Tactile-based Grasping Stability Prediction based on Human Grasp Demonstration for Robot Manipulation

Zhou Zhao[1], Wenhao He[2], and Zhenyu Lu[2†], *Member, IEEE*

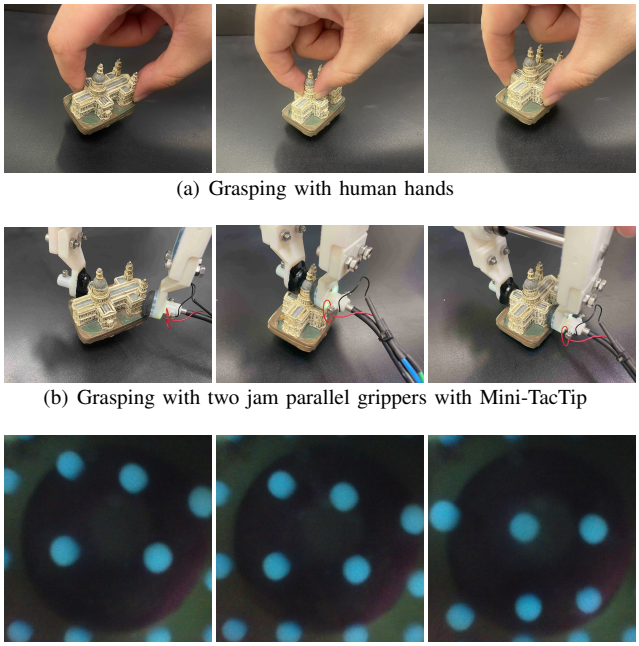
***Abstract*—To minimize irrelevant and redundant information in tactile data and harness the dexterity of human hands. In this paper, we introduce a novel binary classification network with normalized differential convolution (NDConv) layers. Our method leverages the recent progress in visual-based tactile sensing to significantly improve the accuracy of grasp stability prediction. First, we collect a dataset from human demonstration by grasping 15 different daily objects. Then, we rethink pixel correlation and design a novel NDConv layer to fully utilize spatio-temporal information. Finally, the classification network not only achieves a real-time temporal sequence prediction but also obtains an average classification accuracy of 92.97%. The experimental results show that the network can hold a high classification accuracy even when facing unseen objects.**

***Index Terms*—Grasping, deep learning in grasping and manipulation, learning from experience.**

I. INTRODUCTION

THE exploration of robotic grasping has spanned several decades, leading to a wealth of methodologies being developed. Typically, visual systems are utilized to identify a suitable grasping posture [1], [2]. However, robotic grasping involves not only visual challenges but also tactile problems [3]. Over the past decade, robotic hands have attained heightened levels of dexterity [4]. This advancement can be attributed to breakthroughs involving soft tactile sensors that enable the perception of touch across robotic hands [5], [6]. Recently, numerous research employs soft tactile sensors to achieve stable robotic grasping [7], [8], [9].

Naturally, in order to prevent grasping failures like the slipping of objects, detecting unstable grasps is crucial to activate corrective actions. As shown in Fig. 1, facing the same object, human hands can achieve stable grasps with different forces at different positions (see Fig 1(a)). However, to achieve a human-like grasping ability, the softness of vision-based tactile sensors should be as close as possible to human hands. Based on our previous work [10], [11], the two jam parallel grippers with Mini-TacTip designed by us also can satisfy stable grasps at different positions (see Fig. 1(b)), and obtain different captured images (see Fig. 1(c)) corresponding to different stable grasping states.



(a) Grasping with human hands

(b) Grasping with two jam parallel grippers with Mini-TacTip

(c) Captured images from with Mini-TacTip sensors

Fig. 1. Human stable grasps different positions on the same object with different forces (Fig. 1(a). Here, we use two jam parallel grippers with visual-based tactile sensors called **Mini-TacTip** (Fig. 1(b)) to capture tactile images (Fig. 1(c)) to evaluate grasp stability throughout the grasping process.

Most previous work predicting grasp stability is mainly based on tactile images [12] or tactile temporal sequences [3] acquired by visual-based tactile sensors. However, the way to capture tactile data is mainly based on robotic hands. It is well known that robotic hands need to be pre-programmed to perform grasping tasks [13], [14], [15], which reduces the possibility of generalizing the learned grasping capabilities to other tasks. Hence, we collect tactile data from human demonstrations by using a wearable parallel hand exoskeleton from our previous work [11], which not only obtains human grasping experience but also reduces irrelevant or redundant information, achieving fast and efficient tactile data collection.

With the development of deep learning methods, many novel

Manuscript received: September 4, 2023; Revised November 13, 2023; Accepted January 14, 2024.

This paper was recommended for publication by Editor Tetsuya Ogata upon evaluation of the Associate Editor and Reviewers' comments. This work was supported in part by the Vice Chancellor's Early Career Researcher (VC ECR) 2023-2025 award of the University of the West of England and in part by the National Funding Program of China for Post-Doctoral Researchers under Grant GZC20230924. *(†Corresponding author: Zhenyu Lu.)*

[1]Zhou Zhao is with School of Computer Science, Central China Normal University, Wuhan, China. zhaozhou@ccnu.edu.cn

[2]Zhenyu Lu, and Wenhao He are with the Faculty of Environment and Technology and Bristol Robotics Lab at the University of the West of England, Bristol, BS16 1QY, UK. Zhenyu.Lu@uwe.ac.uk, fhkitval@gmail.com

Digital Object Identifier (DOI): see top of this page.

network frameworks have achieved a promising performance in robotic grasping [16], [17]. However, network frameworks that simply consist of the standard layers of deep learning (convolutional layers, max-pooling layers, dense layers, etc.) are not sufficient for exploring tactile data. Therefore, in this paper, we will rethink pixel correlation to fully use spatio-temporal information from tactile data.

