IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Self-Attention Based Visual-Tactile Fusion Learning for Predicting Grasp Outcomes

Shaowei Cui[1], Rui Wang[2], Junhang Wei[1], Jingyi Hu[1], and Shuo Wang[3]

***Abstract*— Predicting whether a particular grasp will succeed is critical to performing stable grasping and manipulating tasks. Robots need to combine vision and touch as humans do to accomplish this prediction. The primary problem to be solved in this process is how to learn effective visual-tactile fusion features. In this paper, we propose a novel Visual-Tactile Fusion learning method based on the Self-Attention mechanism (VTFSA) to address this problem. We compare the proposed method with the traditional methods on two public multimodal grasping datasets, and the experimental results show that the VTFSA model outperforms traditional methods by a margin of 5+% and 7+%. Furthermore, visualization analysis indicates that the VTFSA model can further capture some position-related visual-tactile fusion features that are beneficial to this task and is more robust than traditional methods.**

I. INTRODUCTION

The study of dexterous grasping and manipulation skills of robots has attracted increasing attention in the robot community [1], [2]. Predicting grasp outcomes before lifting is critical to achieving a stable grasp [3], which can help the robots formulate re-grasping strategies. [4]. Many studies have proposed various methods to accomplish this task [5], [6], [7]. Early studies focus on physical modeling of the grasping objects, grippers, and environment, while recent studies are more inclined to build predictive models in a data-driven and supervised manner, typically using visual or depth observations [8], [9]. Some more recent studies indicate that the addition of tactile information can significantly improve the prediction accuracy of this task [4], [5].

Humans make extensive use of multi-modal perception when grasping, including visual, tactile, and other modalities sensing [10]. Robots also need to perceive both visual and tactile modalities to evaluate the results of the current grasp. Intuitively, vision provides the grasp poses, and touch perceives the detailed contact information, both of which are helpful in determining whether the grasp will succeed. Some previous studies [5] has tried to tackle this problem by Visual-Tactile Fusion Perception (VTFP). These studies demonstrate that VTFP achieves better performance than single modal perception. In this paper, we focus on predicting grasp outcomes using VTFP, as shown in Fig. 1.

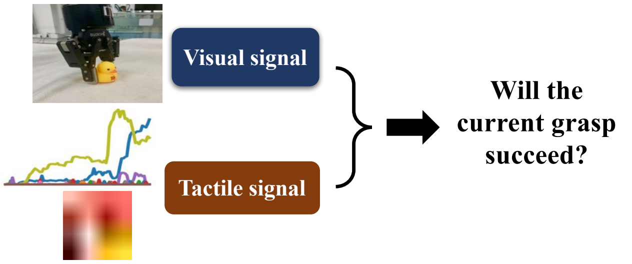

Fig. 1: A scenario framework for predicting grasp outcomes using visual-tactile fusion perception.

VTFP has long been used for a variety of tasks, such as surface classification [11], object recognition [12], object 3D shape perception [13]. Especially for robotic grasping and manipulating tasks, tactile perception can provide direct contact information, which can be used to evaluate grasping state and make timely adjustments. However, the most existing VTFP methods concatenate features from visual modality and tactile modality directly (e.g. [14], [4]), which we called Direct-Fusion method (DF). Unfortunately, this method may not deeply explore the modeling work of the information interaction of visual and tactile modality [15].

In this study, we focus on how to learn effective Visual-Tactile Fusion (VTF) features in the task of predicting grasp outcomes. Inspired by the remarkable success of the Self-Attention mechanism (SA) in the machine translation task [16], we propose a VTF representation learning method based on SA mechanism (VTFSA) to accomplish this task. Additionally, we validate the performance of different visual-tactile fusion method on two public grasping datasets. The experimental results indicate that the proposed method achieves better performance than traditional methods, and it can also address tactile signals captured by different tactile sensors, which extends its application scenarios.

Our primary contributions are three-fold:

- A novel visual-tactile fusion learning method is proposed to achieve better performance of the grasp out-

This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103003, in part by the National Natural Science Foundation of China under 61773378, in part by the Opening Project of Guangdong Provincial Key Lab of Robotics and Intelligent System, and in part by the Youth Innovation Promotion Association CAS.

[1] Shaowei Cui, Junhang Wei, and Jingyi Hu are with with School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China, and with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China `cuishaowei2017@ia.ac.cn`

[2] R. Wang is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China, and with Guangdong Provincial Key Lab of Robotics and Intelligent System, Shenzhen Institutes of Advanced Technology Chinese Academy of Sciences, Shenzhen, 518055, China

[3] S. Wang is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and with the University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Excellence in Brain Science and Intelligence Technology Chinese Academy of Sciences, Shanghai 200031, China (corresponding author, `shuo.wang@ia.ac.cn`)

comes prediction task.

- The experimental results demonstrate that the proposed model has better performance than the traditional methods in the grasp outcomes prediction task and can address different forms of tactile signals.
- Visualization analysis indicates that the VTFSA model can further capture some position-related visual-tactile fusion features that are beneficial to this task and is more robust compared to traditional methods.

