2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

STNet: Spatio-Temporal Fusion-Based Self-Attention for Slip Detection in Visuo-Tactile Sensors

Jin Lu, Bangyan Niu, Huan Ma, Jiafeng Zhu, and Jingjing Ji*, *Member, IEEE*

***Abstract*—Slip detection plays a pivotal role in the dexterity of robotics, improving the reliability and precision of manipulations but also contributing to safety, efficiency, and adaptability. Deep learning-based slip detection algorithms commonly difficult to concentrate on key features when faced with dense 3D shape data obtained by visuo-tactile sensors. Data from noncontact locations can interfere with slip judgements and the ignorance of interframe linkage can also lead to slip detection failure. In this paper, a new spatio-temporal sequences fusion-based self-attention, STNet, is proposed to perform slip detection by allocating more attention to the object-sensor contact area when processing complex 3D shape data. A binocular visuo-tactile system (BVTS) is designed and fabricated for dataset construction. The entire 3D shape dataset containing 4 motion patterns, including stationary, pressing, rolling and slipping. Self-attention architecture with and without spatio-temporal sequences fusion mechanism (denoted as STNet and TemNet, respectively) are trained based on the same dataset. The experiments show the validity of STNet, which can reach 98.91% slip detection accuracy. Meanwhile, the ablation studies confirm the effectiveness of the spatio-temporal sequences fusion mechanism.**

I. INTRODUCTION

Human hand can increase the gripping force in time to achieve a more stable grip when it senses the object is about to slip [1]. Robots are also expected to have this ability in their grasping. Slip refers to the unintended movement or grip failure between the robot's end-effector (such as a gripper or tool) and the object it is manipulating [2-3]. Detecting slips provides valuable feedback for robots to make real-time adjustments and optimize force control performance.

Robots performing slip detection are mainly based on vision [4] and tactile senses [5]. Vision can only tell whether contact and deformation have been made from an external perspective. While a relative displacement of contact surfaces can be determined directly by tactile senses to judge whether slippage has occurred. Tactile sensors have flourished in recent years thanks to manufacturing development and tremendous work has been devoted to their corresponding slip detection. Francomano *et al.* [6] summarize a detailed survey of slip detection methods using various tactile sensors.

This work was supported in part by the National Key R&D Program of China under Grant 2021YFB3200700; in part by the National Science Foundation of China under Grant 52175510, Grant 52188102; and in part by Hubei Provincial Natural Science Foundation of China under Grant 2023AFA085.

Jin Lu, Bangyan Niu, Huan Ma, Jiafeng Zhu, and Jingjing Ji are with the State Key Laboratory of Intelligent Manufacturing Equipment and Technology (SKL-IMET), Huazhong University of Science and Technology (HUST), Wuhan, Hubei 430074, China. (Email: jinlu@hust.edu.cn; niubangyan@hust.edu.cn; huanm@hust.edu.cn; zhujiaf@hust.edu.cn; jijingjing@hust.edu.cn).

*Corresponding author: Jingjing Ji

Among various types of tactile sensors, visuo-tactile sensors [7] are able to provide the richest tactile perception, including high-resolution 3D surface shape data (point cloud) [8], shown in 1st row of Fig. 1. Displacement field and force field [9] can be calculated accordingly. The multi-modal sensing capabilities also provide more options for slip detection, which can be classified into two categories: model-based and deep learning-based methods.

Model-based approach can significantly increase the success rate of slip judgment in a given situation. Yuan *et al.* [10] introduce a method of inferring the state of the contact interface based on analysis of the image sequence of visuo-tactile sensors elastomer medium. Dong *et al.* [11] explore a new design of Gelsight, which can detect translational and rotational slip in general cases. Raj *et al.* [12] propose a model-based algorithm that detects rotational slip patterns and measures rotational displacement. However, contact conditions vary and the model will greatly limit its ability to generalize.

Deep learning-based methods commonly achieve good performance when facing large amounts of data, so as to slip detection for visuo-tactile sensors. Li *et al.* [13] propose a new method based on deep neural network to detect slip and achieve a detection accuracy as high as 88.03%. Zhang *et al.* [14] develop a novel optical-based tactile sensor FingerVision and propose a slip classification framework based on Convolution Long Short Term Memory (ConvLSTM) network. Cui *et al.* [15] present a bimodal PE-ConvLSTM network to effectively utilize 2D displacement data and 3D contact shape data.

Visuo-tactile sensors can provide high-resolution perception information, but this is not necessarily a good thing for slip detection. Most existing algorithms use whole 3D shape data or entire physical fields as direct input to the network, which seems that the network obtains more information, but the redundancy is likely to mislead slip judgment. Specifically, data from non-contact locations can interfere with slip judgements. Slip is only related to the area where the sensor is in contact with the object. Non-contact areas may also generate deformation due to factors such as the gravity of gripped object, which have no effect on slip detection and may distract the network. Besides, slip is judged based on temporal signals, and the concept of slip cannot be generated from a single frame image. The ignorance of inter-frame linkage can also lead to slip detection failure. Therefore, a successful network for visuo-tactile sensors' slip detection should focus on the continuity change of the contact position.