Thus, the main contributions of this paper are:

1) **Data collection from human demonstrations**. Unlike the dexterity of human hands, using robotic hands to acquire tactile data often requires pre-programming, which is likely to break the object with excessive force or yield tactile data with limited generalizability. So, to improve the above cons, we collect tactile temporal sequences from human demonstrations to transfer human grasping experience to robotic grasping systems.
2) **Normalized differential convolution (NDConv)**. We rethink pixel correlation and design a new convolutional layer called NDConv. NDConv not only makes full use of spatio-temporal information but also improves the generalization ability of the classification network.
3) **Classification network**. We propose a binary classification model to extract features of different grasping phases, which helps the robotic grasping system provide relevant operation strategies when facing different grasping states.

II. RELATED WORK

Various visual-based tactile sensors integrated with deep learning techniques have been rapidly developed and used in robotic manipulation [18]. Many studies have been done about robotic grasping stability relying on visual-based tactile feedback. Hence, we will introduce some previous work on visual-based tactile sensors and deep learning methods in grasping, respectively.

A. Visual-based Tactile Sensors in Grasping

Visual-based tactile sensors play a vital role in enhancing robotic grasping capabilities by providing tactile feedback to robotic systems [19]. This feedback allows the robot to better understand its interactions with objects and adjust its grasping force. Typically designed to mimic human skin, they can sense various parameters such as pressure, vibration, and temperature, and are often arrayed on robotic fingers or grippers to provide fine-grained information about the surface of an object and forces applied during grasping. For example, tactile sensors were employed by Bekiroglu et al. [20] to estimate grasp stability, while Li et al. [21] presented integrating tactile feedback into dynamics models of objects to enhance the capabilities of a dexterous hand.

Robotic grasping mainly includes three phases: approaching, grasping and lifting objects. Most research detects grasping stability in the lifting phase. James et al. [22] designed a biomimetic optical tactile sensor for rapid slip detection. Veiga et al. [23] proposed a novel method of slip prediction to achieve stabilizing objects. Calandra et al. [24] monitored incipient slip to realise stable grasps. Yan et al. [25] presented a multi-phase, multi-output framework to accomplish stability

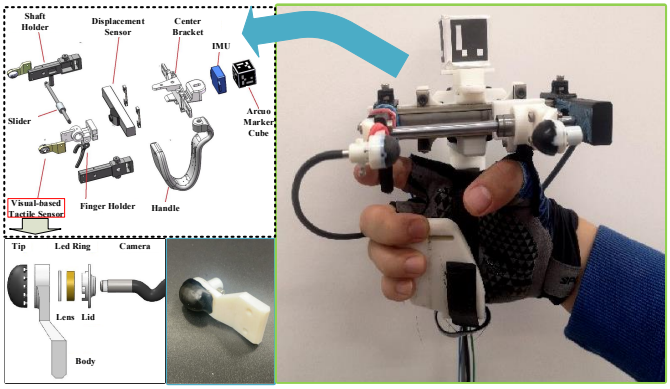


Fig. 2. **The Wearable tool-like parallel hand exoskeleton structure**. Visual-base tactile sensors are designed to mimic human skin and can acquire the movement of 16 pins (see Fig. 6(a))

prediction and slip detection. Yi et al. [26] extracted tactile features from multimodal tactile data and presented a novel ensemble approach for the grasp stability recognition task. To solve the problem of blind grasping, Dang et al. [27] only used tactile feedback without visual and geometric information to predict the robotic grasping stability. However. in our case, we design a visual-based tactile sensor, called **Mini-TacTip**, to mimic human tactile perception, and use it to collect real human-like tactile data.

B. Deep Learning in Grasping

With the rise of deep learning, researchers try to combine deep learning methods with tactile grasping data. Lots of evidence has indicated that deep learning methods could boost the grasping performance [1], [28]. To date, most deep learning methods are built based on some convolutional-based milestone architectures such as VGG [29], GAN [30], ResNet [31], GoogLeNet [32], and Transformer [33], etc. And deep learning methods used in the field of robot grasping represent further effective applications based on foundational network architectures. For example, Yang et al. [34] presented a deep learning method based on a critic-policy format and Rusu et al. [35] presented the utilization of progressive neural networks to tailor an established deep reinforcement learning policy, enabling models learned in simulation to be reliably transferred to real environments and even generalized to novel objects. Gualtieri et al. [36] used simulated data to train the proposed deep-learning method for detecting grasping poses. Facing to instance grasping task in cluttered scenes, Fang et al. [37] proposed a convolutional neural network for multi-task domain adaptation. These work have shown the success of using deep learning methods in robotic grasping.

However, our method differs from the methods mentioned above. First, we add human grasping experience into the grasp stability recognition task by collecting tactile data from human demonstration. Second, we choose to stack three consecutive frames to obtain 3D-like images, which not only reduces irrelevant and redundant information in tactile temporal sequences but also shortens prediction time. Finally, we propose an end-to-end classification network with normalized

differential convolution (NDConv) layers that processes rich spatio-temporal information to predict grasp stability, and also provide a controlled evaluation of whether incorporating tactile information improves grasp success within a robotic system with Mini-TacTip.