II. Related work

A. Learning to Grasp

With the rapid development of representation learning methods, the data-driven approach [2], [17] has gradually replaced the traditional grasping strategies based on physical properties such as object shape, surface material, environment, and gripper characteristics [18]. Saxena *et al.* [9] present a learning algorithm for identifying grasp locations from an image, which is trained via supervised learning. Pinto *et al.* [7] train a Convolutional Neural Network (CNN) for predicting grasp locations without severe overfitting by collecting a large dataset. Levine *et al.* [2] propose a learning-based approach to hand-eye coordination for robotic grasping from monocular images. They train a CNN to predict that will the gripper's task-space motion result in successful grasps. Aktas *et al.* [19] provide a complete solution for deep dexterous grasping of novel objects from a single view. Most of the above studies only perceive visual and depth information while ignoring tactile sensing. Therefore, these methods may have a limited ability to reason about the details of the contact location.

H. Dang *et al.* [20] use tactile feedback to predict the stability of a robotic grasp without visual or geometric information about the grasped object. Murali *et al.* [21] propose a tactile-sensing based approach to grasp novel objects without prior knowledge of their location or physical properties. However, this study focus on object localization by tactile sensing.

B. VTFP in Grasping and Manipulation

Calandra *et al.* [4] investigate whether touch sensing aids in predicting grasp outcomes within a multimodal sensing framework that combines vision and touch. The experimental results indicate that incorporating tactile readings improves grasping performance significantly. They also propose an end-to-end action-conditional model that learns re-grasp policies from visual-tactile data [5]. This method has been proved by experiments to be able to obtain useful and interpretable re-grasp behaviors. Li *et al.* [22] propose a new method based on a deep neural network to detect slip. The visual features and tactile features are extracted by Pretrain CNNs, then are combined directly into fusion features, and finally fed into an LSTM network. Cui *et al.* [23] propose a 3D CNNs based visual-tactile fusion network to assess grasp states of deformable objects.

Fazli *et al.* [24] propose a method to emulate hierarchical reasoning and multi-sensory fusion in a robot that learns to play Jenga, a complex game that requires physical interaction to be played expertly. This model captures latent descriptive structures, and the robot learns probabilistic models of these relationships in force and visual domains through a short exploration phase. Lee *et al.* [25] use self-supervision to learn a compact and multimodal representation of many sensory inputs, which can be used to improve the sample efficiency of policy learning.

These amazing studies illustrate the importance of VTFP in grasping and complex manipulation tasks. Unfortunately, most of the existing deep-learning-based VTFP methods combine the feature of visual and tactile modalities directly and then perform subsequent classification or regression tasks. However, this Direct-Fusion (DF) may not be able to capture effective fusion features because of the simple structure.

C. Self-attention Mechanism

Attention mechanism has recently proven to be a powerful technique in deep-learning models and has been widely used in various tasks such as natural language processing [26]. Self-attention [16] is proposed to attend a word to all other words for learning relations in the input sequence. It significantly improves the performance of machine translation. Furthermore, it has also been introduced in videos to capture long-term dependencies across temporal frames among visual, text, audio modalities [27]. We deeply absorb the idea of global attention and propose a Self-Attention-based Visual-Tactile Fusion learning method (VTFSA), which transforms the temporal context correlation in the original sequence tasks to the relationship reasoning between different modalities and spatial positions in VTFP tasks.

III. Problem statement

Our goal is to learn effective VTF features for a grasp outcomes prediction task. Given the visual (X_V) and tactile (X_T) signals, we first extract visual (M_V) and tactile (M_T) features by visual (E_V) and tactile (E_T) encoder functions.

$$\begin{aligned} M_V &= E_v(X_V) \\ M_T &= E_t(X_T) \end{aligned} \tag{1}$$

Secondly, the VTF feature $M_{V,T}$ is learned from the visual and tactile features by a fusion function (F_F).

$$M_{V,T} = F_F(M_V,\, M_T) \tag{2}$$

Finally, $M_{V,T}$ is input into a classification function F_C to predict the current grasping results y.

$$y = F_C(M_{V,T}) \tag{3}$$

where $y \in \ 0 \ or \ 1$, and 0 denotes the grasp result is failed, while 1 means successful. Thus, the grasping results prediction task is defined as a two-categories classification problem.

In this paper, the visual encoder E_V, tactile encoder E_T, fusion function F_F, and classification function F_C are implemented by neural networks with different structures and parameters, respectively. The detailed network architecture of these networks are described in Sec. IV.

