer implements the guidance for the precise and adaptive selections to achieve better efficiency.

A soft attention across channels is used to adaptively select different modal features, which is guided by the compact feature descriptor $\mathbf{Z} \in \mathbb{R}^{1024\times1}$. $\mathbf{Z}$ is truncated as $\mathbf{Z}^{'} \in \mathbb{R}^{512\times1}$, $\mathbf{Z}^{''} \in \mathbb{R}^{512\times1}$. Note that $\mathbf{Z}_n^{'}$ is the n-th row of $\mathbf{Z}^{'}$ and a_n is the n-th element of a, likewise $\mathbf{Z}_n^{''}$ and b_n. Specifically, a softmax operator is applied on the channel-wise digits [40]:

$$a_n = \frac{e^{\mathbf{Z}_n^{'}}}{e^{\mathbf{Z}_n^{'}} + e^{\mathbf{Z}_n^{''}}}, b_n = \frac{e^{\mathbf{Z}_n^{''}}}{e^{\mathbf{Z}_n^{'}} + e^{\mathbf{Z}_n^{''}}} \tag{15}$$

The final feature summary $\mathbf{V} = [\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_n]$ is obtained through the attention weight of each feature; $\mathbf{V} \in \mathbb{R}^{512\times1}$.

$$\mathbf{V}_n = a_n \cdot \tilde{\mathbf{U}}_n + b_n \cdot \widehat{\mathbf{U}}_n, \quad a_n + b_n = 1 \tag{16}$$

We conducted experiments using Python 3.9 and the PyTorch framework on a system with 4 NVIDIA T4 Tensor Core GPUs. We use Adam on 4 GPUs with a mini-batch size of 20 examples. During the training of the SV fusion network, the learning rate is set to 0.0001, and a total of 10 epochs are completed. The transferred feature layers are initially frozen to better strengthen the training of the backend fusion attention network. After 7 epochs, it is unfrozen for further optimization training.

IV. Result and Discussion

We utilize the aforementioned associated learning architecture to accomplish cross-modal perception tasks. We evaluate the associated learning architecture using the SV dataset described in Section 2. In this section, we will report several metrics for evaluating different aspects of cross-modal learning and fusion. The results demonstrate that utilizing parallel impedance and grasping angle as somatosensory information can enhance pattern recognition in computer vision under associated learning architecture.

A. Somatosensory-Visual Cross-modal Learning Assessment

When humans see and grasp an object, we can rely on two sensory modalities to infer its characteristics. This prompts us to evaluate the model's understanding of objects across multiple sensory interactions. In this experiment, we analyze the results of the somatosensory-visual cross-modal model and assess the feasibility for modal transformation.

Our test set comprises a complete collection of 10 object categories, with 50 data samples for each category. The test set is used to evaluate the quality of feature extraction and the accuracy of transformation for both modalities. In the CGAN of generating visual images from somatosensory data, qualitative comparisons are shown in Fig. 7. The Fréchet Inception Distance (FID) score aims to calculate the distance between the distribution of feature vectors extracted from real and generated images [41], [42]. We use this score to quantitatively assess the similarity between the distributions of real images and generated images across different object categories. A lower score indicates better image quality for the generated images. Most objects have achieved satisfactory image reconstruction. However, the image quality of the tennis ball and apple is relatively poor, with some of them having fuzzy shadows.

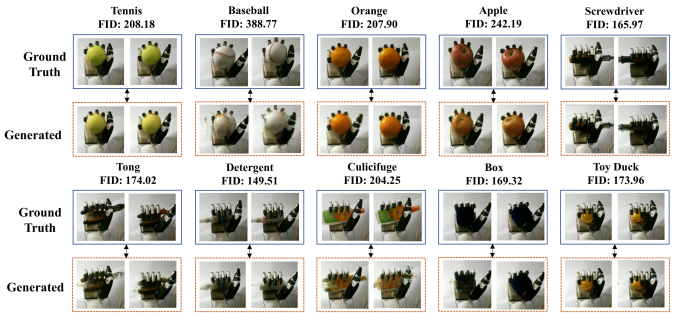


Fig. 7. Somatosensory→Visual cross-modal learning results.