III. PRELIMINARY WORK

A. Mechanical Structure of Wearable Tool-like Parallel Hand Exoskeleton with Mini-TacTip

Fig. 2 shows the mechanical structure of wearable tool-like parallel hand exoskeleton with visual-based tactile sensors (**Mini-TacTip**) from our previous work [11], and it consists of an inertial measurement unit (IMU, WitMotion Bluetooth 2.0, 9 Axis IMU Sensor, BWT901CL, China), a displacement sensor (Greet, Resolution: 0.01mm, Maximum displacement: 75mm, China) in a commercial way, two Mini-TacTip, a aruco maker cube, and some support components that are made by using a 3D printer and PLA materials, etc. The displacement sensor incorporates a gliding mechanism, although it is not designed to withstand direct pushing and pulling forces exerted by human hands. Hence, to mitigate friction and withstand the pinching forces exerted by human hands, we introduce an additional slider (bearing, inner diameter 6mm, outer diameter 10mm) along with a set of holders on either side. Both the finger holder and the handle are interchangeable, allowing them to be adjusted according to the hand size and manipulation preferences of various operators.

The Mini-TacTip is designed with inspiration from [6]. It consists of various components, including a Tip with pins printed with Agilus, a lens, and a camera to match a finger size, etc. The Mini-TacTip is designed to be fingertip-sized and can be assembled onto both the shaft holder and the slider. The chosen assembly method aims to minimize the impact of the slider's repetitive movements on the tactile images. The Mini-TacTip can acquire the movement of 16 pins embedded in the Tip (see Fig. 6(a)). Due to the Mini-TacTip's surface being as soft as human skin, it can rapidly respond to changes in the distribution of 16 pins movement.

B. Data Collection from Human Demonstration

Achieving robotic grasping with human-level dexterity is one of the primary goals of robotics. However, most grasping datasets are captured from robotic hands [38], [39], [40], ignoring the dexterity of human hands. We use a wearable tool-like parallel hand exoskeleton (see Fig. 2) for data collection, which relies on human arm movements. The "dexterity" refers to the nuanced and complex fine motor control capabilities of the human arm compared to robotic manipulators, which extends beyond the mere number of Degrees of Freedom (DoFs). While both human arms and robotic end-effectors operate with 6 DoFs, human arms offer a broader range of intricate movements and sensory capabilities, making them more versatile and adaptable in certain scenarios. As shown in Fig. 3, a dataset for robotic grasping is collected through human demonstrations in this paper, which facilitates the transfer of human grasping expertise to robots.

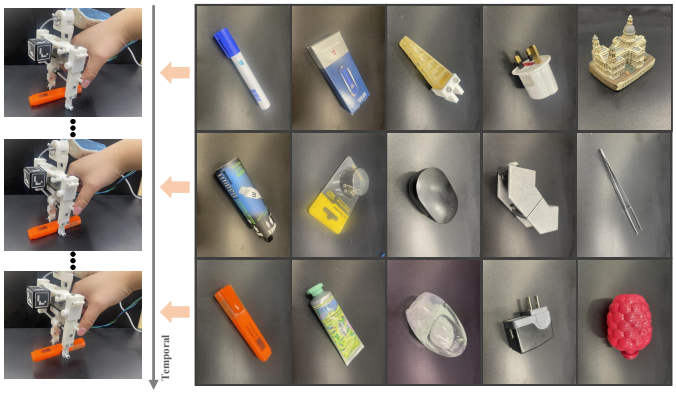


Fig. 3. **Data collection**. The dataset for robotic grasping is collected through human demonstration. We use the wearable tool-like parallel hand exoskeleton (see Fig. 2) to grasp an object at different positions.

Most research focuses on detecting optimal grasp positions. Yet, facing the same object, any position can be tried to grasp for humans. Therefore, we use the wearable tool-like parallel hand exoskeleton (see Fig. 2) to grasp an object at different positions. The entire process of grasping objects mainly includes three phases: approaching, grasping, and lifting objects. For each grasp, the hand exoskeleton should be as close to the object as possible, since approaching the object phase has less impact on the grasp stability prediction, thereby minimizing its contribution to the overall duration of the grasping process. Moreover, human hands manipulate the hand exoskeleton, which naturally adds random perturbations throughout the grasping process, thereby helping to generalize the human grasping experience to robotic systems. To release the problem of class imbalance, we collect tactile data based on a 6:4 ratio for successful and unsuccessful grasps.

The overall sequence of each video is as follows: firstly, according to human grasping experience, the hand exoskeleton is moved to random grasping positions of objects, and the duration is about 2s. Then, two-jaw parallel grippers of the hand exoskeleton start to close until they reach the desired random grasping force, and the duration is also about 2s. Finally, the object is lifted at a slow speed for about 3s, which is enough to observe the lifting result based on human experience and feeling. If the object remains in the two-jaw parallel grippers and does not slip during lifting, it is labelled as a successful grasp manually. Conversely, if the object shows an unstable grasping trend such as slipping or dropping from the two-jaw parallel grippers, it is categorized as an unsuccessful grasp.

Ultimately, the dataset consists of 419 videos, each video annotated with a successful or unsuccessfully grasp. The duration of each video is 7s, with 30 frames per second. These videos are collected from the camera system of Mini-TacTip. Their spatial resolution is 1280×720 pixels. 319 videos are used for training and the remaining 100 for testing. The total number of objects involved in the experiment is 15, out of which 12 are used for the training dataset, and 3 are used for the test dataset. When evaluating the number of grasp attempts,