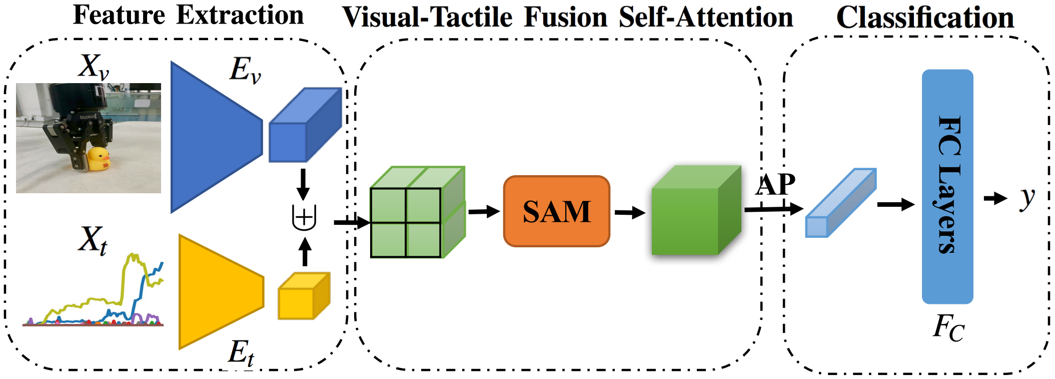

Fig. 2: An overview of our prediction model. The proposed model consists of three components including a feature extraction module, a VTFSA module, and a classification module. The network takes visual and tactile signals as inputs and encodes them into visual and tactile features. The final VTF feature is extracted through the VTFSA module. Finally, the final VTF feature is fed into the classification module to obtain the final prediction. AP means average-pooling operation.

IV. Methods

The overall architecture of our prediction model is shown in Fig. 2. The proposed model consists of three components including a feature extraction module, a visual-tactile fusion self-attention module, and a classification module. Given an image X_v captured by a side camera and a tactile signal X_t (e.g. tactile image, time serial, etc.), the output of our proposed model is whether the current grasp will succeed (y).

A. The Feature Extraction Module

The visual feature M_V is generally extracted by CNNs. While for different tactile sensors, we will use different neural networks to extract tactile feature M_T. For example, for the tactile images acquired by the vision-based tactile sensor (e.g. GelSight sensor [28]), we also use CNNs to extract features, but for six-axis force/torque sensors, causal convolutions [29] may be more suitable.

B. The VTFSA Module

The primary goal of the VTFSA module is to learn an effective VTF representation based on the given visual and tactile features. Most of the existing methods adopt direct concatenation of features from two different modalities, which we called Direct-Fusion (DF) method, as shown in Fig. 3. However, the DF method still stays in the junior stage of multimodal learning [15].

Different from the DF method, VTFSA extracts the fusion features in two steps. Firstly, the early VTF feature is obtained by the “slice-concatenation” of the visual and tactile features. Secondly, a self-attention module is performed on the early VTF feature to obtain the final VTF feature.

1) The early VTF feature: Given the visual M_V and tactile features M_T, the early VTF feature $M^E_{V,T}$ is constructed based on them.

$$M^E_{V,T} = M_V \uplus M_T$$
$$M_V \in \mathbb{R}^{H_V \times W_V \times C_V},\ M_T \in \mathbb{R}^{H_T \times W_T \times C_T} \tag{4}$$

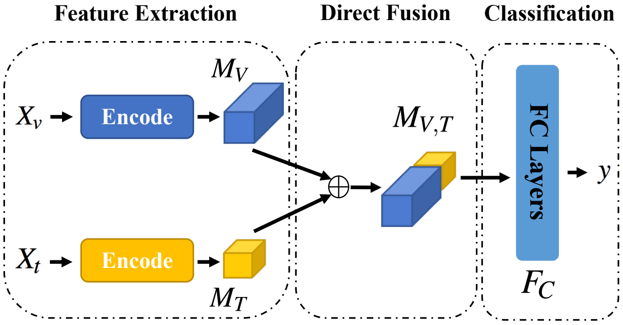

Fig. 3: The network architecture of DF model.

where H_V, W_V and C_V are the height, width and feature channel dimensions of M_V, respectively. These notations are similar for the tactile feature M_T. $\uplus$ operation indicates the “slice-concatenation”, which is described below.

Let p be a spatial location in the feature map M_V, $p \in \{1,2,...,H_V \times W_V\}$. v_p is used to denote the “slice-vector” of the visual feature map at the spatial location p. Similarly, the spatial position and “slice-vector” of the tactile modality are represented by $q \in \{1,2,...,H_T \times W_T\}$ and t_q, respectively. Thus we can define the early VTF feature vector $f_{p,q}$ corresponding to the location p and q in visual and tactile feature map as follows

$$f_{pq} = \text{Concate}(v_p, t_q) \tag{5}$$

where Concate() operation denotes the concatenation of the two “slice-vectors”. The fusion feature vector encodes the combination of a specific location p in the visual feature map and a specific location q in the tactile feature map with a total dimension of $(C_v + C_t)$. As a result, the early VTF feature is expressed as $M^E_{V,T} = \{f_{p,q} : \forall p, \forall q\}$. The dimension of $M^E_{V,T}$ is $(H_V \times H_T) \times (W_V \times W_T) \times (C_V + C_T)$. An illustration of this construction process is shown in Fig. 4.

In this way, the early VTF feature ensures cross-modal interaction of visual and tactile features at each spatial location by combining slice vectors at each position of the

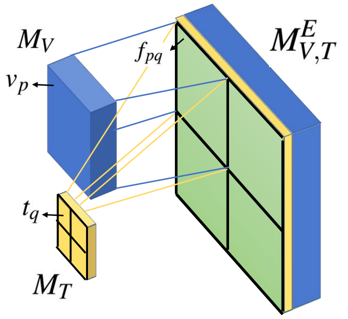

Fig. 4: Illustration of the construction process of the early VTF feature.

