Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.083466

ARTICLE

STR-CMFNet: A Visual-Tactile Spatio-Temporal Rectification and Cross-Modal Fusion Network for Slip Detection

Hao Chu, Xibin Xiao, Song Gao, Chao Zheng and Fei Wang*

Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China

*Corresponding Author: Fei Wang. Email: wangfei@mail.neu.edu.cn

Received: 04 April 2026; Accepted: 16 June 2026; Published: 13 August 2026

ABSTRACT: In recent years, embodied intelligence has become an important research direction. Slip detection is a critical challenge in dexterous robotic manipulation, especially in manipulating deformable objects such as soft fruits. Visual–tactile fusion can provide rich sensory information, but it also introduces significant modality differences. We propose STR-CMFNet, a novel network framework for visual–tactile slip detection. The framework adopts a two-stage design. It consists of Spatio-Temporal Rectification (STR) and Cross-Modal Fusion (CMF). The STR applies temporal attention weights to recalibrate key-frame features. It also refines spatial feature maps. Based on the corrected features, the CMF module is further introduces. The global dependencies between visual and tactile features is modeled through a cross-attention mechanism to achieve effective cross-modal interactions and allows complementary information. Experiments are conducted on a public benchmark dataset and an extended fruit dataset. The results show that STR-CMFNet outperforms existing methods. In particular, while ViViT achieves a classification accuracy of 78.5%, the proposed method achieves 83.1% accuracy, showing a clear and consistent performance improvement. The ablation study further validates the effectiveness of each module. Overall, we provide a practical solution for stable robotic grasping of deformable objects, and demonstrate strong potential in complex interaction scenarios.

KEYWORDS: Slip detection; visual-tactile fusion; spatio-temporal rectification; cross-modal fusion; robotic grasping

1 Introduction

In recent years, embodied intelligence has become a important research focus in artificial intelligence. As a core component of embodied intelligence, dexterous robotic manipulation at the end-effector level has attracted increasing attention [1]. A key challenge is enabling robotic end-effectors to stably grasp soft and fragile objects, such as fruits, eggs, and water-filled disposable paper cups. If the grasping force is too small, the object may slip, and the grasp may fail. If the force is too large, it may damage the object. In addition, the deformation of soft objects during grasping increases the uncertainty. Slip detection is essential for safe manipulation. This allows the system to detect whether the object is slipping. This provides useful feedback to the end-effector. The end-effector can then adjust the grasp in real time and improve force control.

Early research focused primarily on vision-based approaches to evaluate grasp stability [2]. Although visual sensing offers global surface cues, it is often insensitive to local deformations at the contact region, which can lead to mismatches between perceived and actual slip states. Inspired by the sensing mechanisms of human skin, researchers introduced tactile-based methods for slip detection [3,4], which improved sensitivity to small-scale motions and contact variations. However, tactile-only methods struggle when visual context is needed to interpret ambiguous contact events.

With the advancement of multimodal learning, recent studies have explored visual-tactile fusion to leverage the strengths of both modalities. For example, Calandra et al. [5] demonstrated improved grasp performance using a multimodal perception framework, while Li et al. [6] employed a CNN-LSTM architecture to capture spatial and temporal slip features. The Vision-guided tactile frameworks [7] further enabled effective interaction with transparent or visually challenging objects. These frameworks demonstrate promising performance in complex grasping scenarios. However, most existing works adopt early fusion strategies, combining raw features before higher-level processing. Visual and tactile data are inherently heterogeneous. Early fusion often distorts the contribution of each modality, thereby limiting the overall robustness of the system.

To address these limitations, late or hybrid fusion strategies have been explored. Gao et al. [8] proposed a CNN-MSTCN model. The model learns visual and tactile representations separately before fusion. Fusion is performed after modality-specific feature extraction. Other studies proposed more advanced architectures, including Temporal Convolutional Networks (TCNs) [9], Transformer-based models [10], hierarchical fusion [11], and ViViT-based grasping frameworks [12]. These approaches further improve multimodal reasoning and representation learning. In particular, Li et al. [13] introduced VITO-Transformer, a Transformer-based fusion network for visual–tactile object recognition, which significantly improved performance through a specialized fusion mechanism. These works collectively highlight the growing importance of Transformer architectures in capturing cross-modal dependencies.

Recently, some researchers have shared robot grasping datasets [12,14,15]. Meanwhile, the development of vision-based tactile sensors (VBTS) has promoted visual-tactile multimodal grasping research. Examples include Gelsight [16], GelSim [17], and Digit [18]. However, tactile data acquisition and representation depend heavily on the sensor type. It is difficult to generalize across different sensors. In this study, we use the Digit tactile sensor and RealSense camera to conduct various pinch and lift experiments on common soft fruits. We collect tactile and visual data from these experiments.

Although recent works have made significant progress in visual-tactile fusion, challenges remain. The information gap between vision and touch modalities is large. Existing late fusion methods lack early feature interaction mechanisms. Therefore, we proposes a cross-modal spatiotemporal feature interaction module. This module calibrates bimodal features by leveraging spatial and temporal correlations. It helps both modalities focus more on complementary cues from each other. It also reduces uncertainty and noise from different modalities.

Transformer models have strong modeling abilities but lack inductive bias. This makes them prone to overfitting [19]. Inspired by the large receptive field from self-attention [20], we integrates a cross-attention mechanism. We enhance ViViT's multi-head self-attention and develop a cross self-attention module. This module captures effective features at different levels and establishes global dependencies.

Consequently, the main contributions of the study are summarized as follows:

1. Visual-Tactile Spatiotemporal Interaction Module: We propose a spatiotemporal feature interaction module. The module enables early-stage fusion between visual and tactile modalities. It enhances the richness of extracted features.
2. Cross-Attention Feature Extraction: We develop a visual-tactile cross-attention mechanism. It supports global reasoning across modalities. It also enables mutual guidance between visual and tactile features. The mechanism improves cross-modal feature alignment.
3. Soft Fruit Grasping Dataset: We construct a novel dataset of soft fruits grasping interactions. The dataset reflects real-world robotic control tasks and practical manipulation challenges.

The remainder of the paper is organized as follows: Section 2 reviews related work. Section 3 describes the proposed model architecture and key components. Section 4 introduces the dataset and experimental setup. Section 5 presents the experimental results and provides detailed analysis. Section 6 concludes the paper and discusses future directions.

2 Related Work

2.1 Slip Detection

Grasp stability evaluation is a key prerequisite for achieving optimal grasp strategies, including minimum-force grasping. Humans rely on slip-sensitive feedback to manipulate objects dexterously [21]. Without tactile feedback, maintaining a stable grasp is very challenging for humans. Similarly, tactile sensing is critical for robotic manipulation. Slip occurs when the contact force is insufficient. Robots must adjust their grasping plans automatically in these situations.