Additionally, we evaluate the performance of generating somatosensory data from visual images on the test set. We utilize the R^2 value for quantification.

$$R_n^2 = 1 - \frac{\sum \left(\hat{y}_n^i - y_n^i\right)^2}{\sum \left(\bar{y}_n - y_n^i\right)^2} \tag{17}$$

$$R^2 = \frac{1}{N_c} \sum_{n=1}^{N_c} R_n^2 \tag{18}$$

Where y_n^i is the true value of the somatosensory sensor data, $\hat{y}_n^i$ is the corresponding predicted value, and $\bar{y}_n$ is the average sensor value. The sensor channel is represented by n and N_c denotes the total number of sensor channels, which is specified according to the size of the somatosensory data ($N_c = 42$). R^2 is used to measure the fitting prediction degree of the predicted value to the true value. The closer it is to 1, the better the fitting effect. The R^2 value for generating somatosensory data is 0.29, which confirms that the CGAN possesses a certain capability in generating somatosensory data. This indicates that the generation network for somatosensory data is capable of effectively extracting object perception features.

The primary objective of CGAN is to establish feature correspondence for object perception in modality conversion. Therefore, we need to verify the effectiveness of the feature extraction network in the generation network. We utilize the S-CNN and ResNet feature extraction networks trained under CGAN to extract features from visual and somatosensory data, and visualize the feature data using t-SNE. As shown in Fig. 8, the distribution of visual and somatosensory features exhibits good clustering effects. These results indicate that the feature extraction network trained under CGAN can provide valuable feature cues for object perception and can be transferred for use in feature fusion networks.

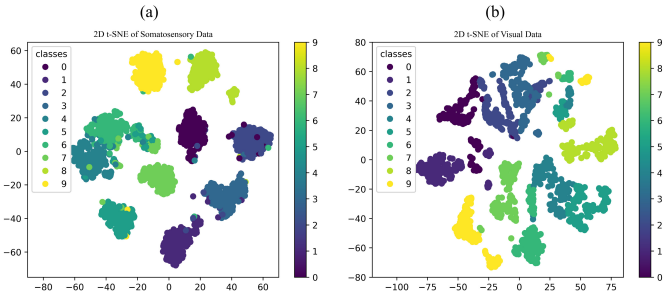


Fig. 8. Feature extraction under CGAN training. (a) Visualization of somatosensory features extracted by S-CNN. (b) Visualization of visual features extracted by V-ResNet18.

B. Somatosensory-Visual Attention Fusion Learning Assessment

When humans grasp and observe objects, we infer their categories through multimodal data fusion. Such multimodal fusion can overcome the impact of challenging environmental conditions (e.g., image blurring, insufficient illumination, etc.). Therefore, we aim to evaluate the contribution of somatosensory data in fusion recognition.

We evaluate the transfer contributions of the feature extraction networks V-ResNet18 and S-CNN and analyze their performance under image interference. The results of the ablation experiments are shown in Table II. The transfer application of V-ResNet18 and S-CNN can consistently improve recognition performance. V-ResNet18 increased the accuracy from 89.6% to 93.1%, while S-CNN brought a 2.2% improvement in accuracy. When both are used simultaneously, the recognition accuracy is further increased to 97.0%, demonstrating the significant importance of feature extraction network transfer in fusion perception. Additionally, even under image interference with noise variance of 0.04, a high recognition accuracy of 94.0% is maintained.

TABLE II
FEATURE EXTRACTION NETWORK TRANSFER ABLATION ANALYSIS

V-Resnet18 (Visual)	**S-CNN** (Somatosensory)	**Accuracy** (Var=0)	**Accuracy** (Var=0.04)
×	×	89.6$_{\pm 1.4}$%	87.3$_{\pm 1.8}$%
✓	×	93.1$_{\pm 0.7}$%	86.7$_{\pm 2.4}$%
×	✓	91.8$_{\pm 1.1}$ %	89.7$_{\pm 0.8}$%
✓	✓	97.0$_{\pm 0.4}$ %	94.0$_{\pm 1.0}$%

Note: The first and second columns indicate the transfer of V-ResNet18 and S-CNN features. The third and fourth columns show recognition accuracy after feature fusion. Var denotes the variance of the applied image noise.