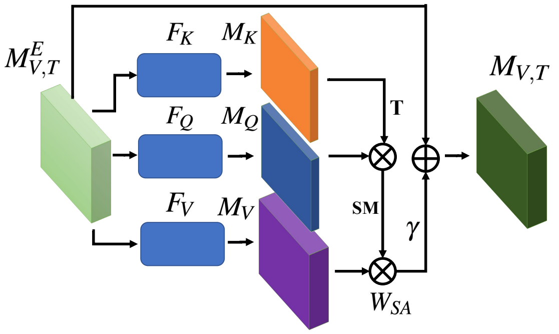

Fig. 5: Detailed schematic of SA mechanism in VTFSA. The $\otimes$ denotes matrix multiplication and $\oplus$ means element-wise addition. SM indicates the SoftMax operation.

two original feature maps, which lays the foundation for the subsequent self-attention module to capture the uni-modal and cross-modal information between arbitrary spatial positions of the visual and tactile features.

2) The self-attention mechanism: The early VTF featuremap $M^E_{V,T}$ is quite large and may contain much redundant information. We adopt the Self-Attention (SA) mechanism to further streamline and extract VTF features that are beneficial to the task. It achieves the effect of attention by adding different weights to different positions of the original feature map, which is motivated by our followed observation.

Different spatial positions in single-modality and feature combinations at different positions across modalities have different importance for the determination of the current grasp result. For example, the contact position between the object and the gripper in the visual image is obviously more useful than other positions. However, $M^E_{V,T}$ contains sufficient uni-modal and cross-modal features while it does not distinguish between different spatial positions and their combinations, which poses a considerable challenge for subsequent classifiers. Therefore, we adopt the SA mechanism to achieve further position-related cross-modal feature extraction.

The detailed architecture of the SA mechanism in VTFSA is shown in Fig. 5. It takes $M^E_{V,T}$ as input, and generates a weighted feature map W_{SA} to distinguish the importance of different spatial positions. W_{SA} is added to $M^E_{V,T}$ by residual connection [30]. Note that $M^E_{V,T}$ and W_{SA} have the same size and can be connected directly by element-wise addition.

The core function of SA module is to generate a position-sensitive weight feature map based on the input feature map, which is implemented by building three feature spaces, including Key, Query, and Value, as shown in Fig. 5. Given the early VTF feature $M^E_{V,T} \in \mathbb{R}^{H_F \times W_F \times C_F}$, where $H_F = H_V \times H_T$, $W_F = W_V \times W_T$, and $C_F = C_V + C_T$ represent the height, width, and channel of $M^E_{V,T}$ respectively, the SA module first produces a set of feature maps M_K, M_Q, and M_V by 1×1 convolutions.

$$\begin{aligned} M_K &= F_K(M^E_{V,T}) = Conv(M^E_{V,T}) \ (M_K \in \mathbb{R}^{H_F \times W_F \times \tilde{C_F}}) \\ M_Q &= F_Q(M^E_{V,T}) = Conv(M^E_{V,T}) \ (M_Q \in \mathbb{R}^{H_F \times W_F \times \tilde{C_F}}) \\ M_V &= F_V(M^E_{V,T}) = Conv(M^E_{V,T}) \ (M_V \in \mathbb{R}^{H_F \times W_F \times C_F}) \end{aligned} \tag{6}$$

where $Conv$ denotes 1×1 convolution operation. $\tilde{C_F}$ indicates the number of channels of M_Q and M_K, which is set to reduce the amount of calculation. $\tilde{C_f}$ is set as $\sqrt{C_f}$ in this paper. M_K and M_Q are used to capture the interdependence between each position and all other positions in the input feature map, and M_V is adopted to generate specific weight values for each position.

Next, the correlation between each "slice vector" of the input feature map and other "slice vectors" is calculated by matrix multiplication of M_K and M_Q.

$$\begin{gathered} a_{i,j} = \frac{\exp\left(s_{i,j}\right)}{\sum_{j=1}^{j=N} \exp(s_{i,j})} \\ s_{i,j} = M^T_{K_j} M_{Q_i}, \ N = H_F \times W_F \end{gathered} \tag{7}$$

where $a_{i,j}$ indicates the attention coefficient between the vector of M_K at the i^{th} location and the vector of M_Q at j^{th} location. Then, the final attention weight of each position $W_{SA,i}$ is obtained by multiplying the weights of other positions by the correlation coefficients and summing them.

$$W_{SA,i} = \sum_{j=0}^{j=N} a_{i,j} M_{V_i} \tag{8}$$

where M_{V_i} indicates the vector of M_V at the i^{th} location.

Finally, the final VTF feature $M_{V,T}$ as the output of the VTFSA module is obtained by

$$M_{V,T} = \gamma W_{SA} + M^E_{V,T} \tag{9}$$

where γ is a scale parameter to adjust the attention mechanism effect on the early VTF feature and is initialized as 0.

In this way, the whole process can be regarded as parameter weighting at each position and channel of $M^E_{V,T}$ without changing its shape. Furthermore, the residual connection allows the insertion of VTFSA module into the backbone network without breaking its behavior and ensure the effectiveness of gradient back-propagation [30].