Traditional slip detection methods [22,23] use conventional numerical techniques and physical models to estimate grasp stability. These methods rely heavily on accurate models and parameters. Machine learning offers effective alternatives. Hyttinen et al. [24] discretized the graspable object space into prototype shapes using data-driven clustering. They used kernel logistic regression to classify grasp stability based on tactile data for each prototype. Zhao et al. [25] put forward a tactile-only slip detection framework. It monitors friction-grip force ratio and tactile signal frequency shift from stable grasping to slipping, and adjusts finger gripping current adaptively. The method realizes calibration-free slip recovery with force control, unaffected by object shapes. Si et al. [26] and Xu et al. [27] trained CNN-LSTM models to predict grasp outcomes from sequential tactile inputs, with the latter further enhancing real-time performance on dexterous hands through dynamic force compensation by capturing spatiotemporal slip dynamics. Zapata-Impata et al. [28] noted that simple stability detection fails to identify slip types and directions. They proposed a ConvLSTM network to classify slip direction by analyzing spatiotemporal tactile features. Li et al. [29] introduced the WACSAN model, which uses weight allocation and cross self-attention. Weight allocation assigns different importance to regions of the feature map. Cross attention extracts local information from neighboring pixels. This improves slip detection when objects move along the gripper edge.

2.2 Visual-Tactile Fusion Perception

In recent years, multimodal feature fusion networks have gained increasing attention. These methods extract information separately from visual and tactile modalities, and then fuse the features to perform slip detection.

Zhang et al. [30] proposed an attention-guided visual-tactile fusion architecture using a cross-modal transformer. Their model processes visual and tactile inputs to predict grasp stability but handles only single-frame data, ignoring temporal dynamics essential for slip detection. Yan et al. [9] proposed a CNN-TCN network to fuse visual and tactile information for slip detection. Their experiments demonstrated that this model outperformed CNN-LSTM when using various pre-trained visual networks. However, the feature fusion design remains limited: visual and tactile features are extracted separately and concatenated without explicit cross-modal learning. Huang et al. [31] proposed an attention-based visuo-tactile fusion model using MS-TCN, achieving 98.98% accuracy in grasp tasks. Similarly, Zheng and Chen [32] introduced a grasp stability prediction module that mimics human behavior, where tactile perception guides visual grasp detection. Cui et al. [33] used a 3D CNN to extract spatiotemporal features from XELA tactile and RGB data, achieving 99.97 accuracy. However, the increased parameters slow training and inference, complicating real-time grasping. Han et al. [12] proposed a Transformer-based framework for rigid grippers using visual and

tactile inputs to grasp deformable objects. Their model outperformed CNN+LSTM on a public dataset [6]. However, existing methods use basic feature-level fusion, limiting complementary information use and ignoring modality interactions.

This paper proposes an attention-guided cross-modal fusion architecture. It aims to fully integrate visual and tactile features for better grasp performance.

3 Proposed Model

3.1 Problem Statement

Our goal is to determine the final grasp state of flexible fruits—*slip*, *safe grasp*, or *over-force*—based on visual-tactile modality fusion.

$$\hat{y} = \arg \max_{c \in \{0,1,2\}} \mathcal{F}_{FC}(X_{final} \parallel P_{force}) \tag{1}$$

where the class label $c = 0$ denotes a slip event (Slip), $c = 1$ represents a stable and secure hold (Safe Grasp), and $c = 2$ indicates excessive deformation or squeezing force (Over-force). A slip is explicitly triggered and reported when $\hat{y} = 0$.

First, a Feature Refinement Module (FRM) is used to exchange information between the two modalities along temporal and spatial dimensions. This allows the model to capture complementary information. Second, the Cross-Attn is applied during feature extraction to build global dependencies between visual and tactile features. All low-level features are then concatenated. Finally, the concatenated features and force control parameters are fed into a fully connected (FC) layer to produce the output. The output is defined as a three-class classification: 0 for slip, 1 for safe grasp, and 2 for over-force.

3.2 Slip Detection Framework

As shown in Fig. 1, the slip detection framework consists of three main components: FRM, Cross-Attn, and a prediction module.

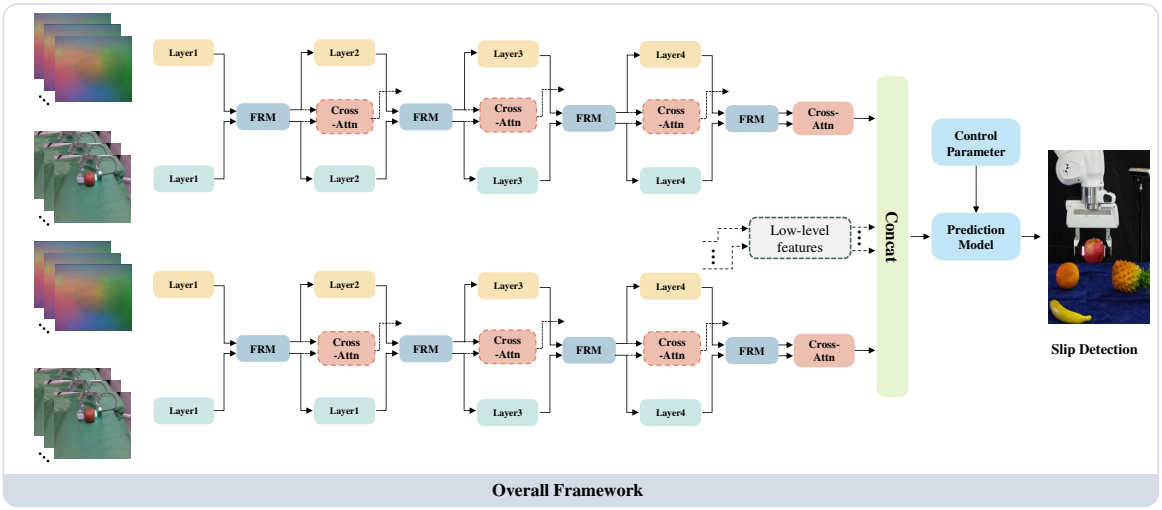

Figure 1: Overview of grasping framework. The robot first performs two explorative actions: 1). Pinching the fruit, 2). Sliding along the fruit surface in the optical axis. Visual and tactile features are extracted in parallel, then corrected and fused via FRM and Cross-Attn. The fused features and force parameters are used to classify the grasping state into slip, stable grasp, or over-force.