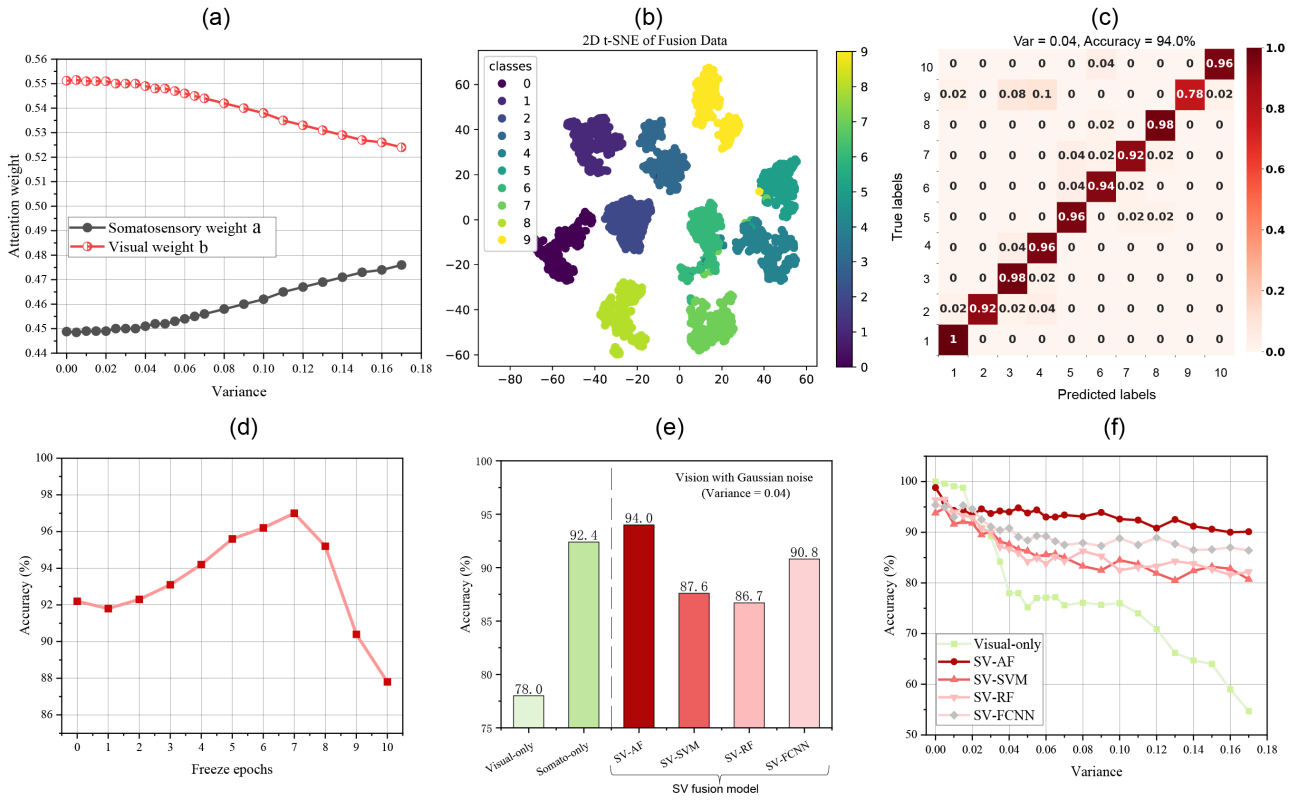


Fig. 9. The experimental results of somatosensory-visual attention fusion. (a) The variation of attention weights with the intensity of image noise. (b) Visualizing the fusion features of the SV-AF network using t-SNE dimensionality reduction. (c) Confusion matrix of recognition under attention fusion perception with images added with noise of variance 0.04. (d) The impact of transfer feature layer freezing epochs on accuracy. (e) The results of unimodal (visual or somatosensory) and multimodal fusion strategies are under vision with Gaussian noise with variance of 0.04. (f) Test results of different feature fusion strategies under visual information defects caused by Gaussian noise of varying intensities.

In the SV-AF network after feature transfer, we experimentally test whether the attention fusion model can adjust different modal weights based on the level of image noise. As shown in Fig. 9(a), as the image noise increases, the attention weight of somatosensory features significantly increases, while the attention weight of image features decreases. The experiment effectively demonstrates the robustness of attention fusion against image noise.