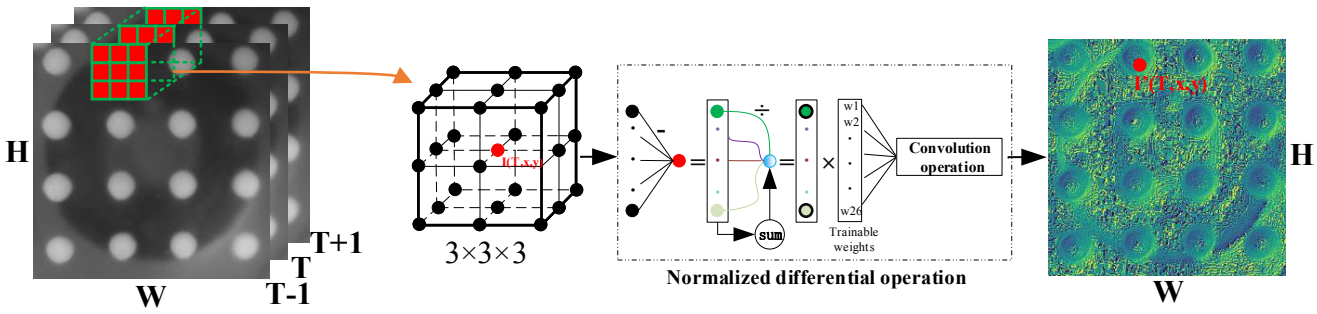


Fig. 4. **Normalized differential convolution (NDConv)**. W and H denote the weight and height of an image, respectively. T denotes temporal. Three consecutive frames (T-1, T, T+1) as input of NDConv layer, we take the minimum convolution operation area of 3×3×3, and learn the relationship between pixels by normalized differential operations.

it's worth noting that we conduct only one attempt per object in the same position and pose. **Objects in the test dataset that do not appear in the training dataset**.

IV. METHODOLOGY

A. Rethinking Pixel Correlation

The standard 2D convolution operation mainly consists of two parts. Initially, input feature maps are sampled by k×k convolution kernels. Subsequently, the sampled values are assigned weights and eventually summed and fused. Let k=3 be considered as an example, and the standard 3×3 convolution operation is defined as follows:

$$\mathbf{Conv}\left(x,y\right)=\sum_{dx=\ 1}^{1}\sum_{dy=\ 1}^{1}\omega\left(dx,dy\right)\mathbf{I}\left(x+dx,y+dy\right) \tag{1}$$

where $\mathbf{Conv}\left(\cdot\right)$ is the feature maps after convolution operation. $\mathbf{I}\left(\cdot\right)$ denotes original feature maps. x and y represent the location of the pixel in the image coordinate system. $\omega\left(dx,dy\right)$ denotes the weight of convolution kernel. Each position of the convolution kernel is designed by $-1\leq dx\leq 1$ and $-1\leq dy\leq 1$.

Different receptive fields allow to handle multiple spatial scales and image resolutions, e.g., atrous convolution [41], [42], depthwise separable convolution [43] or deformable convolution [44], yet assuming such scales are known. These convolution operations calculate the linear sum of learned kernels and ignore the correlation between pixels. Compared with some similar existing works such as Local Binary Pattern (LBP) [45], Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [46], and Pixel Difference Convolution (PDC) [47], when they are used to handling temporal information, LBP [45] and PDC [47] cannot capture temporal changes. Although LBP-TOP [46] is specifically designed to capture temporal information, it involves the conversion of grayscale images into binary patterns, which loses some information related to the exact intensity values of pixels. Hence, this paper rethinks pixel correlation and designs a new convolutional layer called normalized differential convolution (**NDConv**).

1) Normalized Differential Convolution: The NDConv layer also has the same convolution kernel as the standard 2D convolution layer. However, before the convolution operation, a series of operations need to be performed on the convolution area for the NDConv layer. Therefore, in order to simplify the illustration, we will employ a 3×3 convolutional kernel for the modelling process. For three consecutive frames as input, we take the minimum convolution operation area of 3×3×3 as an example (see Fig. 4).

First, for the centre point $\mathbf{p}(T,x,y)$, each neighbouring point around it serves a distinct role. For example, when the pixel value of a neighbouring point differs significantly from that of the pixel value of the centre point $\mathbf{I}(T,x,y)$, there exists substantial contrast between them like the pronounced response typically observed at the edges of objects. Consequently, the neighbouring point exerts significant influence on the centre point, akin to generating substantial responses at edges. Therefore, we replace the pixel value of corresponding neighbouring points by calculating the difference between the neighbouring points and $\mathbf{I}(T,x,y)$. Simultaneously, to effectively incorporate temporal information, we extend these operations into the three-dimensional space. So, the difference based on 3×3×3 convolution operation area can be defined by

$$\begin{aligned}\mathbf{\Psi}(T+dt,x+&dx,y+dy)=\\&\mathbf{I}(T,x,y)-\mathbf{I}(T+dt,x+dx,y+dy)\end{aligned} \tag{2}$$

where $dt,\ dx$ and $dy\in\{+1,0,-1\}$, but $dt=dx=dy\neq 0$. $T\in$ [0, N), and N represents the total number of frames in a video. $0\leq x<W$, $0\leq y<H$.

Then, to refine local information and enhance contrast, local normalization is performed in 3×3×3 convolution operation area based on $\mathbf{p}(T,x,y)$.

$$\begin{aligned}\mathbf{\Theta}(T+dt,x+dx,y+dy)=&\\\frac{\mathbf{\Psi}(T+dt,x+dx,y+dy)}{\sum_{dt=\ 1}^{1}\sum_{dx=\ 1}^{1}\sum_{dy=\ 1}^{1}&|\mathbf{\Psi}(T+dt,x+dx,y+dy)|}\end{aligned} \tag{3}$$