Two predefined actions are used. Each action employs two parallel branches to extract features from visual and tactile inputs. A cross-branch interaction mechanism is introduced, where features from one modality are corrected based on features from the other modality. Additionally, at each stage of the dual-branch architecture, the corrected features from both modalities are exchanged to promote cross-modal feature interaction.

This framework uses a dual-branch structure. It leverages the complementary nature of visual and tactile modalities to enhance slip detection performance. Features from each modality may contain noise. The other modality provides cues to calibrate and correct these errors.

As illustrated in Fig. 2, the FRM aligns and refines both visual and tactile features by leveraging inter-modality correlations. It is inserted between adjacent network stages, allowing the corrected features to be passed forward for deeper feature extraction.

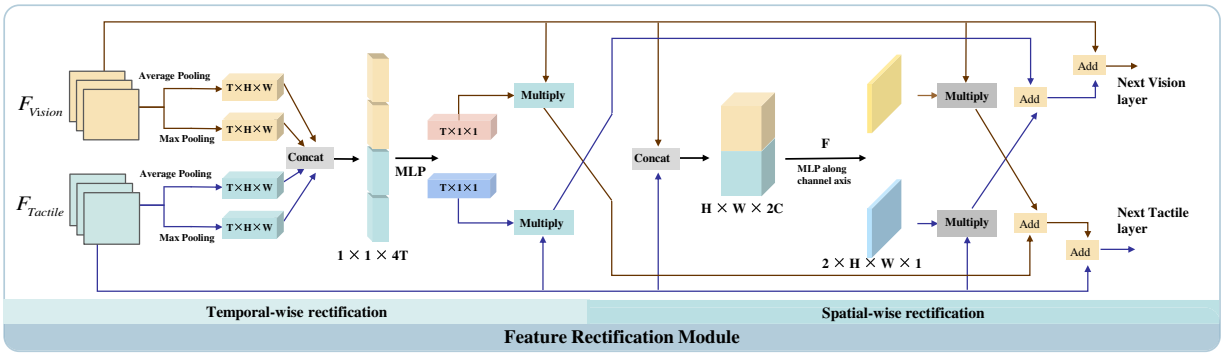

Figure 2: Structure of FRM. The FRM performs cross-modal feature correction along temporal and spatial dimensions. Temporal-wise rectification learns complementary attention weights from the other modality across the time axis, while spatial-wise rectification applies local feature calibration using 1×1 convolutions and sigmoid activation. The corrected features are combined with the original input using weighted residual connections.

Furthermore, as shown in Fig. 3, we design Cross-Attn to fuse visual and tactile information. After encoding, the input images are divided into patches and transformed into token sequences. These token embeddings from both modalities are then processed by a multi-head cross-attention module to capture the dependencies between the modalities.

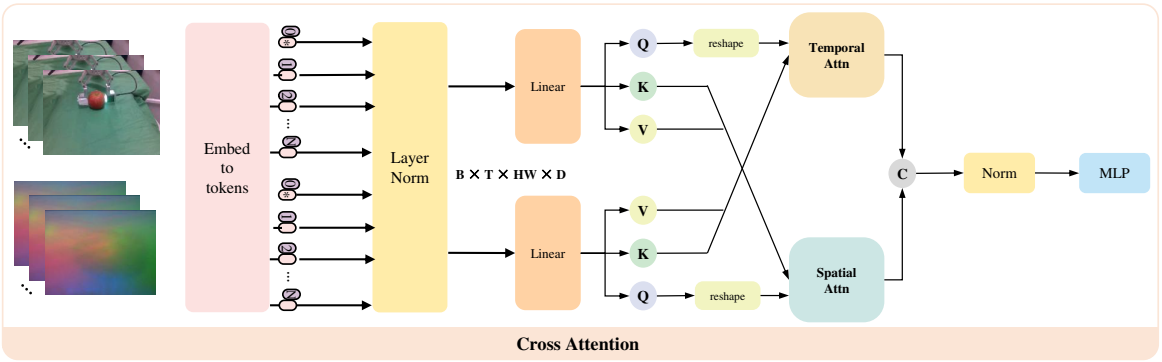

Figure 3: Cross-attn module. Both visual and tactile inputs are partitioned into uniform patches and embedded into high-dimensional feature sequences. The Cross-Attn is then applied. visual queries attend to tactile keys-value pairs, and tactile queries attend to visual key–value pairs. The updated features are concatenated and projected into a unified representation.

3.3 Cross-Modal Feature Rectification

As discussed above, visual and tactile modalities usually provide complementary information. Visual data can help disambiguate tactile signals but often contains measurement noise. This noise can be filtered and calibrated using features from the other modality.

To address this issue, we design a Cross-Modal Spatiotemporal Feature Rectification Module (FRM), as shown in Fig. 2. This module performs feature alignment between the two parallel streams before each stage of feature extraction. To reduce the effects of noise and uncertainty across modalities, the FRM operates in two dimensions: temporal and spatial. This dual-axis calibration supports better multimodal feature interaction and extraction.

3.3.1 Temporal-Wise Feature Rectification

To mitigate the inherent modal discrepancies and suppress measurement noise, the Feature Rectification Module (FRM) establishes an early-stage spatiotemporal calibration mechanism. Formally, we embed the dual-modal input features $\text{Vision}_{in} \in \mathbb{R}^{T\times C\times H\times W}$ and $\text{Tactile}_{in} \in \mathbb{R}^{T\times C\times H\times W}$ along the temporal axis. Unlike previous attention-based temporal models that compress frames naively, our temporal FRM applies both global max pooling (GMP) and global average pooling (GAP) along the temporal dimension to both streams simultaneously. This dual-pooling strategy effectively preserves both the salient peak stimuli of transient slips and the macro contextual continuity of the grasping process.

The four resulting feature vectors are flattened and concatenated to construct a joint spatiotemporal descriptor $Y \in \mathbb{R}^{4T}$. Subsequently, a multi-layer perceptron (MLP) processes Y to capture inter-modality temporal correlations, which is then mapped through a sigmoid activation function to generate the rectified temporal attention weights $W^T \in \mathbb{R}^{2T}$. Finally, a splitting operator divides W^T into distinct modality-specific attention vectors:

$$W^T_{vision}, W^T_{tactile} = \mathrm{T}_{split}(\mathrm{sigmoid}(\mathrm{T}_{mlp}(Y))) \tag{2}$$

where $W^T_{vision} \in \mathbb{R}^T$ and $W^T_{tactile} \in \mathbb{R}^T$. Using these mutually derived weights, the cross-modal temporal rectified features are computed by modulating one modality with the temporal gating vector of the opposite stream:

$$\begin{aligned} \text{Vision}^T_{rec} &= W^T_{tactile} \odot \text{Tactile}_{in} \\ \text{Tactile}^T_{rec} &= W^T_{vision} \odot \text{Vision}_{in} \end{aligned} \tag{3}$$

where the operation $\odot$ denotes broadcasting temporal-wise element-wise multiplication applied sequentially across all time frames.