C. The Classification Module

The final part of the proposed model is a classification module, which classifies the extracted final VTF feature ($M_{V,T}$) and outputs the result (y). The final representation for this task is obtained by average-pooling over all position ($H_F \times W_F$), which can be summarized as:

$$\begin{aligned} f_m &= \text{avg-pool}_n(M_{V,T}) = \frac{\sum_{n=1}^{N} f_{V,T,m,n}}{N} \\ \hat{M}_{V,T} &= \{f_m : \forall m\},\ \hat{M}_{V,T} \in \mathbb{R}^{1\times 1\times C_F} \end{aligned} \tag{10}$$

where m and n denote the channel and position index on $M_{V,T}$, respectively. $f_{V,T,m,n}$ indicates the value of n^{th} position, m^{th} channel of $M_{V,T}$. As a result, we use $\hat{M}_{V,T}$ as the input of the classification layers.

Finally, $\hat{M}_{V,T}$ is fed into a two-layer Fully Connection (FC) layers to predict whether the current grasp will succeed, and the number of nodes in the two FC layers are set to $C_F/4$ and 2, respectively.

V. Experiments: design and setup

In this section, we perform some comparative experiments on two public multi-modal grasping datasets [4], [31] containing visual and tactile sensing data to verify the performance of the proposed model. The goal of our experimental evaluation to answer the following questions:

- Is the proposed method (VTFSA) better than traditional methods in the grasp outcomes prediction tasks?
- Can the VTFSA model handle different forms of tactile signals?
- If the answer to the first question is yes, why is the VTFSA module helpful?

A. Datasets Introduction

*1) **D0**:* The visual and tactile data of the **D0** dataset is collected by a Microsoft Kinect 2 camera and two GelSight sensors [28], respectively. The GelSight tactile sensors provide raw-pixel measurements at a resolution of 1280x960 @30Hz over an area of 24x18mm. The GelSight grasping dataset collects 9269 grasping trials from 106 unique objects. Besides, each routine consists of six images and its corresponding grasp outcome label. These six images are taken by the camera and the two tactile sensors during each grasp routine.

*2) **D1**:* This dataset is built by a novel designed dexterous robot hand and two RealSense RGB-D cameras [31]. The most significant difference between this dataset and the previous is that the visual data is acquired by two views (right side and front side), and the tactile signals are time serials. The dataset contains 2550 sets of data. Similarly, we select the visual images (four images taken by two cameras) before and during the grasp and the tactile time series during grasping to train and evaluate different methods.

B. Experiment Setup

*1) **D0**:* Since the data collected in this dataset are abundant, we first combine different inputs of **D0** to evaluate how different models perform for different inputs. The difference between these inputs is whether the visual and tactile images before grasping are subtracted from the images during grasping. The reason for this is that the tactile images' subtraction will make the change caused by grasping more noticeable, while visual images may not.

- **I1**: The left and right tactile images during grasping are spliced, and the visual and tactile images during grasping are respectively subtracted by the images before grasping to obtain a visual difference image and a tactile difference image, and these two difference images are fed into two CNNs.
- **I2**: Different form **I1**, the visual difference image and spliced tactile image (not subtracted) are fed into two CNNs.
- **I3**: Different form **I1**, the visual image during grasping (not subtracted) and the tactile difference image are fed into two CNNs.

Based on the above different inputs, we compare our model to some baselines:

- **Original [4]**: Six images are combined and sent into three CNNs (the inputs of origin model [4], Inception-v4).
- **DF+I1**: The original three CNNs of [4] are changed to two, and change the corresponding inputs.
- **DF+I2**: Same as **I1+DF** except that the inputs are changed to **I2**.
- **DF+I3**: Same as **I1+DF** except that the inputs are changed to **I3**.
- **VTFSA+I1**: The VTFSA model with inputs **I1**.
- **VTFSA+I2**: The VTFSA model with inputs **I2**.
- **VTFSA+I3**: The VTFSA model with inputs **I3**.

Furthermore, we also perform ablation studies of the proposed model on dataset **D0**. To figure out whether the two modules (early concatenation and SA mechanism) of the VTFSA module are helpful, we also evaluate the VTFSA model without one of them.

- **VTFSA_noec**: After being extracted, the visual and tactile features are fed into SA modules directly and then combined as the DF model does.
- **VTFSA_nosa**: We only combine the visual and tactile features by "slice concatenation" without the SA module.

*2) **D1**:* We combine the images taken at two views of the **D1** dataset into one image as the visual input and use the last 50 readings of the tactile time series as the tactile input. Similarly, we also compare our model to two baselines on the **D1** dataset:

- **Original [31]**: The original model of [31].
- **DF**: The DF model with the given inputs.
- **VTFSA**: The VTFSA model with the given inputs.

Specifically, we use cross-entropy [32] as the loss function and Adam optimizer [33] as the optimizer. In the training process of the VTFSA model, we first initialize the CNNs parameters with the weights trained on the ImageNet and freeze them, then train the other parts of the model with three epoch at a learning rate of 1e-3. After that, we freeze these

TABLE I: Performance comparison of different models with different inputs on **D0** dataset.