Finally, a new feature map is obtained by a standard

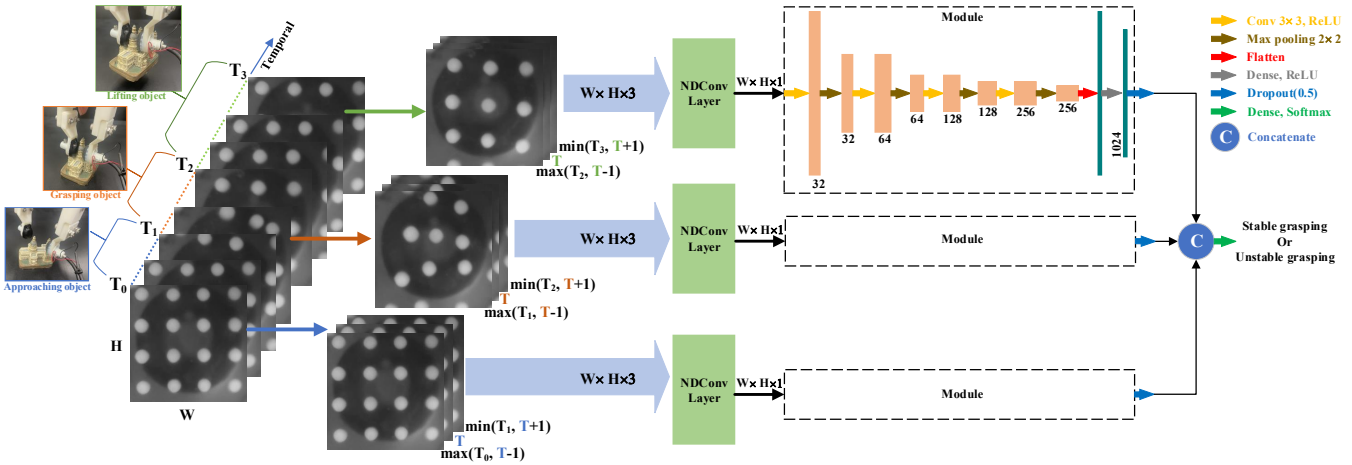
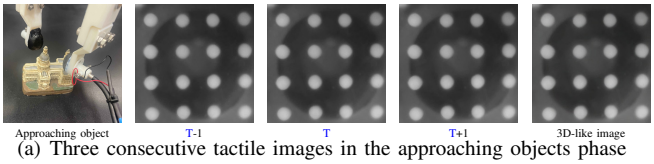
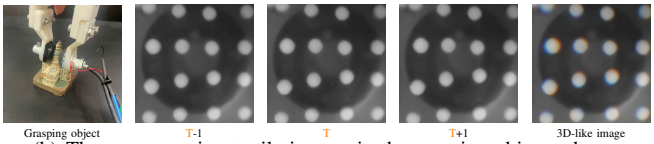


Fig. 5. **Overview of Network Architecture**. The tactile temporal sequences act as input, and it is divided into three phases: approaching ($T_0 \sim T_1$), grasping ($T_1 \sim T_2$), lifting ($T_2 \sim T_3$) objects. We obtain 3D-like images (see Fig. 6) by stacking three consecutive frames, and the 3D-like images act as the input of the proposed NDConv layer (see Fig. 4). Based on the Mini-TacTip camera's sampling frequency of 30Hz, 3D-like images can significantly enhance motion information (see Fig. 6(b)). Therefore, to ensure the prediction accuracy of the model, we should adjust the selection of three consecutive frames for different sampling frequencies. For example, the sampling frequency is changed from 30Hz to 60Hz, three consecutive frames (T-1, T, T+1) also should be changed to (T-2, T, T+2).



(a) Three consecutive tactile images in the approaching objects phase



(b) Three consecutive tactile images in the grasping objects phase

(c) Three consecutive tactile images in the lifting objects phase

Fig. 6. **Tactile images of different grasping phases for a stable grasp**. For the grasping objects phase, the 3D-like image differs with other phases, the motion information is enhanced by stacking three consecutive tactile images. In the case of stable grasps, the acquired tactile images remain unchanged in the approaching and lifting objects phase.

convolution operation:

$$\mathbf{NDConv}(T,x,y) = \sum_{dt=}^{1}\sum_{1\,dx=}^{1}\sum_{1\,dy=}^{1}\omega(T+dt,x+dx,y+dy) \times \mathbf{\Theta}(T+dt,x+dx,y+dy) \quad (4)$$

where $\omega(T+dt,x+dx,y+dy)$ denotes trainable parameters.

B. Overview of Network Architecture

The process of stable robotic grasping can be divided into multiple phases: approaching, grasping and lifting objects (see Fig. 6). For the approaching phase, the pins' distribution on the Mini-TacTip is not changed. When the Mini-TacTip starts to touch objects and transitions into grasping and lifting phases, the pins' distribution begins to change, eventually remaining unchanged to achieve stable robotic grasping. During the lifting phase, the tactile information obtained post-lifting is mainly served for slip detection. For three different phases, firstly, we design three different inputs at the beginning of the network architecture (see Fig. 5). Then, we stack three consecutive 2D images to obtain 3D-like images as inputs of the network, which provides a representation of the dynamic motion of the papillae pins. Some methods like [3] use the entire sequence of a video as inputs of the network, resulting in too much redundant information in the network. However, we choose 3D-like images from three different phases as inputs, which is beneficial to quickly respond to predict and can be applied in practical robotic grasping.

Later, the 3D-like images from three different phases are fed into the proposed NDConv layer, respectively, and then output three feature maps corresponding to the approaching, grasping, and lifting phases. Each feature map from the NDConv layer is fed into one module to continue extracting features at different levels. The module is mainly made of convolutional layers, Rectified Linear Unit (ReLU) layers for non-linear activation function, max-pooling layers, a flattened layer, and a dense layer. Finally, the outputs of three different modules are concatenated and fed into the softmax layer to obtain the prediction results.