3.3.2 Spatial-Wise Feature Rectification

While the temporal rectification branch aligns frame-level sequential correlations, local contact geometries and micro-vibrations require precise spatial calibration. Therefore, a parallel spatial FRM is introduced. The bimodal inputs Vision_{in} and Tactile_{in} are concatenated along the channel dimension and fed into a dual-layer spatial embedding network. This embedding block comprises two 1×1 convolutional layers assembled with a ReLU activation function to preserve local spatial configurations.

By passing the output through a sigmoid function, we obtain a structural embedded weight map $F \in \mathbb{R}^{H\times W\times 2}$, which encapsulates the local spatial boundaries of both object appearance and tactile contact imprints. The splitting function then divides F into two separate spatial weight charts:

$$F = \mathrm{Conv}_{1\times1}(\mathrm{ReLU}(\mathrm{Conv}_{1\times1}(\mathrm{Vision}_{in} \parallel \mathrm{Tactile}_{in}))) \tag{4}$$

$$W^S_{vision}, W^S_{tactile} = \mathrm{T}_{split}(\mathrm{sigmoid}(F)) \tag{5}$$

where $W^S_{vision} \in \mathbb{R}^{H\times W}$ and $W^S_{tactile} \in \mathbb{R}^{H\times W}$. Symmetrical to the temporal branch, the spatial-wise feature rectification applies the spatial map of one modality to calibrate the local feature representation of the other:

$$\begin{aligned} \mathrm{Vision}^S_{rec} &= W^S_{tactile} * \mathrm{Tactile}_{in} \\ \mathrm{Tactile}^S_{rec} &= W^S_{vision} * \mathrm{Vision}_{in} \end{aligned} \tag{6}$$

where $*$ denotes spatial-wise element-wise multiplication. Finally, the temporal and spatial rectified streams are aggregated with the original inputs via residual connections to yield the fully calibrated features Vision_{out} and $\mathrm{Tactile}_{out}$:

$$\begin{aligned} \mathrm{Vision}_{out} &= \mathrm{Vision}_{in} + L_T * \mathrm{Vision}^T_{rec} + L_S * \mathrm{Vision}^S_{rec} \\ \mathrm{Tactile}_{out} &= \mathrm{Tactile}_{in} + L_T * \mathrm{Tactile}^T_{rec} + L_S * \mathrm{Tactile}^S_{rec} \end{aligned} \tag{7}$$

where L_T and L_S are hyperparameters balancing the two rectification axes, both set to 0.5 by default to ensure an equal contribution of time and space constraints.

3.4 Cross-Attention Mechanism for Feature Extraction and Fusion

To capture meaningful features from visual and tactile images, extract local structures, and provide rich information for subsequent attention mechanisms, we employ a Uniform Patch Embedding method. The method first divides each image frame into fixed-size patches, and then maps these patches into a high-dimensional embedding space using convolution operations. The resulting representation is structured and suitable for subsequent modules. Specifically, the input image has dimensions (B, C, T, H, W), where B is the batch size, C is the number of channels, T is the number of frames, and H and W are the height and width of image. Convolution operations divide the image into num-patches $= \frac{H}{P_h} \times \frac{W}{P_w}$ patches, where Ph and Ph are the height and width of each patch. The patches are then flattened into a matrix of size $(B \bullet T, N, D)$, where N is the number of patches, and D is the embedding dimension of each patch. The same processing method is applied to tactile images, which also have dimensions (B, C, T, H, W). These are transformed into corresponding feature representations through the same convolution operations. These features are then fed into the cross-attention module, which enhances the correlation between visual and tactile features and provides rich spatial and temporal information for downstream tasks:

$$x_v, x_t = \mathrm{PatchEmbed}(x_v, x_t). \tag{8}$$

To better integrate visual and tactile modalities, we adopt a cross-attention mechanism. The design places greater emphasis on tactile information. This mechanism enhances feature fusion by enabling interaction between visual and tactile features. Unlike traditional attention, which focuses on a single modality, cross-attention uses visual features as queries and tactile features as keys and values. This allows the model to perform attention computation across modalities and improves the quality of the fused representation.

First, visual and tactile features are transformed separately to generate queries (Q), keys (K), and values (V) using linear projections. Specifically, the visual feature x_v is passed through a linear layer to obtain its Q, K, V. Similarly, the tactile feature x_t is processed by another linear layer to produce its corresponding Q, K, V. The process is formulated as follows:

$$\begin{aligned} q_v &= W_q x_v, \quad k_v = W_k x_v, \quad v_v = W_v x_v, \\ q_t &= W_q x_t, \quad k_t = W_k x_t, \quad v_t = W_v x_t. \end{aligned} \tag{9}$$

In the cross-attention mechanism, the query vector from the visual modality is matched with the key-value vectors from the tactile modality. Specifically, the visual query q_v and the tactile key k_t are used to compute the attention matrix via dot product. This matrix is then normalized using the softmax function to obtain the attention weights:

$$\text{Attn}_v = \text{softmax}\left(\frac{q_v k_t^T}{\sqrt{d}}\right), \tag{10}$$

here, d is the dimension of the key vectors used for scaling. Then, the attention weights are multiplied by the tactile value vectors v_t to obtain the updated information for the visual modality:

$$x_v^{\text{updated}} = \text{Attn}_v v_t. \tag{11}$$

Similarly, the cross-attention between the tactile query q_t and the visual key k_v is computed in a similar manner:

$$\text{Attn}_t = \text{softmax}\left(\frac{q_t k_v^T}{\sqrt{d}}\right), \tag{12}$$

then, the updated tactile feature is:

$$x_t^{\text{updated}} = \text{Attn}_t v_v, \tag{13}$$

finally, the updated visual and tactile features are fused through a projection layer. Specifically, the updated visual and tactile features are concatenated and then passed through a linear transformation to restore the original dimension:

$$x = \text{concat}\left(x_v^{\text{updated}}, x_t^{\text{updated}}\right) \quad \Rightarrow \quad x_{\text{final}} = W_{\text{proj}} x. \tag{14}$$

Through this cross-attention mechanism, visual and tactile information are better integrated. This effectively improves the model's performance in grasp force estimation.

