IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Grasp Independent Indirect Tool Force Estimation using Vision-based Tactile Sensors

Luchen Li[1], and Thomas George Thuruthel[1]

***Abstract*—Humans possess the capability to seamlessly integrate tools into their body schema, enabling precise and adaptive interactions with the environment. This touch-mediated ability allows us to dexterously use tools in everyday tasks, an ability currently lacking in robotic systems. In this work, we propose a novel method for indirect force estimation in robotic tool use, a prerequisite for advanced tool use, leveraging vision-based tactile sensing (VTS) and deep learning techniques. By capturing high-resolution spatial deformations from tactile images, our model implicitly infers force transmission dynamics without requiring explicit knowledge of tool properties or material characteristics. We validate our approach across multiple tool types using a single trained machine learning model, demonstrating its generalization capability. This work represents the first demonstration of indirect force estimation for tool-mediated robotic interactions, offering a pathway toward more dexterous and adaptive robotic tool use in real-world applications.**

***Index Terms*—Force and Tactile Sensing, Soft Sensors and Actuators, Deep Learning in Grasping and Manipulation, Perception for Grasping and Manipulation, Sensorimotor Learning.**

I. Introduction

HUMANS demonstrate an unparalleled ability among all creatures to use in-hand tools effectively, which enables us to extend the reach and capabilities of the physical body. This remarkable skill allows the engagement in complex interactions with the environment, ranging from daily utility tasks such as writing, cutting, and wiping to professional activities such as drawing, stitching, surgical procedures, etc. [1]. Compared to the general manipulation tasks like grasping, moving, and arranging that involve direct contact and straightforward dynamics, research has shown that during tool-use actions, the brain integrates external tools into the body schema, treating them as extensions of the body [2]. This integration enhances precision and spatial awareness, enabling more effective interaction with the environment. For instance, when a blind person uses a walking stick to probe the ground, vibrations and resistance felt at the tip of the stick are transmitted to the hand, allowing the user to "sense" the surface beyond their physical reach. This perceptual extension relies on neural adaptation and tactile feedback, where mechanoreceptors relay touch signals to somatosensory neurons for interpretation [3], which enables the brain to treat tool-mediated sensations as part of the body, facilitating seamless and intuitive tool use.

In robotics, enabling artificial agents to utilize tools in a generalized manner is a fundamental challenge. While robots are increasingly deployed across diverse domains, including industrial automation and robot-assisted surgery, their tool-use capabilities remain largely task-specific. These systems typically rely on specialized end-effectors designed for predefined applications or operate within highly structured environments [4]. Therefore, granting robots the ability to use general-purpose tools designed for human hands can greatly expand their applicability. Comprehensive skills in perception, manipulation and cognition are required in robotic tool use to achieve unique functions that differ from general robot-object interaction [3]. For instance, pose adaptation becomes inherently more complex, requiring the consideration of both the gripper-tool and tool-object relationships. Moreover, unlike in-hand manipulation, where contact forces are explicitly measured and controlled, indirect force estimation plays a key role in achieving precise tool control, which involves the inference of the forces applied at the tool tip rather than exerted by the gripper.

In order to emulate human sense of touch, significant efforts have been devoted to endowing robotic systems with tactile sensing capabilities to provide rich contact information during interaction, such as textures, contact forces, stiffness and other invisible physical properties [5]. Recent advancements in visuotactile sensing technology, exemplified by sensors such as TacTip [6] and GelSight [7], have addressed the limitations of traditional tactile sensors in terms of resolution and modality compatibility. These sensors capture high-resolution, fine-grained spatial deformations, making them highly effective for providing multi-modal contact force perception [8]. Both physical-based and deep learning approaches are utilized by researchers to model the relationship between the tactile imaging and the force measurement, achieving promising estimation of normal force [9], [10], 3D force [11], [12] and force distribution [13] across various measurement ranges. However, these works mostly focus on retrieving the direct contact force between the sensor and the object, and there lacks sufficient study on estimating forces transmitted through tools.

In this work, we address the challenge of force estimation in tool-mediated robotic interactions using vision-based tactile sensing (VTS) and data-driven techniques. Traditional force control methods in tool use—such as drilling, cutting, or

Manuscript received: May, 1, 2025; Revised September, 18, 2025; Accepted October, 19, 2025.

This paper was recommended for publication by Editor Tamim Asfour and Cecilia Laschi upon evaluation of the Associate Editor and Reviewers' comments.