Additionally, we use t-SNE to visualize the dimensionality reduction of features from both modalities after attention fusion. As shown in Fig. 9(b), the SV-AF network effectively integrates features from both modalities, resulting in distinct clustering of extracted features. These results indicate that attention fusion significantly enhances object recognition performance. Fig. 9(c) shows the comparison between predicted results and true labels under attention fusion perception, providing the classification accuracy confusion matrix of the test dataset. It is evident that under the cross-modal fusion of the SV-AF network, object recognition maintains an accuracy of over 94.0% even when the images are affected by noise interference. Class 9 objects, which are foam boxes, have the lowest recognition accuracy. Due to the diverse grasping postures, dark colors, and high impedance of these objects, feature overlap easily occurs.

Moreover, we include ablation experiments to investigate the impact of different cross-modal data on recognition performance. As shown in Table III, incorporating finger bend angle data improved object recognition accuracy from 78.0% to 85.1% under low-quality imaging conditions. Furthermore, with the addition of parallel impedance data, object recognition accuracy increased by an additional 8.9%.

We also conduct experiments to train the transfer feature layer under different freeze epochs to determine the appropriate setting for the freeze epochs of the transfer layer. Fig. 9(d) demonstrates that, with a total of 10 epochs of training, freezing the transfer feature layer for 7 epochs before resuming training achieves the highest accuracy of 97.0%.

TABLE III
THE INFLUENCE OF SOMATOSENSORY MODAL DATA ON OBJECT RECOGNITION

Visual	**Somatosensory** (Finger bending)	**Somatosensory** (Parallel impedance)	**Accuracy** (Var=0.04)
✓	✗	✗	$78.0_{\pm 2.1}\%$
✓	✓	✗	$85.1_{\pm 2.4}\%$
✓	✓	✓	$94.0_{\pm 1.0}\%$

Note: The first to third columns represent data from different modalities. The fourth column shows the recognition accuracy under modality fusion.

We compare our fusion model with the most commonly used machine learning feature fusion recognition models, such as Support Vector Machine (SVM), Random Forest

(RF), and neural network models like Fully Connected Neural Network (FCNN). After extracting features from the same feature extraction layer, the features from both modalities were concatenated as input to the decision model. Compared to unimodal recognition and multimodal fusion recognition of other decision models, the fusion recognition accuracy of the SV-AF network is significantly higher under low image quality conditions (Fig. 9(e)). Notably, the SV-AF network demonstrates stronger resistance to image noise interference. As depicted in Fig. 9(f), when images are disturbed, the recognition accuracy of the single visual modality significantly decreases. However, the SV-AF network maintains high recognition accuracy even as image noise intensity increases, significantly outperforming other fusion decision models.

TABLE IV
PERFORMANCE COMPARISON OF DIFFERENT CROSS MODAL FEATURE NETWORKS UNDER LOW IMAGE QUALITY.

Feature network (Visual-Somato)	**Precision**	**Recall**	**F1-score**	**Accuracy**	**Params**
MobileNetV2-MobileNetV2	93.6%	92.8%	92.6%	92.8%	9.7 M
MobileViT-MobileViT	92.5%	91.2%	90.7%	91.2%	6.5 M
MobileViT-MobileNetV2	92.4%	91.4%	91.1%	91.4%	8.1 M
Ours	94.1 %	94.0%	94.0 %	94.0%	13.1 M

Note: Retain the cross-modal learning architecture but substitute the modal feature extraction with different lightweight networks, such as MobileNetV2 [43] and MobileViT [44].

Table IV summarizes the multimodal recognition performance of lightweight feature extraction networks under low image quality conditions. Compared to MobileNetV2 [43] and MobileViT [44], the proposed method achieves an improvement of over 1.2% in recognition accuracy, further validating the higher robustness and superior adaptability of our approach in real-world scenarios. Although our model has slightly larger parameters, we believe the current model efficiency is acceptable, and improving this will be a key focus for our future work.