4 Experimental Datasets and Setups

4.1 The Dataset Introduction

In this study, we use the publicly available vision-tactile grasping dataset introduced in [12], which includes five types of fruits: apples, oranges, tomatoes, lemons, and plums, as shown in Fig. 4. Each fruit is grasped under different force thresholds. Since tactile data requires direct contact with the object, the dataset collects tactile signals through two predefined actions: pinching and sliding on the fruit surface. At the same time, an external camera captures visual images in real time. The grasping outcomes are classified into three types: slip, safe grasping, and damage. Due to differences in fruit hardness and surface texture, force thresholds vary across fruit types. To maintain a balanced distribution of the three outcome categories, the threshold ranges are determined empirically. For example, the force thresholds for apples are sampled as integers from 4 to 16.

Due to the limited size of the public dataset and the poor generalization of data collected under different experimental conditions, we expands the original dataset by including additional fruit types, as illustrated

in the Fig. 4. In addition to force thresholds, we also consider the width of the gripper's end-effector. The model is first trained using the public dataset. Then, pretrained weights are loaded and updated using the extended dataset.

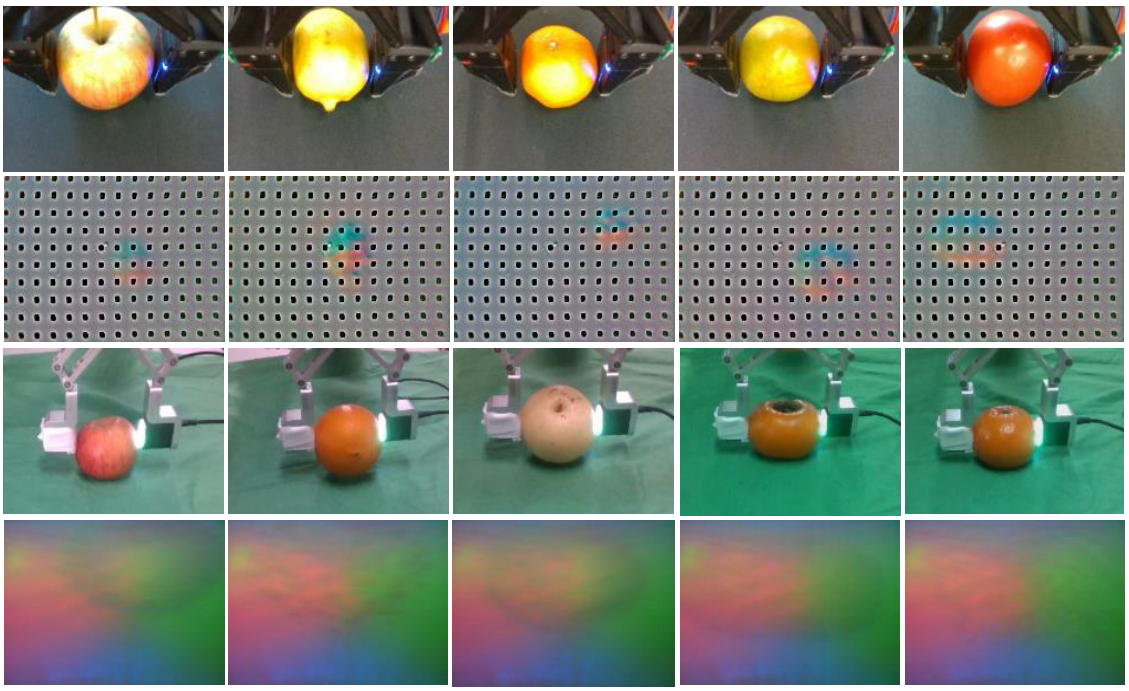

Figure 4: Visual and tactile samples from public and extended fruit grasping datasets. Top row: RGB images from the public dataset, showing (from left to right): apple, lemon, orange, plum, tomato. Second row: corresponding Gelsight tactile images collected at the final pinching frame. Third row: RGB images from the extended dataset, showing (from left to right): apple, orange, pear, persimmon, tangerine. Bottom row: corresponding tactile images collected using Digit sensors. The visual and tactile differences reflect the diversity in fruit hardness, surface texture, and deformation characteristics across datasets.

4.2 Experimental Setup

The visual-tactile dataset were collected using an Intel RealSense camera and a DIGIT tactile sensor, operating at 60 and 30 Hz, respectively. The original resolutions were 640 × 480 for visual data and 200 × 150 for tactile data. The experimental setup is shown in the Fig. 5. To reduce computational cost, visual images were resized to 160 × 120. For the patch embedding module, fixed patch sizes of 16 × 12 for visual images and 20 × 15 for tactile images were used.

The model consists of four layers with embedding dimensions of 64, 128, 256, and 512. The number of layers is set to 8, and the number of attention heads is 16. During training on all datasets, we applied data augmentation techniques, including random horizontal flipping, color jitter, and Gaussian blur, to both visual and tactile data.

Training was conducted using the AdamW optimizer with a weight decay of 0.01 and an initial learning rate of 1×10^{-4}, combined with a learning rate scheduler. The cross-entropy loss function was used for optimization. To evaluate slip detection performance, we used success rate, F1-score, recall, and inference time as the main evaluation metrics.

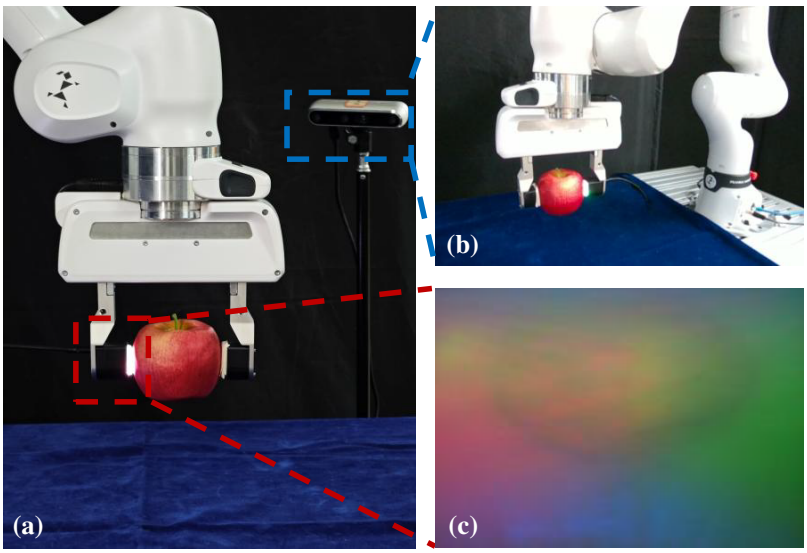

Figure 5: (**a**) Experimental setup: Franka robot arm and parallel gripper. One finger of the gripper is replaced with a Digit sensor. An external camera (RealSense) is mounted on the outside of the arm. The gripper is holding a toy bear. (**b**) Image captured by the external camera. (**c**) Image from the Digit sensor.