Models	Precision %	Recall %	F1 score %
Original [4]	79.97	79.36	79.65
DF+I1	81.28	79.34	80.17
DF+I2	77.70	74.44	75.63
DF+I3	81.65	78.95	80.04
VTFSA+I1	85.08	79.81	81.67
VTFSA+I2	81.29	76.94	78.84
VTFSA+I3	**87.06**	**83.95**	**85.23**

TABLE II: Performance comparison of different models on **D1** dataset.

Models	Precision %	Recall %	F1 score %
Original [31]	94.60	/	/
DF	95.74	91.78	93.58
VTFSA	**98.55**	**97.37**	**97.93**

parts and train five epochs on the two CNNs with a learning rate of 1e-4. Finally, the model is completely unfrozen to train 5 epochs at a learning rate of 1e-5. All models are built by using the PyTorch-1.2 development package and trained on an NVIDIA DGX server. The batch size is set as 64 in this paper. For detailed parameter settings and source code of these models, please refer to the supplementary materials and source code at `https://github.com/swchui/VTFSA`.

VI. Experiments: results and analysis

A. Results

1) Statistics Results: To more comprehensively and accurately evaluate the performance of the proposed model, we compare the Precision, Recall, and F1 score of different models with different inputs.

Table I presents the comparison results on **D0** dataset. The results show that the original model [4] with six images inputs achieves 79.97 % precision, which is nearly consistent with the original results (77.8 %). The DF model with I3 inputs achieves 81.65 % precision, which indicates that I3 is more suitable for this grasp outcomes prediction task. Additionally, the VTFSA model can obtain better prediction performance than traditional methods on different input combinations. VTFSA model with I3 inputs realizes 87.06 % precision, which is **5+%** higher than the DF model and **7+%** higher than the original model. Table II gives the performance comparison results on **D1** dataset. The VTFSA model also achieves **4+%** Precision and **6+%** Recall improvement over the DF model on **D1** dataset.

These results answer our first and second question. The answers are both **positive**. The results show that the proposed model not only has better performance than the traditional methods in the grasp outcomes prediction tasks but also can address different forms of tactile signals.

Note that the proposed VTFSA model's network parameters are 23.38 M more than the traditional DF models, which makes the proposed model relatively inefficient. However, compared to the feature extraction CNNs, the extra computational memory can be negligible when inference.

TABLE III: Ablation studies results of VTFSA model on **D0** dataset.

Models	Precision %	Recall %	F1 score %
Original [4]	79.97	79.36	79.65
DF	81.65	78.95	80.04
VTFSA_noec	82.32	79.87	80.89
VTFSA_nosa	82.12	79.94	80.86
VTFSA	**87.06**	**83.95**	**85.23**

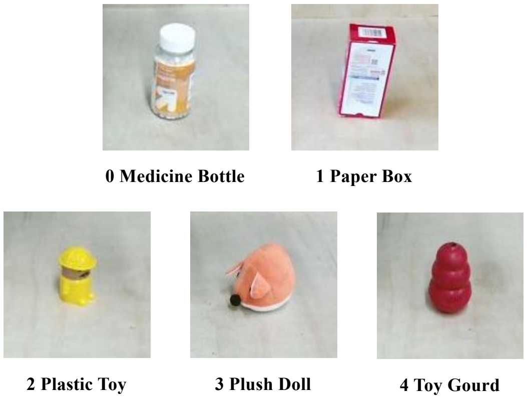
0 Medicine Bottle
1 Paper Box
2 Plastic Toy
3 Plush Doll
4 Toy Gourd

Fig. 6: The selected five household objects with different sizes, shapes, and materials in the test set of **D0**. Please note that these images are from dataset **D0** [4].

2) Ablation Studies: The early concatenation operation and SA mechanism are two critical components of the VTFSA module. According to the ablation studies results shown in Table III, both VTFSA_noce and VTFSA_nosa achieve better prediction accuracy than the traditional methods while far form the performance of VTFSA. These findings indicate that both of the early concatenation operation and the SA mechanism are critical to the proposed model.

The results of ablation studies are also consistent with our intuition that task-related cross-modal fusion features can be captured by constructing the fusion featuremap firstly and then conducting the SA operation. If only the fusion featuremap is constructed without SA operation, only the permutations of features are changed without further feature extraction. Similarly, if the SA operation is performed on the uni-modal featuremap, it is equivalent to building more complex uni-modal features extraction networks, which does not help to extract the cross-modal fusion features.

3) Performance on specific objects: We select five household objects with different sizes, shapes, and materials in the test-set of **D0**, as shown in Fig. 6. We study the applicability of such VTFP methods to different objects by comparing the accuracy of their evaluating results on these five objects.