In this paper, we propose a new self-attention based architecture, STNet, illustrated in the 2nd row of Fig. 1, to combine spatial and temporal sequences for visuo-tactile

sensors' slip detection. The multi-head self-attention (MSA) mechanism in transformers [16] allows them to capture subtle patterns and dependencies in the data, which can be crucial for slip detection and stable object manipulation. The 3D shape data are input into an encoder-decoder module to extract the contact feature from entire cloud map, which constructs the spatial part. This extraction strengthens the spatial features and weakens the temporal variations. To ensure that the network also focuses on the temporal information at the same time, unprocessed shape data is also fed directly, which forms the temporal part. The spatial part and temporal part are fused and fed into self-attention block simultaneously to consider the motion sequence and complete slip judgment. The experiments and ablation studies confirm the validity of the network and the effectiveness of the spatio-temporal sequence fusion mechanism.

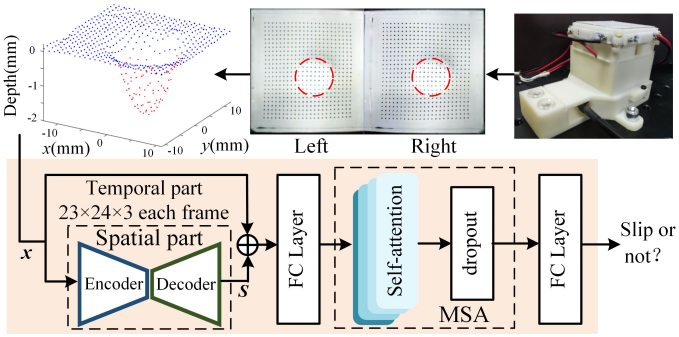

Fig. 1. The slip detection flowchart.

The remainder of this paper offers the followings.

1) The principle and structure of binocular visuo-tactile sensor is briefly described. Then the STNet architecture is detailed and the spatio-temporal sequence fusion mechanism is formulated to ensure the network focuses on the continuity change of the contact position.

2) A visuo-tactile sensor prototype is designed and fabricated for data collection. The dataset is constructed and analyzed while experimental configurations and training strategies for STNet are presented.

3) The experimental results and ablation studies are illustrated, confirming the validity of STNet and the effectiveness of spatio-temporal sequence fusion mechanism.

II. Sensor and Algorithm Structure

STNet is implemented in the binocular visuo-tactile system (BVTS) [17] proposed in our previous work. A brief introduction of BVTS is given first and the STNet architecture will be detailed.

A. Binocular Visuo-Tactile System (BVTS)

A binocular camera-based visuo-tactile sensor is used for data collection in this paper. The exploded view of the designed BVTS with a quarter section is shown in Fig. 2(a). The sensor has a flexible contact surface with coated markers and a shading layer placed on the top. Internal LEDs provide supplemental light to ensure the sensor operates independently of the external environment. When in contact with an object, the elastomer pad deforms and conforms to the object's shape. Surface contours can be clearly reflected in the deformation of the markers. As the deformation proceeds, the embedded binocular cameras continuously capture subsequent images in 20Hz. Compared with other popular visuo-tactile sensors (e.g. GelSight), the marker points on binocular visuo-tactile sensors provide greater measurement accuracy.

The transparent elastomer, coated with 23×24 markers (surface range of 22 mm×21 mm, with a depth error of less than 4%) [17], is the part that comes into contact with the object during gripping process. Four sets of LEDs located around the elastomer are powered by a controllable voltage/current device RIGOL DP1308A to supply uniform and controlled illumination to the object's surface. Acrylic supporting plate underneath the elastomer ensures that the elastomer will not be damaged by excessive pressing. The two cameras $\boldsymbol{C_l}$, $\boldsymbol{C_r}$ capture images of the markers from their respective viewpoints and the shape of the contact surface is reconstructed consecutively through parallax. Kanade-Lucas-Tomasi [18] method is used for sparse optical flow marker tracking, which is completed by *PointTracker* function in MATLAB.

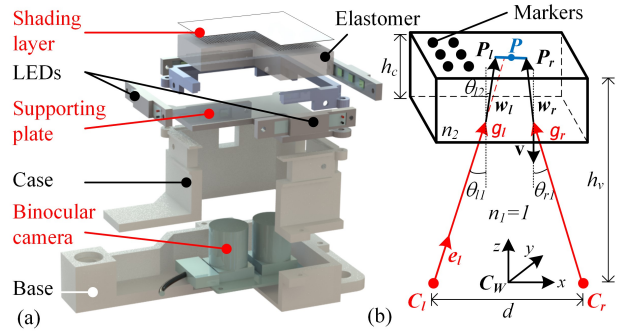

Fig. 2. The structure of the visuo-tactile sensor. (a). CAD model. (b). refraction model.