V. EXPERIMENT AND RESULTS

A. Implementation and Experimental Setup

We implement our experiments on Keras/TensorFlow by using NVIDIA GeForce RTX 3090 GPU servers. The categorical crossentropy of Keras is used as the loss function of the entire network and predicts one probability distribution over

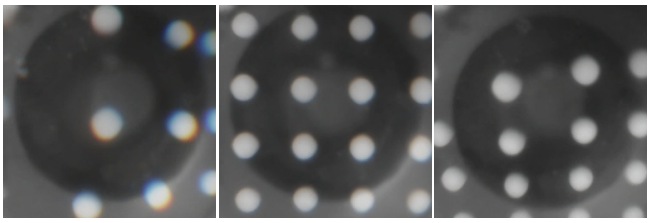


Fig. 7. **3D-like images in the lifting object phase**.

classes by one softmax function. For the optimizer, we use an Adam [48] (batchsize = 32, $\beta 1$ = 0.9, $\beta 2$ = 0.999, ε = 0.001, lr = 0.0001), and we do not use learning rate decay. Based on the collected dataset, we train the network on 30 epochs. We perform some preprocessing on these videos: 1) we crop these videos into the spatial resolution 680×680 pixels; 2) we convert colour images to grayscale images; 3) we resize each frame of these videos into images size 256×256 pixels.

B. Evaluation Methods

The assessment of our method is conducted through classification accuracy. The metric is defined by a ratio of the number of correct predictions and the total number of predictions made. As we prioritize a balanced distribution of data during the data collection process, achieving a high classification accuracy from the combined impact of both the proposed classification network and the collected dataset.

C. Results and Discussion

The classification accuracy is shown in Table. I. Compared with these state-of-the-art classification methods that are pre-trained on ImageNet, our proposed classification network achieves the best results in an average accuracy of 92.97% for predicting grasp stability. Since the objects in the test dataset never appear in the training dataset, the experimental results can explain the generalization ability of the classification model. To prove the superiority of our proposed method, we further perform a perturbation experiment on the test dataset. Captured images are susceptible to light intensity and noise. Hence, to test the performance of the proposed method on contrast and noise variations, we use the contrast function Eq. 5 of image augmentation tool [52] to change the contrast of the captured images, and add gaussian noise from a normal distribution N(0, β) to images by gaussian noise function of image augmentation tool [52], where β is sampled per image and varies between 0 and $\beta \times 255$. Finally, some changed images are shown in Fig. 8.

$$\mathbf{I}_c\left(T,x,y\right) = 255 \times \left(\frac{\mathbf{I}\left(T,x,y\right)}{255}\right)^{\alpha} \tag{5}$$

where α is used to change the image contrast, and we set the α to 0.4, 0.6, and 0.8, respectively.

In Table. I, the classification accuracy of the proposed method does not decrease with contrast changes. The proposed method has good stability to contrast changes. The classification accuracy decreases when adding Gaussian noise into images, but the classification accuracy of the proposed method degrades more slowly than other state-of-the-art methods, which proves that the proposed method is robust to noise.

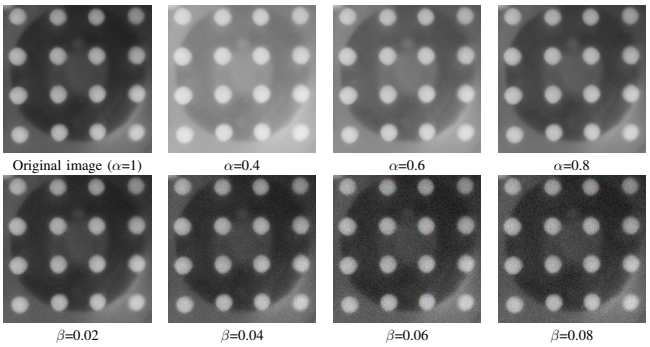


Fig. 8. **Images for different α and β.**

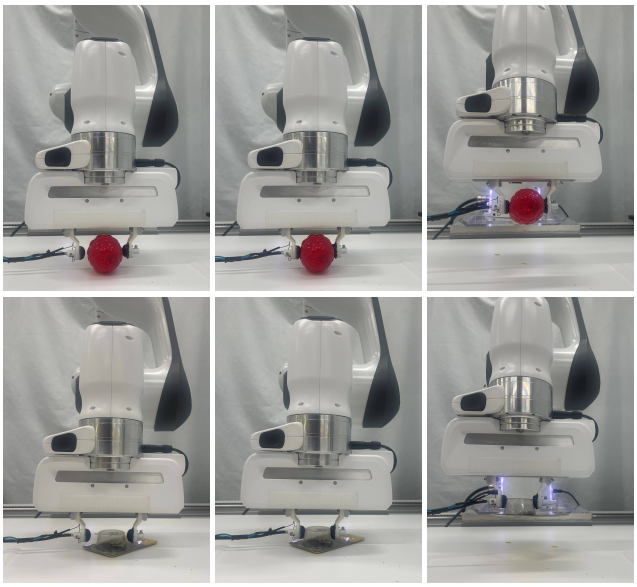


Fig. 9. **Grasp stability prediction on the robotic platform**. Ignoring the optimal grasping positions of objects, random grasping is done by the two jaw parallel grippers with our designed Mini-TacTip.

We have also conducted an ablation study to analyze the significance of our proposed NDConv in enhancing the generalization ability of the classification model in Table. II. First, we compare the impact of using/not using NDConv on classification accuracy. Utilizing NDConv outperforms non-use by a margin of 3.34% in terms of classification accuracy. We then replace NDConv with two standard convolutional layers with 3×3 convolution kernels, each producing an output feature map, however, our proposed method maintains a 1% performance advantage even when the number of parameters is almost the same. Moreover, the classification results prove that the proposed NDConv layer plays an important role in facing contrast and noise variations.