5 Experiment and Analysis of Results

5.1 Comparative Experiments

A series of baseline experiments were conducted to compare the proposed method with existing approaches. These experiments were performed on both the public dataset and the extended dataset. For each dataset, all methods were evaluated under the same experimental setup. The results are summarized in Table 1. The results show that our proposed method demonstrates strong performance on both the public and extended datasets. Compared to CNN+LSTM, TimeSformer, and ViViT, our model consistently outperforms all baseline methods. In particular, our method reaches an accuracy of 83.1%, whereas ViViT reaches 78.5%.

Table 1: Performance comparison of different models on both datasets.

Dataset	Model	Accuracy	Precision	Recall	F1-Score
Public Dataset	CNN + LSTM [26]	66.2% (0.3%)	0.70	0.65	0.66
	TimeSformer [34]	70.7% (0.2%)	0.60	0.62	0.60
	ViViT [12]	78.5% (0.4%)	0.73	0.72	0.72
	ours	83.1% (0.3%)	0.75	0.76	0.75
Extended Dataset	CNN + LSTM [26]	74.8% (0.4%)	0.76	0.75	0.76
	TimeSformer [34]	66.4% (0.2%)	0.65	0.64	0.63
	ViViT [12]	76.6% (0.3%)	0.71	0.69	0.70
	ours	80.4% (0.3%)	0.81	0.80	0.80

To evaluate the impact of the FRM and Cross-Attn module at different depths, we conducted a series of controlled experiments under the same training settings. We constructed three model variants with two, four, and six FRM with Cross-Attn layers. The experimental results are shown in Table 2. The corresponding slip detection accuracies were 76.9%, 83.1%, and 83.0%.

Table 2: Effect of FRM and Cross-Attn layer depth on slip detection accuracy.

Layer Configuration	Accuracy (%)
2 layers of FRM + Cross-Attn	76.9
4 layers of FRM + Cross-Attn (Ours)	83.1
6 layers of FRM + Cross-Attn	83.0

The results show a clear performance improvement when the number of layers increases from two to four. This indicates that deeper cross-modal interactions help fuse visual and tactile features more effectively and improve representation quality. However, when the number of layers increases to six, performance slightly decreases. This may result from redundant features or optimization difficulties in deeper networks, which can reduce generalization ability.

Based on these analysis, we selected the four-layer structure as the optimal configuration for the FRM and Cross-Attn modules. The setting provides a balance between feature richness and computational efficiency.

5.2 Ablation Experiment

We conducted a series of ablation experiments to investigate the impact of different architectural components on slip detection performance. Unless otherwise specified, all experiments in this section are performed on the public dataset.

5.2.1 Vision-Tactile Multimodal Ablation

To evaluate the effectiveness of multimodal fusion, we compared three configurations: a *vision-only* baseline, a *tactile-only* baseline, and the proposed *vision-tactile fusion* model. To ensure a fair comparison, all experiments use the same backbone architecture. The results demonstrate that fusing visual and tactile information significantly improves slip detection performance compared to use either modality alone.

To investigate the contribution of each modality, we fine-tuned the proposed network using single-modality inputs. The overall architecture remained largely unchanged. When input dimensions differed, we applied simple resizing operations without altering the core parameters. This flexibility allows the model to be easily adapted for fusion with different data modalities. Evaluation results, including accuracy, are summarized in Table 3. Our multimodal fusion model consistently outperforms the single modal baseline across all metrics.

Table 3: Accuracy (%) comparison under different modality configurations and models.

Modality	CNN+LSTM	ViViT	Ours
Vision-only	70.7% (0.4%)	76.4% (0.7%)	81.5% (1.2%)
Tactile-only	72.3% (0.8%)	55.3% (0.5%)	76.9% (0.5%)
Vision & Tactile	74.8% (0.3%)	78.5% (0.4%)	83.1% (0.3%)

Notably, the visual-only model consistently outperformed the tactile-only model. This is likely because visual data captures most of the global appearance and contextual information of the object, while tactile

data provides more subtle and local contact cues. These results demonstrate that the proposed multimodal fusion framework effectively leverages the complementary strengths of visual and tactile modalities.

5.2.2 Without FRM

We designed the FRM to optimize features extracted from the visual and tactile branches. To evaluate its impact, we conducted an ablation experiment by removing the FRM. As shown in Table 4, without the FRM, feature extraction in each branch is performed independently and without rectification. Compared with the *No FRM* and *Cross-Attn only* baselines, incorporating the FRM alone improves slip detection accuracy by 4.6%.

Table 4: Ablation results on FRM and Cross-Attn modules (Accuracy %).

Model Variant	Accuracy (%)
w/o FRM, w/o Cross-Attn	74.8
FRM only	78.5
FRM (Time-only)	73.2
FRM (Space-only)	72.6
Cross-Attn only	80.0
FRM + Cross-Attn (Ours)	83.1

The improvement is due to the complementary characteristics of visual and tactile information and the time-sensitive nature of slip detection. The FRM emphasizes temporal information around slip events using temporal correction. It also applies a spatial correction mechanism to enhance tactile features by referencing cross-modal spatial patterns.

We attribute this gain to the complementary nature of visual and tactile information and the temporally sensitive nature of slip detection. The FRM leverages temporal rectification to emphasize frames before and after slip events and applies spatial rectification to enhance tactile features by referencing spatial patterns across modalities.

We further analyzed two FRM variants. The *Time-only* variant applies only temporal rectification ($\lambda_T = 1$, $\lambda_S = 0$), and the *Space-only* variant uses only spatial rectification ($\lambda_T = 0$, $\lambda_S = 1$). As indicated in Table 4, both variants lead to sub-optimal performance compared to the full FRM. This confirms that the combination of temporal and spatial rectification is essential for robust multimodal alignment and performance.

5.2.3 Without Cross-Extraction and Fusion (Cross-Attn)

To evaluate the contribution of the Cross-Attn module for multimodal feature interaction and fusion, we removed it from the network architecture. After removing the Cross-Attn module, we used a standard feature extraction process followed by simple feature concatenation. According to Table 4, the model with Cross-Attn only, without the FRM, improves performance by 3.1% compared with the base model.