When light passes through the interface between air and the elastomer, it will bend, causing refraction and leading to distortions which greatly constrain the 3D reconstruction accuracy and surface texture reconstruction. To obtain more precise 3D coordinates, the parameters of the refraction model are split from those of stereovision during calculation in our previous work. Taking the left camera $\boldsymbol{C_l}$ as an example, as shown in Fig. 2(b). The ray starting from $\boldsymbol{C_l}$ follows the direction of the unit vector $\boldsymbol{e_l}$, reaches the bottom of the elastomer, and then emits along $\boldsymbol{w_l}$. Similarly, the light captured by the $\boldsymbol{C_r}$ along the $-\boldsymbol{w_r}$ and then the $-\boldsymbol{g_r}$. The two endpoints of the shortest perpendicular bisector between the two rays are $\boldsymbol{P_l}$ and $\boldsymbol{P_r}$. The light path can be divided into two segments in different media, $\boldsymbol{g_l}$ and $\boldsymbol{w_l}$ with corresponding refractive indexes n_1 and n_2. These two segments are calculated as (1).

$$\boldsymbol{g_l} = \frac{h_v - h_c}{\|\boldsymbol{e}_{lz}\|_2}\boldsymbol{e_l} \tag{1a}$$

$$\boldsymbol{w_l} = \frac{[\boldsymbol{e_l} + (n_r \cos\theta_{l2} - \cos\theta_{l1})\mathbf{v}]}{\|\boldsymbol{e_l} + (n_r \cos\theta_{l2} - \cos\theta_{l1})\mathbf{v}\|_2} \tag{1b}$$

where (h_v, h_c) represent the distance between camera and shading layer and the thickness of elastomer and supporting plate. The e_{lz} is the component of the unit vector $\boldsymbol{e_l}$ on the z-axis. θ_{l1}, θ_{l2} are the angles of the ray to the normal vector $\mathbf{v}$ before and after it transmitting through the elastomer, respectively. $n_r \sin\theta_2 = \sin\theta_1$ based on the Snell's law, where $n_r = n_2/n_1$. The $\boldsymbol{P_r}$ detected by $\boldsymbol{C_r}$ can be calculated via the same procedure. The positions of $\boldsymbol{P_l}$, $\boldsymbol{P_r}$ are presented in (2) related to the world coordinate $\boldsymbol{C_w}$, which is located in the middle of

the two cameras. However, two unknown variables, λ_l and λ_r, still exist in (2).

$$\begin{bmatrix} \boldsymbol{P_l} \\ \boldsymbol{P_r} \end{bmatrix} = \begin{bmatrix} \boldsymbol{g_l} - \boldsymbol{d} \\ \boldsymbol{g_r} + \boldsymbol{d} \end{bmatrix} + \begin{bmatrix} \lambda_l \boldsymbol{w_l} \\ \lambda_r \boldsymbol{w_r} \end{bmatrix} \text{ where } \boldsymbol{d} = [d/2, 0, 0]^{\mathrm{T}}. \quad (2)$$

According to the geometric constraints, the vector $\overrightarrow{P_l P_r}$ is perpendicular to both $\boldsymbol{w_l}$ and $\boldsymbol{w_r}$, so the $[\lambda_l, \lambda_r]$ constraint is established as (3). Actual position of $\boldsymbol{P}$ is located on the $\overrightarrow{P_l P_r}$ which can be present as $\lambda(\boldsymbol{P_l} - \boldsymbol{P_r})$.

$$[\lambda_l \quad \lambda_r \quad \lambda]^{\mathrm{T}} = [-\boldsymbol{w_l} \quad \boldsymbol{w_r} \quad \boldsymbol{P_l} - \boldsymbol{P_r}]^{-1}(\boldsymbol{g_l} - \boldsymbol{g_r}) \quad (3)$$

where each vector is composed of $[x, y, z]^{\mathrm{T}}$

After obtaining $\boldsymbol{P_l}$ and $\boldsymbol{P_r}$, the exact position of the marker $\boldsymbol{P}$ can be approximated as the midpoint of $\boldsymbol{P_l P_r}$. By explicitly modeling the refraction effects to estimate the 3D depth map and faithfully reconstructing the points' height, the erroneous appearance can be eliminated. For a specific derivation of the refraction model, please refer to [17]. The reconstructed 3D shape data (point cloud) captured by the binocular camera is the basis for subsequent network training.

B. Spatio-Temporal Network

Aiming at constructing a network focusing on the continuity change of the contact position, 3D shape data with and without an encoder-decoder module are fused. The integrated data are fed into self-attention to simultaneously make full use of its significant advantage for temporal signal processing and emphasize the importance of the object-sensor contact region for slip detection. The whole network is composed of two parts: spatio-temporal fusion and self-attention (Fig. 3 and 4).