In order to verify the effect of transferring our proposed method to a real robot platform. We installed our designed Mini-TacTip on the end effector of the Franka Emika Panda as shown in Fig. 9. Facing different objects, we complete three phases: approaching, grasping and lifting objects. Ran-

TABLE I
CLASSIFICATION ACCURACY/% OF STATE-OF-THE-ART CLASSIFICATION METHODS ON TEST DATASET

Method	Accuracy%							
	α				β			
	1	0.8	0.6	0.4	0.02	0.04	0.06	0.08
VGG16 [29]	89.78	78.73	69.03	60.56	87.66	80.67	75.33	56.43
VGG19 [29]	85.57	70.21	61.53	53.22	81.85	73.55	58.78	53.93
ResNet50 [31]	90.11	81.19	72.31	63.55	89.04	85.34	78.90	69.32
InceptionV3 [49]	89.97	79.17	64.12	55.13	88.04	84.79	72.33	62.03
DenseNet121 [50]	90.62	80.26	70.18	59.51	89.93	86.19	80.11	69.41
LSTM [51]	89.05	82.43	67.22	58.48	88.91	85.15	77.24	57.23
Ours	**92.97**	**92.13**	**91.93**	**91.66**	**92.25**	**91.48**	**88.36**	**80.45**

TABLE II
CLASSIFICATION ACCURACY/% OF OUR PROPOSED CLASSIFICATION NETWORK WITH/WITHOUT NDCONV ON TEST DATASET

Method	Accuracy%								Parameters
	α				β				
	1	0.8	0.6	0.4	0.02	0.04	0.06	0.08	
Without NDConv	89.63	77.64	68.75	61.36	87.49	79.38	73.22	53.44	202,501,058 (∼20M)
With Conv.+Conv.	91.90	79.11	69.99	59.67	90.95	81.25	74.50	59.46	202,499,444 (∼20M)
With NDConv (ours)	**92.97**	**92.13**	**91.93**	**91.66**	**92.25**	**91.48**	**88.36**	**80.45**	202,499,489 (∼20M)

dom grasping is performed without considering the optimal grasping positions of objects, and then the grasping stability is predicted by our proposed method. Finally, the accuracy of grasp stability prediction remains at around 92%.

We need to emphasise the importance of considering all three phases - approaching, grasping, and lifting - when predicting grasping stability in robotic systems. This is a valid point, as each of these phases contributes unique information and challenges to the overall stability of the grasp:

1) **Approaching Phase**: The approaching phase involves the robot's movement toward the grasping object. This phase is crucial because the robot must align itself correctly with the object before attempting to grasp it. Misalignment during the approaching phase can lead to instability during grasping and lifting. By including this phase, it ensures that the robot's initial position and orientation are appropriate for a stable grasp.
2) **Grasping Phase**: The grasping phase is when the robot's end effector (such as a gripper) makes contact with the object and attempts to secure it. During this phase, temporal information from Mini-TacTip can be valuable for assessing the stability of the grasp. This information can indicate an unstable grasp.
3) **Lifting Phase**: The lifting phase occurs after the object has been grasped, and the robot begins to lift it. The stability of the grasp during lifting is vital to avoid dropping the object. Factors like the object's weight, shape, and the robot's control strategy play a role in the lifting phase's stability. Neglecting this phase may lead to unsuccessful manipulation. As shown in Fig. 7, facing unstable grasps (dropping or slipping), obtaining 3D-like images enhances motion information compared with the stable grasp.

Furthermore, robots lack the intuition and adaptability of humans in assessing and adjusting their grasp. Humans naturally incorporate all three phases when handling objects, making the grasping process robust and stable. Robots, being less intuitive, must rely on sensors and algorithms to ensure stability at each step. Incorporating information from all three phases allows the robot to make real-time adjustments, ensuring a stable grasp and successful manipulation. Failing to consider any of these phases might result in a less accurate prediction of grasping stability, which can lead to potential issues or failures in robotic manipulation tasks.

VI. CONCLUSIONS

In this paper, we propose a novel binary classification network with normalized differential convolution (NDConv) layers for predicting grasp stability. Firstly, we build a grasping dataset on 15 daily objects by using Mini-TacTip and the dexterity of human hands, which can transfer human grasping experience to robotic systems. Then, to distinguish grasping features from different phases and enhance the motion information, we stack three consecutive tactile images as inputs, and use the proposed NDConv layer to fully learn spatio-temporal information of tactile data. Finally, we use the proposed classification network to combine features of different grasping phases, achieving the grasp stability prediction. In our experiment, the proposed model achieves an average classification accuracy of 92.97% and provides an assessment of grasp stability. In the future, we will continue to provide corrective action for unstable grasp to improve grasp success rates.

REFERENCES

[1] R. Newbury, M. Gu, L. Chumbley *et al.*, "Deep learning approaches to grasp synthesis: A review," *IEEE Transactions on Robotics*, 2023.
[2] B. Wei, X. Ye, C. Long *et al.*, "Discriminative active learning for robotic grasping in cluttered scene," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1858–1865, 2023.
[3] G. Yan, A. Schmitz, S. Funabashi *et al.*, "Sct-cnn: A spatio-channel-temporal attention cnn for grasp stability prediction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 2627–2634.
[4] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019.

[5] N. F. Lepora, "Soft biomimetic optical tactile sensing with the tactip: A review," *IEEE Sensors J.*, 2021.

[6] B. Ward-Cherrier, N. Pestell, L. Cramphorn *et al.*, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[7] Z. Si, Z. Zhu, A. Agarwal *et al.*, "Grasp stability prediction with sim-to-real transfer from tactile sensing," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 7809–7816.

[8] J. Kwiatkowski, M. Jolaei, A. Bernier *et al.*, "The good grasp, the bad grasp, and the plateau in tactile-based grasp stability prediction," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 4653–4659.