The results are shown in Fig. 7, which indicates that the performance of VTFSA on the paper box, the plastic toy, and the plush doll is significantly better than other methods. All three VTF models perform poorly on the toy gourd, which

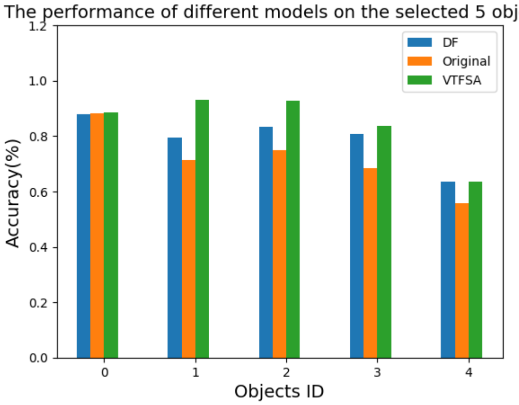

Fig. 7: The performance of different models on the selected five objects.

may be caused by the irregular shape of the gourds. For the other four objects, the prediction results indicate that VTFSA can achieve accuracy close to or exceeding 90%, which provides the cornerstone for its further application to dexterous grasping tasks. Note that these objects are not involved in the model training process, proving that our proposed model has a certain generalization performance.

B. *Visualization Analysis*

To figure out why the VTFSA module is helpful for this task, we visualize the heatmaps of the main network parameters of the VTFSA model at different stages of the inference process and overlay them on the original inputs for presentation. These heatmaps are obtained by normalizing the strongest activated channel of different featuremaps, which are up-sampled to match the shape of origin visual and tactile input images.

We visualize the attention weights learned in the VTFSA module during an inference process, and the results are shown in Fig. 8. These heatmaps indicate that the VTFSA module can further capture some cross-modal position-related features that are beneficial to the task. For example, the visual attention weights are more distributed in the object and gripper part, and the tactile weights are more concentrated on contact parts (Red circled areas in Fig. 8), which demonstrates that the cross-modal combination of these parts is more critical for this task than other positions. Thus, we can answer our third question: The VTFSA module can further learn some cross-modal position-dependent features, which may be helpful for prediction.

Furthermore, we find that the DF model often predicts a failed grasp as a success by statistical analysis of the predicted results, while the VTFSA model can predict correctly. We also visualize one of the results with heatmaps, as shown in Fig. 9. The DF heatmaps show that the DF model only focuses on the contact parts (Red circled areas in Fig. 9) of the left GelSight image while ignores the right GelSight image. However, the influence of the lack of solid contact on the right side of the grasping result is decisive, which directly leads to the failure of DF model prediction. Fortunately, the VTFSA model pays more attention to the contact part of the left GelSight image and the whole right GelSight tactile image, and it successfully predicts this grasp result, which proves that it has a more comprehensive and robust prediction performance compared to DF methods.

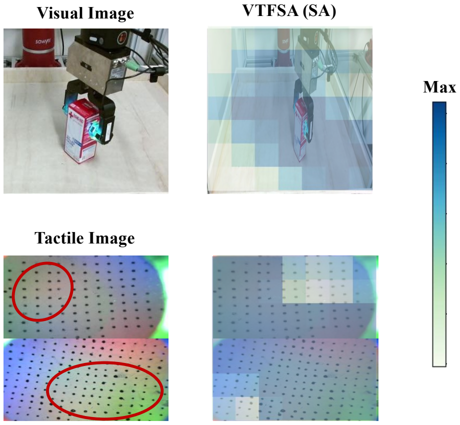

Fig. 8: Heatmaps visualization of the VTFSA module. The attention featuremap obtained by the VTFSA module is separated and averaged according to the modality, and then the strongest activated channel is selected to generate these heatmaps. Please note that these images are from dataset **D0** [4].

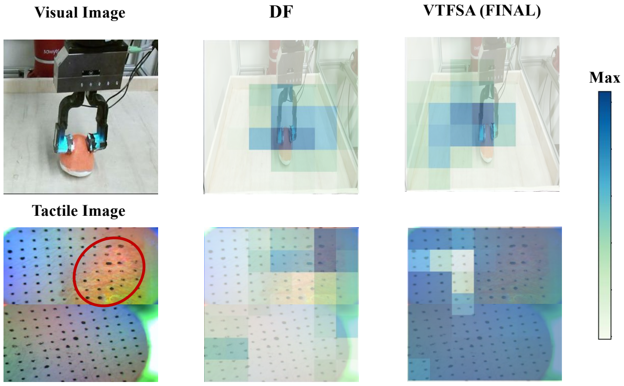

Fig. 9: Heatmaps visualization of the DF and VTFSA model. The DF heatmaps are obtained by normalizing the strongest activated channel of the uni-modal featuremaps. The upper part of the tactile image corresponds the left GelSight image. Please note that these images are from dataset **D0** [4].

In conclusion, visualization analysis indicates that the VTFSA model can further capture some cross-modal position-related features and have better robustness and not easy to overfit, which is very helpful for improving prediction accuracy.

VII. Conclusion and Discussion

This study proposes a new visual-tactile learning method based on the Self-Attention mechanism for a grasp outcomes prediction task. We compare our method with traditional methods on two public multimodal grasping datasets. The experimental results not only show that VTFSA model has better performance than the traditional methods in this grasp outcomes prediction task, but also turn out that it can address different forms of tactile signals. Additionally, the proposed model achieves an accuracy close to or exceeding 90% on household objects with regular shapes. Ablation studies indicate that both modules of the VTFSA module are critical for this task. Further visualization analysis shows that the VTFSA module can further learn some cross-modal position-dependent features, which may be helpful for prediction. Moreover, we find that the proposed model obtains a more comprehensive and robust prediction performance than traditional methods.