This improvement is attributed to the spatio-temporal cross-attention mechanism. It enables effective alignment and interaction between visual and tactile modalities. It also helps the model capture cross-modal dependencies and dynamic patterns, which are critical for accurate slip detection.

5.3 Attention Analysis

To verify whether the model focuses on important spatio-temporal features for slip detection and to evaluate the effectiveness of multimodal fusion, we visualize the attention mechanism. The visualization helps determine whether the model attends to meaningful regions. It also reveals potential issues such as overfitting and attention drift. In addition, it shows the contribution of each network layer to the final decision and improves model interpretability.

The qualitative difference in attention landscapes between the two modalities carries clear physical insights into the dynamic slip process. As visualized via the Attention Rollout method in Fig. 6, visual attention patterns appear more dispersed across the frames. Physically, this dispersion occurs because the external camera captures a complex background (e.g., the robotic gripper limbs, lighting shadows, and macro object surfaces); hence, the visual stream naturally focuses on extracting global deformation context and macro appearance changes. Conversely, the tactile attention patterns are tightly concentrated around local contact sub-regions. Since the DIGIT sensor operates in an enclosed mechatronic space that isolates external environmental noise, it yields a much higher signal-to-noise ratio. The sharp focus of tactile attention reflects its physical role: capturing micro-scale shear displacements and local contact area boundaries that are visually occluded. This bimodal variance demonstrates that STR-CMFNet successfully establishes a 'global semantic guidance' and 'local slip perception' synergy.

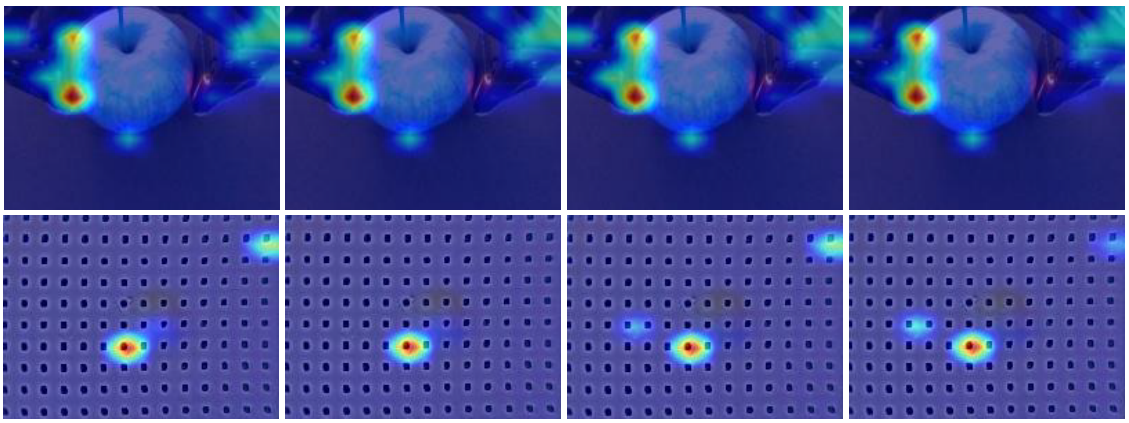

Figure 6: Visualization of multimodal attention. The top rows show attention maps over 4 visual frames, while the bottom correspond to 4 tactile frames. Red indicates higher attention weights. The model exhibits clear temporal focus around key contact moments, such as before and after slippage. Visual attention is more dispersed due to complex backgrounds, whereas tactile attention is concentrated around contact areas.

6 Conclusion

The paper proposes STR-CMFNet, a Visual-Tactile Spatio-Temporal Rectification and Cross-Modal Fusion Network for slip detection during robotic grasping of soft fruits. By integrating a Spatio-Temporal Rectification Module (STR) and a Cross-Modal Fusion Module (CMF), our method aligns visual and tactile features in both space and temporal, enabling more accurate and robust slip prediction. Experiments show that STR-CMFNet achieves a classification accuracy of 83.1% and an F1-score of 0.75, outperforming existing approaches on both public benchmarks and our newly collected soft fruit dataset. Despite these results, our work has limitations. First, the proposed network has not yet been validated on a physical robotic system, which is essential for assessing real-world robustness and deployment feasibility. Second, our model is trained on data from a Digit tactile sensor and a RealSense camera, differences in sensor modalities may limit generalization to other hardware setups. In line with the vision of embodied intelligence, our future work

aims to extend the framework to multi-finger dexterous hands equipped with tactile feedback, enabling more complex manipulation of deformable objects.

Acknowledgement: We thank the Human-Robot Collaboration, Cooperation and Cognition (HRC3) Laboratory at Northeastern University for providing the DIGIT sensors and computational resources. We also appreciate the helpful guidance on the manuscript provided by Yi Guo.

Funding Statement: This work was supported in part by the National Natural Science Foundation of China under Grant 62373087, in part by the Liaoning Provincial Applied Basic Research Program under Grant 2025JH2/101300009, and in part by the Liaoning Revitalization Talents Program under Grant XLYC24110114.

Author Contributions: Supervision, Hao Chu and Fei Wang; funding acquisition, Hao Chu and Fei Wang; methodology, Xibin Xiao; software, Xibin Xiao; writing—original draft preparation, Xibin Xiao; data curation, Song Gao; writing—review and editing, Chao Zheng. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The datasets generated during the current study are not publicly available due to privacy restrictions but are available from the corresponding author on reasonable request.