The overall process is as follows:

Firstly, the labeled point cloud data $\boldsymbol{x}_i$ (i = 1, …, n) is obtained by the above-mentioned BVTS and then flattened and combined into a sequence to form a sample ($x_1 \sim x_n$). A sample can be stored as two-dimensional data in frame number × flattened point cloud data ($n \times x_i$) format.

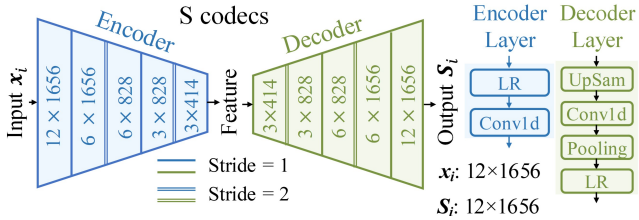

Fig. 3. S codecs structure. LR: Leaky ReLU, UpSam: UpSampling.

Secondly, the data are harmonized in the same format and fed into STNet. The raw data is augmented with location information extracted by spatial codecs (S codecs) composed of encoder and decoder. As illustrated in Fig. 3, both parts are consist of multilayer structures, the blue block being the encoder and the green one being the decoder. In this application, $\boldsymbol{x_i}$ (i=1, 2, …, 12) is the input data of length 1656. The encoder uses Leaky ReLU (4) as the activation function and extracts the feature by one-dimensional convolution (Conv1d). The original data are compressed by four layers to get one-sixteen$^{\text{th}}$ size regarded as features.

$$f(t) = \begin{cases} t, & t > 0 \\ kt, & t \le 0 \end{cases} \text{ where } k = 0.01 \quad (4)$$

The green decoder layer is employed to restore the data to the original size $\boldsymbol{S_i}$ (i=1, 2, …, 12), which contains an upsampling layer, a Conv1d layer, a pooling layer, and an activation function Leaky ReLU. The S codecs preprocesses the data. Specifically, the encoder compresses the data to extract single-sample features and amplify variations in the point cloud; the decoder upsamples to recover features back to their original size in preparation for spatio-temporal fusion.

Thirdly, the enhanced data $\boldsymbol{S_i}$ are concatenated with the original sequence $\boldsymbol{x_i}$ (5) and compressed by a fully connected layer. Thus, the fused data $\boldsymbol{a_i}$ is generated by combining the original temporal information with the enhanced spatial information.

$$\boldsymbol{a_i} = Concate(\boldsymbol{S_i}, \boldsymbol{x_i}) \quad (5)$$

S codecs enhance the spatial information of the original point cloud data, while weakening the timing-related info. Slip is essentially a continuous process and the temporal information cannot be ignored. The fusion emphasises spatial information while retaining sufficient temporal information.

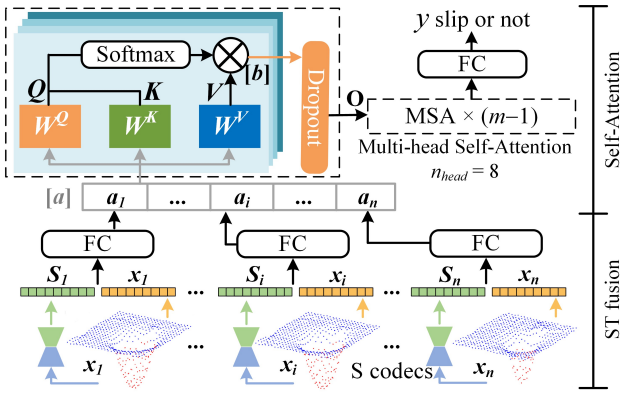

Fig. 4. Network structure of the proposed STNet for slip detection.

The fourth step involves aligning the fusion sequence into a series $[\boldsymbol{a}]$ and feeding it into the multi-head self-attention blocks (m in total). An MSA block is composed of 8 heads and a corresponding dropout layer. Fig. 4 exhibits the detailed mechanism of one head in the first MSA block. The remaining (m−1) MSA blocks is connected in series. The input $[\boldsymbol{a}]$ passes through the embedding layer, and then multiplied separately with the three matrices $\boldsymbol{W^Q}$, $\boldsymbol{W^K}$, and $\boldsymbol{W^V}$ to get $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ (6a). Then $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are computed through (6b) to get the output $[\boldsymbol{b}]$ of the current attention head. d_k in (6b) is the dimension of $\boldsymbol{Q}$ and $\boldsymbol{K}$. The output $\mathbf{O}$ of this MSA block is derived by concatenating the output of 8 heads. Then $\mathbf{O}$ is fed and processed by the following (m−1) MSA. After m MSA, the output sequence $[\boldsymbol{b}]$ maintains the same dimensions as $[\boldsymbol{a}]$. Finally, $[\boldsymbol{b}]$ is fed into the fully connected layer for binary classification to achieve slip detection.