Although we have made some improvements in visual-tactile fusion learning, this progress is not apparent, possibly due to the complex structure of the VTFSA module. In the furture, we will try to simplify this model and explore more new methods in the field of visual-tactile fusion learning.

References

[1] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6336, pp. 1149-+, Jun. 2019.

[2] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, Apr. 2018.

[3] J. M. Romano, K. Hsiao, G. Niemeyer, S. Chitta, and K. J. Kuchenbecker, "Human-Inspired Robotic Grasp Control With Tactile Sensing," *IEEE Transactions on Robotics*, vol. 27, no. 6, pp. 1067–1079, Dec. 2011.

[4] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?," *arXiv preprint arXiv:1710.05512*, 2017.

[5] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3300–3307, Jul. 2018.

[6] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, Apr. 2015.

[7] L. Pinto and A. Gupta, "Supersizing Self-supervision: Learning to Grasp from 50K Tries and 700 Robot Hours," in *IEEE International Conference on Robotics and Automation*, Stockholm, Sweden, May 2016, pp. 3406–3413.

[8] I. Kamon, T. Flash, and S. Edelman, "Learning to grasp using visual information," in *Proceedings of IEEE International Conference on Robotics and Automation*, Apr. 1996, vol.3, pp. 2470–2476.

[9] A. Saxena, J. Driemeyer, and A. Ng, "Robotic grasping of novel objects using vision," *International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, Feb. 2008.

[10] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Reviews Neuroscience*, vol. 10, no. 5, pp. 345–359, May. 2009.

[11] Y. Gao, L. A. Hendricks, K. J. Kuchenbecker, and T. Darrell, "Deep learning for tactile understanding from visual and haptic data," in *IEEE International Conference on Robotics and Automation*, Stockholm, Sweden, May. 2016, pp. 536–543.

[12] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual-tactile fusion for object recognition," *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 2, pp. 996–1008, Apr. 2017.

[13] S. Wang, J. Wu, X. Sun *et al.*, "3d shape perception from monocular vision, touch, and shape priors," *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Madrid, Spain, Oct. 2018, pp. 1606–1613.

[14] H. Zheng, L. Fang, M. Ji, M. Strese, Y. Ozer, and E. Steinbach, "Deep learning for surface material classification using haptic and visual information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2407–2416, Dec. 2016.

[15] A. Zadeh, S. Poria, P. Liang *et al.*, "Memory fusion network for multi-view sequential learning," *Thirty-Second AAAI Conference on Artificial Intelligence*, Louisiana, USA, Feb. 2018, pp. 5634–5641.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, California, USA, Dec. 2017, pp. 5998–6008.

[17] C. Goldfeder and P. K. Allen, "Data-driven grasping," *Autonomous Robots*, vol. 31, no. 1, pp. 1–20, Jul. 2011.

[18] K. B. Shimoga, "Robot grasp synthesis algorithms: a survey," *International Journal of Robotics Research*, vol. 15, no. 3, pp. 230–266, Jun. 1996.

[19] U. R. Aktas, C. Zhao, M. Kopicki, A. Leonardis, and J. L. Wyatt, "Deep Dexterous Grasping of Novel Objects from a Single View," *arXiv preprint arXiv:1908.04293*, 2019.

[20] H. Dang and P. K. Allen, "Learning grasp stability," in *IEEE International Conference on Robotics and Automation*, Saint Paul, MN, May. 2012, pp. 2392–2397.

[21] A. Murali, Y. Li, D. Gandhi, and A. Gupta, "Learning to grasp without seeing," *arXiv preprint arXiv:1805.04201*, 2018.

[22] J. Li, S. Dong, and E. Adelson, "Slip detection with combined tactile and visual information," in *IEEE International Conference on Robotics and Automation (ICRA)*, Brisbane, QLD, 2018, pp. 7772–7777.

[23] S. Cui, R. Wang, J. Wei, F. Li, and S. Wang, "Grasp state assessment of deformable objects using visual-tactile fusion perception," *arXiv preprint arXiv:2006.12729*, 2020.

[24] N. Fazeli, M. Oller, J. Wu, Z. Wu, J. B. Tenenbaum, and A. Rodriguez, "See, feel, act: Hierarchical learning for complex manipulation skills with multisensory fusion," *Science Robotics*, vol. 4, no. 26, Jan. 2019.

[25] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah *et al.*, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," *arXiv preprint arXiv:1810.10191*, 2018.

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[27] X. Wang, R. Girshick, A. Gupta, and K. M. He, "Non-local neural networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, UT, USA, Jun. 2018, pp. 7794–7803.

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

[29] A. v. d. Oord, S. Dieleman, H. Zen *et al.*, "Wavenet: a generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, NV, USA, Jun. 2016, pp. 770–778.

[31] C. Yang, P. Du, F. Sun, B. Fang, and J. Zhou, "Predict Robot Grasp Outcomes based on Multi-Modal Information," in *IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Kuala Lumpur, Malaysia, Dec. 2018, pp. 1563–1568.

[32] P. De Boer, D. P. Kroese, Owens, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[33] D. Kingma, and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.