Ethics Approval: This study did not involve human participants or animals, and therefore ethical approval was not required.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Niu M, Lu Z, Chen L, Yang J, Yang C. VERGNet: visual enhancement guided robotic grasp detection under low-light condition. IEEE Robot Autom Lett. 2023;8(12):8541–8. doi:10.1109/lra.2023.3330664.
2. Xu Z, Wu J, Zeng A, Tenenbaum J, Song S. DensePhysNet: learning dense physical object representations via multi-step dynamic interactions. In: Proceedings of the Robotics: Science and Systems XV (RSS 2019); 2019 Jun 22–26; Freiburg im Breisgau, Germany.
3. Wang C, Wang S, Romero B, Veiga F, Adelson E. SwingBot: learning physical features from in-hand tactile exploration for dynamic swing-up manipulation. In: Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2020 Oct 24–2021 Jan 24; Las Vegas, NV, USA. p. 5633–40.
4. She Y, Wang S, Dong S, Sunil N, Rodriguez A, Adelson E. Cable manipulation with a tactile-reactive gripper. Int J Robot Res. 2021;40(12–14):1385–401. doi:10.1177/02783649211027233.
5. Calandra R, Owens A, Upadhyaya M, Yuan W, Lin J, Adelson EH, et al. The feeling of success: does touch sensing help predict grasp outcomes? In: Proceedings of the 1st Conference on Robot Learning (CoRL 2017); 2017 Nov 13–15; Mountain View, CA, USA. p. 314–23.
6. Li J, Dong S, Adelson E. Slip detection with combined tactile and visual information. In: Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA); 2018 May 21–25; Brisbane, Australia. p. 7772–7.
7. Jiang J, Cao G, Butterworth A, Do TT, Luo S. Where shall I touch? vision-guided tactile poking for transparent object grasping. IEEE/ASME Trans Mechatron. 2023;28(1):233–44. doi:10.1109/tmech.2022.3201057.
8. Gao J, Huang Z, Tang Z, Song H, Liang W. Visuo-tactile-based slip detection using a multi-scale temporal convolution network. arXiv:2302.13564. 2023.
9. Yan G, Schmitz A, Tomo TP, Somlor S, Funabashi S, Sugano S. Detection of slip from vision and touch. In: Proceedings of the 2022 International Conference on Robotics and Automation (ICRA); 2022 May 23–27; Philadelphia, PA, USA. p. 3537–43.
10. Zhang J, Liu H, Yang K, Hu X, Liu R, Stiefelhagen R. CMX: cross-modal fusion for RGB-X semantic segmentation with transformers. IEEE Trans Intell Transport Syst. 2023;24(12):14679–94.

11. Reed S, Zolna K, Parisotto E, Colmenarejo SG, Novikov A, Barth-Maron G, et al. CoAtNet: marrying convolution and attention for all data sizes. In: Proceedings of the 35th International Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. p. 3965–77.
12. Han Y, Yu K, Batra R, Boyd N, Mehta C, Zhao T, et al. Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer. IEEE/ASME Trans Mechatron. 2025;30(1):554–66. doi:10.1109/tmech.2024.3400789.
13. Li B, Bai J, Qiu S, Wang H, Guo Y. VITO-transformer: a visual-tactile fusion network for object recognition. IEEE Trans Instrum Meas. 2023;72:2530810. doi:10.1109/tim.2023.3326241.
14. Wang T, Yang C, Kirchner F, Du P, Sun F, Fang B. Multimodal grasp data set: a novel visual–tactile data set for robotic manipulation. Int J Adv Rob Syst. 2019;16:1729881418821571.
15. Yan G. Hard dataset and Normal dataset for robotic tactile sensing. IEEE Dataport. 2022. doi:10.21227/94km-0873.
16. Yuan W, Dong S, Adelson EH. GelSight: high-resolution robot tactile sensors for estimating geometry and force. Sensors. 2017;17(12):2762. doi:10.3390/s17122762.
17. Taylor IH, Dong S, Rodriguez A. GelSlim 3.0: high-resolution measurement of shape, force and slip in a compact tactile-sensing finger. In: Proceedings of the 2022 International Conference on Robotics and Automation (ICRA); 2022 May 23–27; Philadelphia, PA, USA. p. 10781–7.
18. Lambeta M, Chou PW, Tian S, Yang B, Maloon B, Most VR, et al. DIGIT: a novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. IEEE Robot Autom Lett. 2020;5(3):3838–45. doi:10.1109/lra.2020.2977257.
19. Tu Z, Talebi H, Zhang H, Yang F, Milanfar P, Bovik A, et al. MaxViT: multi-axis vision transformer. arXiv: 2204.01697. 2022.
20. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA. p. 1–7.
21. Lu Z, Chen L, Dai H, Li H, Zhao Z, Zheng B, et al. Visual-tactile robot grasping based on human skill learning from demonstrations using a wearable parallel hand exoskeleton. IEEE Robot Autom Lett. 2023;8(9):5384–91. doi:10.1109/lra.2023.3295296.
22. Zhao Z, Li W, Li Y, Liu T, Li B, Wang M, et al. Embedding high-resolution touch across robotic hands enables adaptive human-like grasping. Nature Mach Intell. 2024;7(6):889–900. doi:10.1038/s42256-025-01053-3.
23. Cravetz M, Vyas P, Grimm C, Davidson JR. Slip detection for compliant robotic hands using inertial signals and deep learning. Front Robot AI. 2025;12:1698591.
24. Hyttinen E, Kragic D, Detry R. Learning the tactile signatures of prototypical object parts for robust part-based grasping of novel objects. In: Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA); 2015 May 26–30; Seattle, WA, USA. p. 4927–32.
25. Zhao C, Yu Y, Ye Z, Tian Z, Zhang Y, Zeng LL. Universal slip detection of robotic hand with tactile sensing. Front Neurorobot. 2025;19:1478758. doi:10.3389/fnbot.2025.1478758.
26. Si Z, Zhu Z, Agarwal A, Anderson S, Yuan W. Grasp stability prediction with sim-to-real transfer from tactile sensing. In: Proceedings of the 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2022 Oct 23–27; Kyoto, Japan. p. 7809–16.
27. Xu H, Nashit Arshad SM, Peng S, Xu H, Yin H, Li Q. Slip detection and stable grasping with multi-fingered robotic hand using deep learning approach. IET Cyber Syst Robot. 2025;7(1):e70036. doi:10.1049/csy2.70036.
28. Zapata-Impata BS, Gil P, Torres F. Learning spatio temporal tactile features with a ConvLSTM for the direction of slip detection. Sensors. 2019;19(3):523. doi:10.3390/s19030523.
29. Li Y, Wu P, Niu M, Chen W, Gao G. Visual–tactile slip detection via weight allocation and cross-self-attention net. IEEE Sens J. 2025;25(2):3750–60. doi:10.1109/JSEN.2024.3501372.
30. Zhang Z, Zhou Z, Wang H, Zhang Z, Huang H, Cao Q. Grasp stability assessment through attention-guided cross-modality fusion and transfer learning. In: Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2023 Oct 1–5; Detroit, MI, USA. p. 9472–9.

31. Huang Z, Gao J, Tang Z, Song S, Guo J. Slip detection for robot grasping based on attention mechanism and visuo-tactile fusion. Inf Control. 2024;53(2):191–8.
32. Zheng D, Chen Y. Enhancing robotic grasping detection using visual-tactile fusion perception. Sensors. 2026;26(2):724. doi:10.3390/s26020724.
33. Cui S, Wang R, Wei J, Li F, Wang S. Grasp state assessment of deformable objects using visual-tactile fusion perception. In: Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31–Aug 31; Paris, France. p. 538–44.
34. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? In: Proceedings of the 38th International Conference on Machine Learning (ICML 2021); 2021 Jul 18–24; Virtual Event. p. 813–24.