$$\begin{bmatrix} \boldsymbol{Q} \\ \boldsymbol{K} \\ \boldsymbol{V} \end{bmatrix} = \begin{bmatrix} \boldsymbol{W^Q} \\ \boldsymbol{W^K} \\ \boldsymbol{W^V} \end{bmatrix} [\boldsymbol{a}] \quad (6a)$$

$$[\boldsymbol{b}] = \boldsymbol{V} \cdot \sigma\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right) \quad (6b)$$

where $\sigma(z) = e^z / \sum_{j=1}^{K} e^{z_j}$

MSA can consider the information contained in all sequences simultaneously, which greatly shortens the path of information propagation back and forth to improves the learning ability on long sequences. The 3D shape data obtained by visuo-tactile sensors are all long sequences, so potential exists to improve accuracy using MSA for slip detection. Meanwhile, 3D shape data with and without the process of the S codecs (spatial and temporal part,

respectively) is fused as the input of STNet to ensure that the network can focus on spatial information in the object-sensor contact region without losing temporal information.

III. Experiment and Result

In this section, the reliability and practicality of the STNet will be demonstrated. First, the dataset which involved objects of different shapes, materials and movement patterns was established by BVTS presented by Section II.A. Then the validity of STNet and effectiveness of the spatio-temporal sequences fusion mechanism are verified by experiments and ablation studies. Meanwhile, why this mechanism can improve slip detection accuracy is analyzed.

A. Setup and Dataset

A visuo-tactile sensor prototype is designed and fabricated based on the principle of BVTS. Data were collected with six easily accessible objects which varied in size, shape, and stiffness. The prototype and objects are shown in Fig. 5 (a) and Fig. 5 (b), respectively.

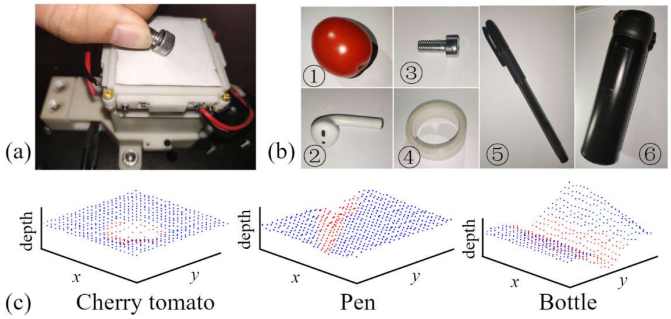

Fig. 5. Data collection platform. (a) The visuo-tactile sensor prototype. (b) Example objects for data collection: cherry tomato, earphone, screw, tape, pen, and bottle. (c) Three typical reconstructed 3D point clouds.

The binocular camera system begins to take pictures at a frame rate of 20Hz when the object comes into contact with the elastomer. The relative movement that may occur during gripping is manually controlled to ensure different motion patterns are precisely presented.

TABLE I. Sample Number and Dataset Composition

	Object	Stationary	Pressing	Rolling	Slipping
①	Cherry tomato	1787	1202	1290	1582
②	Earphone	1302	1597	1273	1708
③	Screw	1323	1595	1319	1642
④	Tape	1619	1641	1517	1515
⑤	Pen	1269	1095	1580	1580
⑥	Bottle	1372	1418	1367	1546
Total number		8672	8548	8346	9573
Dataset		**Training set**		**Test set**	
Proportion		80%		20%	

The state when no object is pressed against the elastomer and the state when no relative displacement occurs between the object and elastomer are denoted as stationary. Besides stationary, the motion of the object relative to the sensor was classified into three categories during data collection: dynamic press, rolling and slipping. These four motion patterns are all basic types of movement used in daily operations. Of all the movement patterns, only slips fall into the category of grasp failures. In order to make the network more focused on the judgment of slips, these movements are labeled into two categories, slips and non-slips. As an object keeps contact with the elastomer and moves in a single-motion form, the parallax images are recorded continuously. The 3D shape data were then reconstructed from the parallax images using BVTS. Three typical reconstruction results are demonstrated in Fig. 5(c).

To ensure all the samples are clear and easy to categorize, about 5% of the frames at the beginning or the end of each imaging process along with the problematic frames that may lead to labeling errors are omitted. The effective time duration of a single motion and single object is about 70s. To ensure efficient utilization of the existing video data and to recognize the slip as early as possible, 12 consecutive frames are extracted from the initial frame as one sample ($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ..., $\boldsymbol{x}_{12}$), and ($\boldsymbol{x}_2$, ..., $\boldsymbol{x}_{13}$) as another. The total number of samples for different movement modes of each object are listed in Table I.