These results indicate that equipping robots with somatosensory information obtained from parallel impedance and grasp angle systems can significantly enhance their object perception capabilities in complex environments. With the assistance of the associated learning architecture, the fusion of visual and somatosensory data features enables robots to accurately perceive the external world, particularly in challenging visual environments. Furthermore, it enables intelligent robots to comprehend and recognize objects across different modes.

V. CONCLUSION

We have reported an associated learning architecture equipped with parallel impedance sensing strategy, which integrates visual and somatosensory information to achieve cross-modal feature transfer learning and fusion. The designed parallel impedance and finger bending angle measurement system provides reliable somatosensory data of the manipulator, while the camera captures visual images. A fast impedance measurement under FDM synchronous excitation is established using a phase-locked amplifier to quickly capture the impedance characteristics of the object. The SV cross-modal feature learning utilizes the CGAN framework to achieve mutual transformation of two types of sensory data, enabling the associated learning of object modal features. The SV attention fusion learning adopts a dynamic attention fusion mechanism, which can adaptively adjust weights based on the different modal feature outputs from the feature extraction network transferred from the CGAN framework, enabling dynamic feature fusion. Feature learning and fusion recognition of two sensory data were achieved on a dataset consisting of 10 categories of everyday items. Under non-ideal conditions with noisy images containing Gaussian noise, the recognition accuracy of SV attention fusion surpassed that of vision-only recognition significantly. Our work demonstrates that the associated learning architecture with SV information can enhance robots' perception of the external complex multimodal world, particularly in challenging visual environments.

ACKNOWLEDGMENT

These authors contributed equally: Zhibin Li, Yuxuan Zhang. The authors would like to thank all the members of Intelligent Sensor Lab at Southern University of Science and Technology for their help during experiments.

REFERENCES

[1] H. Tan, Y. Zhou, Q. Tao, J. Rosen, and S. van Dijken, "Bioinspired multisensory neural network with crossmodal integration and recognition," *Nature communications*, vol. 12, no. 1, pp. 1–9, 2021.
[2] A. R. Sobinov and S. J. Bensmaia, "The neural mechanisms of manual dexterity," *Nature Reviews Neuroscience*, vol. 22, no. 12, pp. 741–757, 2021.
[3] B. He, Q. Miao, Y. Zhou, Z. Wang, G. Li, and S. Xu, "Review of bioinspired vision-tactile fusion perception (vtfp): From humans to humanoids," *IEEE Transactions on Medical Robotics and Bionics*, 2022.
[4] C. Bartolozzi, L. Natale, F. Nori, and G. Metta, "Robots with a sense of touch," *Nature materials*, vol. 15, no. 9, pp. 921–925, 2016.
[5] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10609–10618, 2019.
[6] M. Wang, Z. Yan, T. Wang, P. Cai, S. Gao, Y. Zeng, C. Wan, H. Wang, L. Pan, J. Yu, *et al.*, "Gesture recognition using a bioinspired learning architecture that integrates visual data with somatosensory data from stretchable sensors," *Nature Electronics*, vol. 3, no. 9, pp. 563–570, 2020.
[7] Z. Lu and H. Yu, "Gtac-hand: A robotic hand with integrated tactile sensing and extrinsic contact sensing capabilities," *IEEE/ASME Transactions on Mechatronics*, 2023.
[8] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, "Learning the signatures of the human grasp using a scalable tactile glove," *Nature*, vol. 569, no. 7758, pp. 698–702, 2019.
[9] Z. Li, J. Yang, Y. Zhang, P. Geng, J. Feng, B. Chen, X. Zhang, G. Yuan, X. Chen, and T. Wang, "Ultrafast readout, crosstalk suppression iontronic array enabled by frequency-coding architecture," *npj Flexible Electronics*, vol. 8, no. 1, p. 9, 2024.
[10] W. Lin, B. Wang, G. Peng, Y. Shan, H. Hu, and Z. Yang, "Skin-inspired piezoelectric tactile sensor array with crosstalk-free row+ column electrodes for spatiotemporally distinguishing diverse stimuli," *Advanced Science*, vol. 8, no. 3, p. 2002817, 2021.
[11] J. Ge, X. Wang, M. Drack, O. Volkov, M. Liang, G. S. Cañón Bermúdez, R. Illing, C. Wang, S. Zhou, J. Fassbender, *et al.*, "A bimodal soft electronic skin for tactile and touchless interaction in real time," *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.
[12] H. Zhao, K. O'Brien, S. Li, and R. F. Shepherd, "Optoelectronically innervated soft prosthetic hand via stretchable optical waveguides," *Science robotics*, vol. 1, no. 1, p. eaai7529, 2016.