[9] M. Su, D. Huang, Y. Guan *et al.*, "Soft tactile sensing for object classification and fine grasping adjustment using a pneumatic hand with an inflatable palm," *IEEE Transactions on Industrial Electronics*, 2023.

[10] Z. Zhao and Z. Lu, "Multi-purpose tactile perception based on deep learning in a new tendon-driven optical tactile sensor," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 2099–2104.

[11] Z. Lu, L. Chen, H. Dai *et al.*, "Visual-tactile robot grasping based on human skill learning from demonstrations using a wearable parallel hand exoskeleton," *IEEE Robotics and Automation Letters*, vol. 8, no. 9, pp. 5384–5391, 2023.

[12] R. Calandra, A. Owens, D. Jayaraman *et al.*, "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, 2018.

[13] M. Tavassoli, S. Katyara, M. Pozzi *et al.*, "Learning skills from demonstrations: A trend from motion primitives to experience abstraction," *IEEE Transactions on Cognitive and Developmental Systems*, 2023.

[14] M. Kyrarini, M. A. Haseeb, D. Ristić-Durrant *et al.*, "Robot learning of industrial assembly task via human demonstrations," *Autonomous Robots*, vol. 43, pp. 239–257, 2019.

[15] J. Bohg, A. Morales, T. Asfour *et al.*, "Data-driven grasp synthesis—a survey," *IEEE Transactions on robotics*, vol. 30, no. 2, pp. 289–309, 2013.

[16] H.-S. Fang, C. Wang, H. Fang *et al.*, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, 2023.

[17] L. Fu, M. Danielczuk, A. Balakrishna *et al.*, "Legs: Learning efficient grasp sets for exploratory grasping," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8259–8265.

[18] Z. Xie, X. Liang, and C. Roberto, "Learning-based robotic grasping: A review," *Frontiers in Robotics and AI*, vol. 10, p. 1038658, 2023.

[19] S. E. Navarro, S. Mühlbacher-Karrer, H. Alagi *et al.*, "Proximity perception in human-centered robotics: A survey on sensing systems and applications," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1599–1620, 2021.

[20] Y. Bekiroglu, J. Laaksonen, J. A. Jorgensen *et al.*, "Assessing grasp stability based on learning and haptic data," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 616–629, 2011.

[21] M. Li, Y. Bekiroglu, D. Kragic *et al.*, "Learning of grasp adaptation through experience and tactile sensing," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Ieee, 2014, pp. 3339–3346.

[22] J. W. James, N. Pestell, and N. F. Lepora, "Slip detection with a biomimetic tactile sensor," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3340–3346, 2018.

[23] F. Veiga, H. Van Hoof, J. Peters *et al.*, "Stabilizing novel objects by learning to predict tactile slip," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 5065–5072.

[24] R. Calandra, A. Owens, M. Upadhyaya *et al.*, "The feeling of success: Does touch sensing help predict grasp outcomes?" *arXiv preprint arXiv:1710.05512*, 2017.

[25] G. Yan, A. Schmitz, S. Funabashi *et al.*, "A robotic grasping state perception framework with multi-phase tactile information and ensemble learning," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6822–6829, 2022.

[26] Z. Yi, T. Xu, W. Shang *et al.*, "Genetic algorithm-based ensemble hybrid sparse elm for grasp stability recognition with multimodal tactile signals," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 3, pp. 2790–2799, 2022.

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

[28] M. Niu, Z. Lu, L. Chen, J. Yang, and C. Yang, "Vergnet: Visual enhancement guided robotic grasp detection under low-light condition," *IEEE Robotics and Automation Letters*, 2023.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] I. Goodfellow, J. Pouget-Abadie, M. Mirza *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[31] K. He, X. Zhang, S. Ren *et al.*, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[32] C. Szegedy, W. Liu, Y. Jia *et al.*, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2015, pp. 1–9.

[33] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[34] Y. Yang, H. Liang, and C. Choi, "A deep learning approach to grasping the invisible," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2232–2239, 2020.

[35] A. A. Rusu, M. Večerík, T. Rothörl *et al.*, "Sim-to-real robot learning from pixels with progressive nets," in *Conference on robot learning*. PMLR, 2017, pp. 262–270.

[36] M. Gualtieri, A. Ten Pas, K. Saenko *et al.*, "High precision grasp pose detection in dense clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 598–605.

[37] K. Fang, Y. Bai, S. Hinterstoisser *et al.*, "Multi-task domain adaptation for deep learning of instance grasping from simulation," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3516–3523.

[38] S. Levine, P. Pastor, A. Krizhevsky *et al.*, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International journal of robotics research*, vol. 37, no. 4-5, pp. 421–436, 2018.

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

[40] R. Wang, J. Zhang, J. Chen *et al.*, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 359–11 366.

[41] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

[42] L.-C. Chen, G. Papandreou, F. Schroff *et al.*, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[43] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 1251–1258.

[44] J. Dai, H. Qi, Y. Xiong *et al.*, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision (ICCV)*, 2017, pp. 764–773.

[45] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[46] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[47] Z. Su, W. Liu, Z. Yu *et al.*, "Pixel difference networks for efficient edge detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5117–5127.

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

[49] C. Szegedy, V. Vanhoucke, S. Ioffe *et al.*, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 2818–2826.

[50] G. Huang, Z. Liu, L. Van Der Maaten *et al.*, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2017, pp. 4700–4708.

[51] J. Toskov, R. Newbury, M. Mukadam *et al.*, "In-hand gravitational pivoting using tactile sensing," in *Conference on Robot Learning*. PMLR, 2023, pp. 2284–2293.

[52] A. B. Jung, K. Wada, J. Crall *et al.*, "imgaug," https://github.com/aleju/imgaug, 2020, online; accessed 01-Feb-2020.