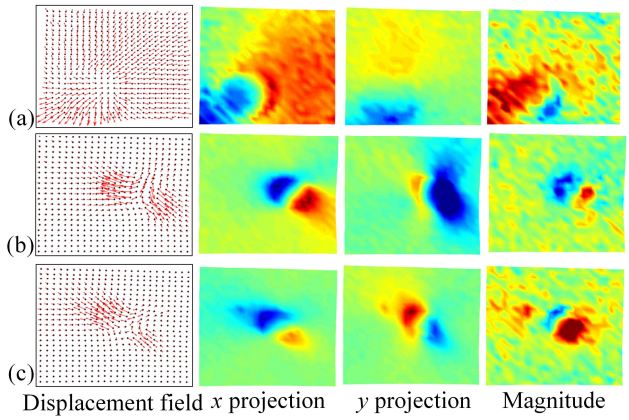

Fig. 6. Displacement field at the sensor contact surface (Colormap to value: Green: neutral, Red: positive, Blue: negative). (a) pressing. (b) rolling. (c) slipping.

In order to have a more intuitive understanding of the elastomer deformation with different motion patterns, samples from the dataset are first analyzed. The displacement field can be obtained by differencing the position of the corresponding point over successive frames. Fig. 6 shows the displacement fields of pressing, rolling, and slipping in the first column. The second and third columns are the projections of the displacement field in the x and y directions. Magnitude refers to the pressing depth. Two typical problems can be observed in Fig. 6:

① Fig. 6(a) exists pronounced vibration, which commonly occurs by the friction between the gripped object and the elastomer during the operation. Vibration-induced noise makes the displacement field complex and inhomogeneous, as shown in x-projection in Fig.6(a).

② The displacement fields of rolling and slipping are extremely proximate and difficult to judge rapidly even by the human visual, as illustrated in Fig. 6(b, c). The similarity between rolling and slipping makes it difficult to tell them apart.

The first problem, as we suggest in the introduction, is that rich shape data may interfere with slip detection, and the network needs to focus its attention more on regions of large magnitude. To address the second question, we first analyze how the displacement field in Figure 6 is generated. Zooming

into the first column of Fig. 6(b, c), when rolling, there are two opposite symmetrical displacements perpendicular to the forward direction close to the contact region. However, when slippage, the displacement on both sides is asymmetrical. The elastomer on the one side produces a larger displacement than the other side and more markers display displacement, as shown in Fig. 6(b, c). This is because during slippage, the object will create a pull on the elastomer, causing tension in the direction of motion and elasticity of the elastomer on the other. So there will be more displacement on the tension side. The ability to determine the difference in the displacement field becomes critical to perform accurate slip detection.

To summarize, a successful network is expected to be able to focus more attention on the part of the large displacement field that is in contact with the object, in order to discriminate between different motion patterns and at the same time avoid noise interference from vibrations.

B. Results

In this section, in order to verify that the spatio-temporal sequences fusion network structure is valid and reliable, the ablation experiment is conducted on the complete STNet network structure. Thus, the necessity of spatio-temporal information fusion is demonstrated.

The ablation experiment contains two experimental groups, full STNet structure and STNet removes S codecs (TemNet). The difference between these two models is that STNet utilizes the initial data twice ($\boldsymbol{x_i}$ and $\boldsymbol{S_i}$). While TemNet inputs the original sequence $\boldsymbol{x_i}$ directly into the MSA for attention detection. The inputs of both models are 12×1656 point cloud sequences with batch size 64.

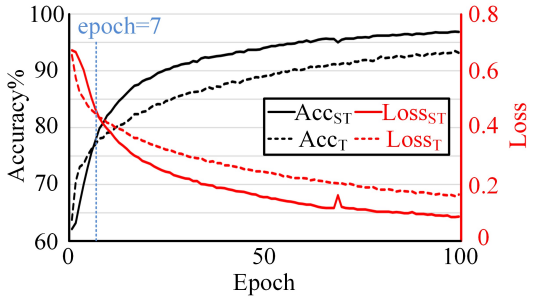

Fig. 7. The accuracy and loss of STNet and TemNet.

Figure 7 demonstrates the accuracy and the loss of the models focusing on the prediction accuracy for the slip category. In the beginning, TemNet training is slightly better, but after a few epochs, STNet overtakes it. Initially, STNet needs to manipulate the complex information that has been encoded and enhanced, which results in a lower accuracy rate compared to TemNet. After that, both have increasing accuracy and decreasing losses, with accuracy above 90% and losses below 0.2 after 100 timings. Rolling accounts for a third in the non-slip category, but the majority of the samples are still able to be correctly categorized, indicating that STNet is able to effectively distinguish between two similar motion patterns.

In order to evaluate the performance of the proposed models more comprehensively and precisely, we compared the precision, recall, and F1 scores, which represent statistical measures of the accuracy of the binary classification models. Table II lists the results of the two structures trained on the same dataset. The simpler structure TNet is also capable of achieving very high precision (96.97%) even after ablation of the S codecs. The MSA recognizes most of the slip samples, but its recall still needs to be enhanced. By reinforcing the location information through a combination of encoder and decoder, fusing it with the timing signal, and then feeding it into the training, the final recall obtained is improved by nearly 6.7% compared to the lower one, 86.39%, and the precision is also improved by about 2%. The resulting F1 score is also elevated to 95.37%. Therefore, it can be adequately clarified that location-coded data enhancement (S codecs) with information fusion can effectively enhance the accuracy and recall of slip detection.