[13] Z. Song, J. Yin, Z. Wang, C. Lu, Z. Yang, Z. Zhao, Z. Lin, J. Wang, C. Wu, J. Cheng, *et al.*, "A flexible triboelectric tactile sensor for simultaneous material and texture recognition," *Nano Energy*, vol. 93, p. 106798, 2022.

[14] E. Ravagli, S. Mastitskaya, N. Thompson, F. Iacoviello, P. R. Shearing, J. Perkins, A. V. Gourine, K. Aristovich, and D. Holder, "Imaging fascicular organization of rat sciatic nerves with fast neural electrical impedance tomography," *Nature communications*, vol. 11, no. 1, pp. 1–10, 2020.

[15] K. Park, H. Yuk, M. Yang, J. Cho, H. Lee, and J. Kim, "A biomimetic elastomeric robot skin using electrical impedance and acoustic tomography for tactile sensing," *Science Robotics*, vol. 7, no. 67, p. eabm7187, 2022.

[16] Q. Hua, Y. Li, M. W. Frost, S. Kold, O. Rahbek, and M. Shen, "Machine learning-assisted equivalent circuit characterization for electrical impedance spectroscopy measurements of bone fractures," *IEEE Transactions on Instrumentation and Measurement*, 2024.

[17] A. F. Neto, N. C. Olivier, E. R. Cordeiro, and H. P. de Oliveira, "Determination of mango ripening degree by electrical impedance spectroscopy," *Computers and Electronics in Agriculture*, vol. 143, pp. 222–226, 2017.

[18] P. Ibba, C. Tronstad, R. Moscetti, T. Mimmo, G. Cantarella, L. Petti, Ø. G. Martinsen, S. Cesco, and P. Lugli, "Supervised binary classification methods for strawberry ripeness discrimination from bioimpedance data," *Scientific reports*, vol. 11, no. 1, pp. 1–13, 2021.

[19] L. M. Vela, H. Kwon, S. B. Rutkove, and B. Sanchez, "Standalone iot bioimpedance device supporting real-time online data access," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 9545–9554, 2019.

[20] C. Zhou, M. Chen, D. Tang, and Y. Han, "A 3d-printed electrical impedance flow cytometer array for parallel detection of cellular biomarkers," in *2021 IEEE 34th International Conference on Micro Electro Mechanical Systems (MEMS)*, pp. 490–493, IEEE, 2021.

[21] D. Liu, D. Smyl, D. Gu, and J. Du, "Shape-driven difference electrical impedance tomography," *IEEE Transactions on Medical Imaging*, vol. 39, no. 12, pp. 3801–3812, 2020.

[22] H. Gong, Z. Cui, Y. Wang, C. Shen, D. Zhang, and S. Luo, "eglove: Designing interactive fabric sensor for enhancing contact-based interactions," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021.

[23] K. Hu, Z. Wang, K. A. E. Martens, M. Hagenbuchner, M. Bennamoun, A. C. Tsoi, and S. J. Lewis, "Graph fusion network-based multimodal learning for freezing of gait detection," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 3, pp. 1588–1600, 2021.

[24] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *Ieee Access*, vol. 7, pp. 63373–63394, 2019.

[25] Y. Jiang and M. Soleimani, "Capacitively coupled electrical impedance tomography for brain imaging," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2104–2113, 2019.

[26] J. Chen, L. Xu, Z. Cao, and H. Zhou, "Four-terminal imaging using a two-terminal electrical impedance tomography system," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 2, pp. 432–440, 2013.

[27] D. Yang, G. Huang, B. Xu, X. Wang, Z. Wang, and Z. Wei, "A dsp-based eit system with adaptive boundary voltage acquisition," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 5743–5754, 2022.