TABLE II. SLIP DETECTION ACCURACY OF DIFFERENT STRUCTURE

Structure	Precision(%)	Recall(%)	F1 score(%)
TemNet	96.97	86.39	89.56
STNet	**98.91**	**93.06**	**95.37**

All models were constructed via PyTorch and model training was executed on an Nvidia GeForce 3090 (24 GB graphics memory). The batch size of the network structure shown in this paper is 64, the network weights are initialized by Xavier normal distribution, the number of attention-head is 8, and cross-entropy loss is used as the loss function. Network parameters were updated by Adam optimizer with a learning rate of 1×10^{-5}. For more detailed settings and source code for the models, please refer to https://github.com/jinLhust/STNet.

C. Analysis

From the results presented in Section III. B, the high accuracy of slip judgment indicates the ability of STNet to settle the redundant noise brought by massive dense data and to differentiate between two similar types of motion, slipping and rolling. Meanwhile, the self-attention module has good interpretability, so in this section, the reason why STNet can achieve better classification for different samples is analyzed. Three representative sets of data are shown in Fig. 8 to illustrate the effects and advantages of STNet more intuitively.

Each set of images comes from the same sample, drawing frame by frame makes it difficult to detect the difference between each frame even through human eyes. Therefore, one image is taken every three frames to visualize the motion pattern better. The first two sets show the slippage of a cherry tomato (small contact area) and a pen (strip contact face), and the last one shows rolling with a larger contact area (bottle). Each subplot contains three rows, from top to bottom, the externally captured scene, the deformation field of the elastomer $\boldsymbol{x_i}$ (i=2, 5, 8, 11), and the self-attention heat map **[*b*]**.

Among them, in the second row's deformation field images, green is the neutral plane, blue is negative, and yellow is positive. Due to the large amount of deformation of the flexure, the reconstructed image is not a perfect flat square. The third row is the attentional heat map which is composed of the data obtained after the MSAs, with redder colors indicating higher attentional weights. Analyzing Fig. 8, STNet can solve the two problems raised in III. A reasonably.

In Fig. 8, three figures show three objects with small to large contact areas to verify the classification of the motion state by the sensor under different working conditions. The attention heat map shows a significant increase in attention to the front and back ends of the contact surfaces when the object slips or rolls. Even if the object just contacts the boundary of the flexible body and fails to map the complete deformation

field, the reconstructed attention weights are able to match the front and back ends of the deformation field pressing region well, no matter slipping or rolling, which is consistent with the analysis of the deformation field in Fig. 6(c). This further validates the reliability of STNet in classifying highly similar motion patterns on different scales.

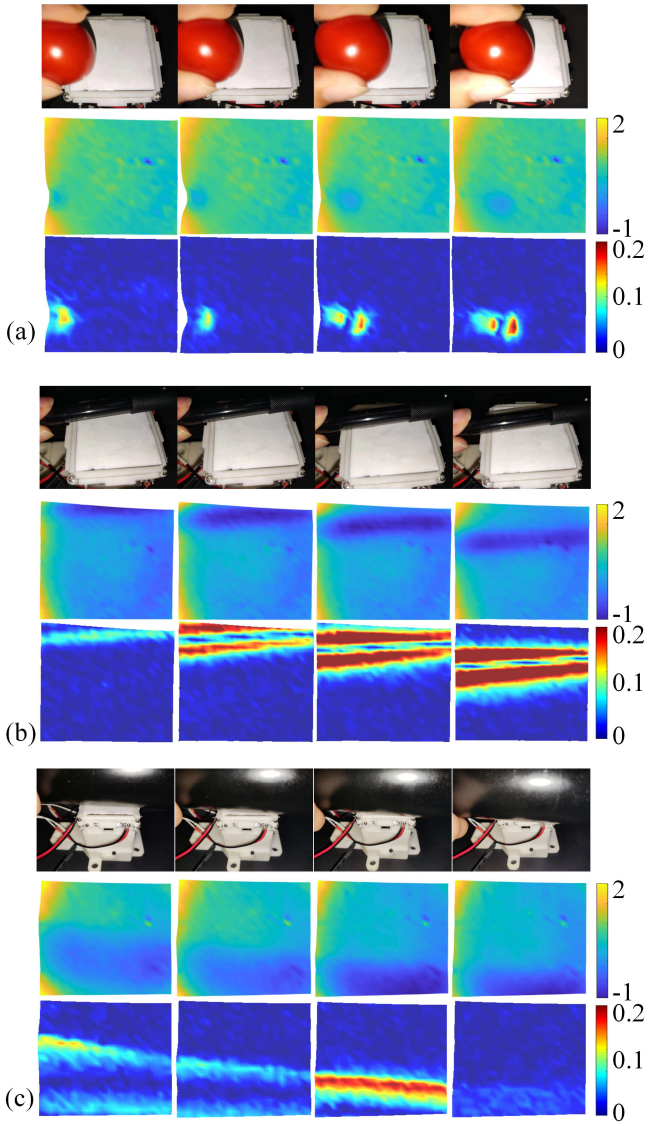

Fig. 8. Examples, 1st row: realistic scene, 2nd row: deformation field (mm), 3rd row: the weight of attention. (a) Cherry tomato slipping. (b) Pen slipping. (c) Bottle rolling

Also, there are yellow patches in the deformation field images of all the scenes which are vibration noises generated on the surface of the elastomer during object contact. However, compared to the area of attention the vibrations at the remaining locations are substantially suppressed. It indicates that STNet has a strong robustness to identify the disturbed and the truly attentive part.

IV. Conclusion

In this paper, a new self-attention based architecture, STNet, is proposed for slip detection in visuo-tactile sensors. The input of STNet is consist of spatial part and temporal part, which refers to 3D shape data collected by visuo-tactile sensors with and without the process of an encoder-decoder module, respectively. The two parts are fused as the input to ensure that the network can focus on spatial information in the object-sensor contact region without losing temporal information. A BVTS prototype is designed and fabricated for dataset construction. The dataset for STNet training contains four motion patterns, including stationary, pressing, rolling and slipping, are classified into two categories, slip or not. The experiment results indicate STNet can make high-precision slip detection by focusing on the object-sensor contact area and neglecting the potential vibration in the remaining region. Meanwhile, the ablation studies confirm the effectiveness of the spatio-temporal sequence fusion mechanism, which can effectively improve the F1 score to 95.37% (5.81% higher than the algorithm without codecs).

These numerous samples of indeterminate motion forms provide a broader validation space for STNet. We will expand the structure of this network in the future to distinguish between multiple classes of motion patterns, providing a more reliable algorithmic basis for robotics dexterous manipulation.

References

[1] G. Westling, and R. S. Johansson, "Factors influencing the force control during precision grip", *Experimental Brain Research*, vol. 53, pp. 277-284, 1984.

[2] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile sensing—from humans to humanoids", *IEEE Trans. on Robotics*, vol. 26, no. 1, pp. 1-20, 2009.

[3] R. A. Romeo, and L. Zollo, "Methods and sensors for slip detection in robotics: A survey". *IEEE Access*, vol. 8, pp. 73027-73050, 2020.

[4] A. Ikeda, and et al., "Grip force control for an elastic finger using vision-based incipient slip feedback", *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pp. 810-815, 2004.

[5] C. Melchiorri, "Slip detection and control using tactile and force sensors", *IEEE/ASME Trans. on Mechatronics*, vol. 5, no. 3, pp. 235-243, 2000.

[6] M. T. Francomano, D. Accoto, and E. Guglielmelli, "Artificial sense of slip—A review," *IEEE Sensors J.*, vol. 13, no. 7, pp. 2489–2498, 2013

[7] A. C. Abad, and A. Ranasinghe, "Visuotactile sensors with emphasis on gelsight sensor: A review", *IEEE Sensors J.*, vol. 20, no. 14, pp. 7628-7638, 2020.

[8] M. Li, T. Li, and Y. Jiang, "Marker displacement method used in vision-based tactile Sensors—from 2D to 3D-a review". *IEEE Sensors J.*, vol. 23, pp. 8042-8059, 2023.

[9] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a GelSight tactile sensor," *IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 304–311, 2015

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

[11] S. Dong, W. Yuan, and E. H. Adelson, "Improved GelSight tactile sensor for measuring geometry and slip," *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, pp. 137–144, 2017.

[12] R. Kolamuri, Z. Si, Y. Zhang, A. Agarwal, and W. Yuan, "Improving grasp stability with rotation measurement from tactile sensing", *IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2021.

[13] J. Li, S. Dong, and E. Adelson, "Slip detection with combined tactile and visual information," *arXiv preprint arXiv:1802.10153*, 2018.

[14] Y. Zhang, Z. Kan, Y. A. Tse, Y. Yang, and M. Y. Wang, "Fingervision tactile sensor design and slip detection using convolutional lstm network", *arXiv preprint arXiv:1810.02653*, 2018.

[15] S. Cui, S. Wang, R. Wang, S. Zhang, and C. Zhang, "Learning-based slip detection for dexterous manipulation using GelStereo sensing." *IEEE Trans. on Neural Networks and Learning Systems*, pp. 1-10, 2023.

[16] A. Vaswani and et al., "Attention is all you need." *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] H. Ma, J. Ji, and K.M. Lee, "Effects of refraction model on binocular visuotactile sensing of 3-D deformation", *IEEE Sensors J.*, vol. 22, no. 18, pp. 17727-17736, 2022.

[18] S. Birchfield, "Derivation of kanade-lucas-tomasi tracking equation", unpublished notes, 1997.