[28] H. Chen, X. Yang, P. Wang, J. Geng, G. Ma, and X. Wang, "A large-area flexible tactile sensor for multi-touch and force detection using electrical impedance tomography," *IEEE Sensors Journal*, vol. 22, no. 7, pp. 7119–7129, 2022.

[29] Y. Granot, A. Ivorra, and B. Rubinsky, "Frequency-division multiplexing for electrical impedance tomography in biomedical applications," *International journal of biomedical imaging*, vol. 2007, 2007.

[30] S. Sun, L. Xu, Z. Cao, W. Yang, *et al.*, "Signal demodulation methods for electrical tomography: A review," *IEEE Sensors Journal*, vol. 19, no. 20, pp. 9026–9035, 2019.

[31] "500 khz / 5 mhz lock-in amplifier." [Online], April 2022. https://www.zhinst.cn/china/cn/products/mfli-lock-amplifier.

[32] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols," *arXiv preprint arXiv:1502.03143*, 2015.

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[34] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, pp. 214–223, PMLR, 2017.

[35] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[36] 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*, pp. 1125–1134, 2017.

[37] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[39] D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. J. DiCarlo, "Performance-optimized hierarchical models predict neural responses in higher visual cortex," *Proceedings of the national academy of sciences*, vol. 111, no. 23, pp. 8619–8624, 2014.

[40] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 510–519, 2019.

[41] 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," *Advances in neural information processing systems*, vol. 30, 2017.

[42] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[43] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

[44] S. Mehta and M. Rastegari, "Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer," *arXiv preprint arXiv:2110.02178*, 2021.

Zhibin Li is currently working toward the Ph.D. degree at Southern University of Science and Technology, China. His research interests include deep learning, integrated sensor systems, and intelligent human-computer interaction.

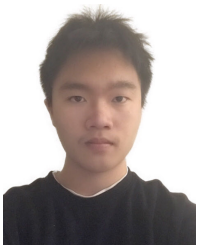
He won the Third Prize in the 2018 National Undergraduate Electronic Design Contest. In 2018, he was designated as a Meritorious Winner at the Interdisciplinary Contest in Modeling. He won the First Prize in the 2019 National Undergraduate Electronic Design Contest. In 2020, he won the Second Prize in the National Post-Graduate Mathematical Contest in Modeling.

Yuxuan Zhang is an undergraduate student at the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, China. His current research interests are intelligent perception and sensing signals analysis.

Weirong Dong is an undergraduate student at the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, China. His research areas are health monitoring and analog circuit design.

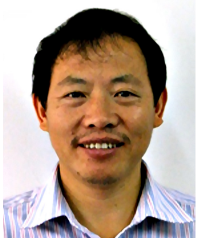
Jing Yang received her bachelor's degree from the college of Chemistry and Chemical Engineering, Hunan University in 2016 and master's degree from Pen-Tung Sah Institute of Micro-Nano Science and Technology of Xiamen University in 2019. Her current research interests involve the development and application of tactile sensors.

Jiansong Feng is currently working toward the Ph.D. degree at Southern University of Science and Technology, China. His current research interests involve the design and fabrication of microsensor.



Chengbi Zhang is an undergraduate student at the Department of Electronic and Electrical Engineering, Southern University of Science and Technology, China. His research areas are tactile perception and analog circuit design.

Xiaolong Chen received the Ph.D. in Hong Kong University of Science and Technology in 2014. He is currently an Assistant Professor at the Southern University of Science and Technology, Shenzhen, China. His current research interests involve the novel electronic and photonic devices.



Taihong Wang received the Ph.D. in Institute of physics, Chinese Academy of Sciences, China, in 1993. He is currently a Chair Professor at the Southern University of Science and Technology, Shenzhen, China. His area of expertise is Sensing and Intelligent Perception.

He was selected as the "Highly Cited Researchers" of Clarivate in 2018. In 1998, he was selected into the "Hundred Talents Program" of the Chinese Academy of Sciences. In 1999, he was funded by the National Science Fund for Distinguished Young Scholars. In 2005, he was hired as the Yangtze River scholar Professor. In 2006, he was awarded Special Subsidies of the Government of the State Council. In 2007, he was hired as Chief Scientist of National 973 Project. In 2008, he won the First Prize of Natural Science Award of Ministry of Education.