[1]L. Li and T.G. Thuruthel are with the Department of Computer Science, University College London, London, United Kingdom luchen.li.23@ucl.ac.uk

The source code and data used in this work can be found here: https://github.com/orionsor/tool_force_estimation

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

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

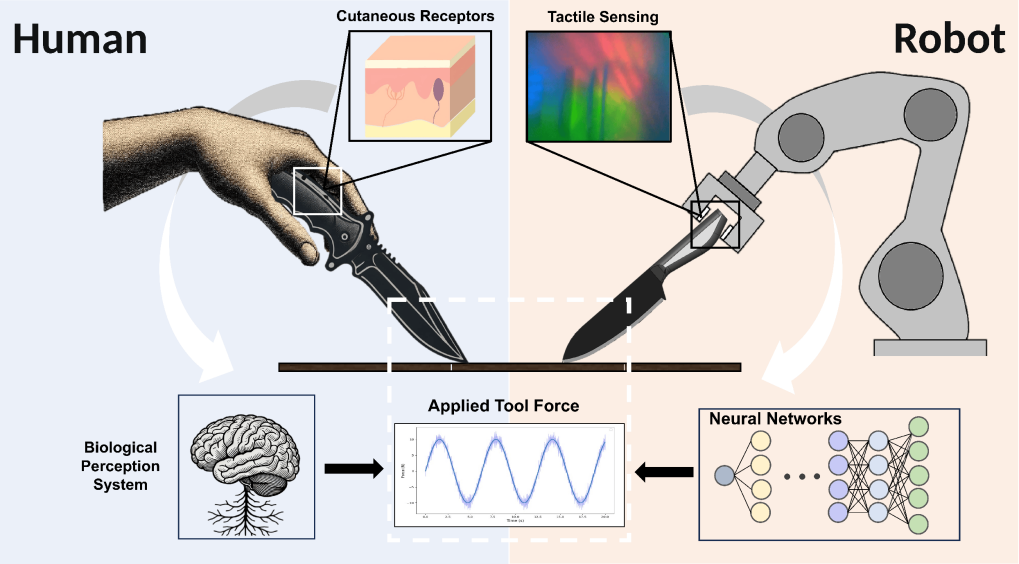

Fig. 1. Humans integrate tools into their body schema using tactile perception, where cutaneous receptors relay touch information to the brain for adaptive control. This work focuses on force estimation in tool-mediated interactions for robotics, leveraging vision-based tactile sensors and data-driven approaches.

hammering—rely on predefined interaction constraints and complex physics-based modeling of tool properties and grasps [14], [15]. However, these approaches lack adaptability and generalization, struggling with variations in tool geometry, material, and dynamic interaction poses. In contrast, leveraging the high-resolution spatial perception of VTS and deep learning enables the extraction of meaningful representations that implicitly capture tool properties and contact interactions [16], allowing force transmission dynamics to be modeled without explicit knowledge of contact mechanics. This enhances generalization across tools and interaction conditions, improving robotic dexterity in real-world scenarios. We present the first demonstration of indirect force estimation for tool-use applications using robotic systems. Our grasp-independent approach relies solely on tactile sensing and is validated on multiple tools with a single trained model. While focusing primarily on normal force estimation, we also extend our study to shear forces to demonstrate the broader applicability and robustness of the proposed method.

II. Related Work

A. Vision-based Tactile Sensing

Traditionally, resistive and capacitive sensors are widely used in research to simulate tactile responses [17]. However, they suffer from limited spatial resolution and require complex circuitry, which constrains their integration into robotic systems and hinders overall perception and operational performance [18]. Recently, Vision-based tactile sensors (VBTS) have emerged as a powerful alternative by leveraging camera-based imaging techniques to high-resolution, spatially dense contact information, which can be generally categorized into two representative types: Tactip [6] and Gelsight [7]. The TacTip sensors are based on biomimetic morphology that features embedded pins within a black, deformable membrane, utilizing an internal camera to track pin displacements. By analyzing the marker movement, Tactip sensors have demonstrated promising performance in object localization [19], edge following [20] and shear estimation [21]. Alternatively, GelSight sensors are more compact, which utilize a transparent elastomer coated with a reflective surface, whose geometry and deformations are captured with high spatial resolution using an LED illumination system and an internal CCD camera [22], enabling precise object recognition [23], contact force prediction [24], pose estimation [25], etc. In this work, we employ DIGIT, a Gelsight-like sensor known for its compact size and cost-effectiveness. Its ease of integration into various robotic systems facilitates human-inspired tactile perception in tool-use scenarios.

B. Force Sensing and Estimation with Vision-based Tactile Sensors

Achieving precise sensing and reasoning about contact forces remains a fundamental challenge in robotic manipulation and interaction. Vision-based tactile sensors (VBTS) have attracted significant attention due to their high-resolution contact representation and their ability to capture rich, multi-modal tactile information. Force estimation techniques using VBTS can be broadly categorized into physics-based models and deep learning approaches. Physics-based models aim to explicitly relate tactile image properties to force values, leveraging elasticity theory and finite element methods to derive forces from features such as marker displacement [8], RGB variations [26], indentation area [27], and optical flow vectors [28]. While these methods offer interpretability, they require precise calibration and often struggle to model complex deformations and interactions accurately.

Deep learning approaches, in contrast, construct nonlinear input/output relationships by learning implicit representations directly from tactile images [29]. For instance, MLPs have

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

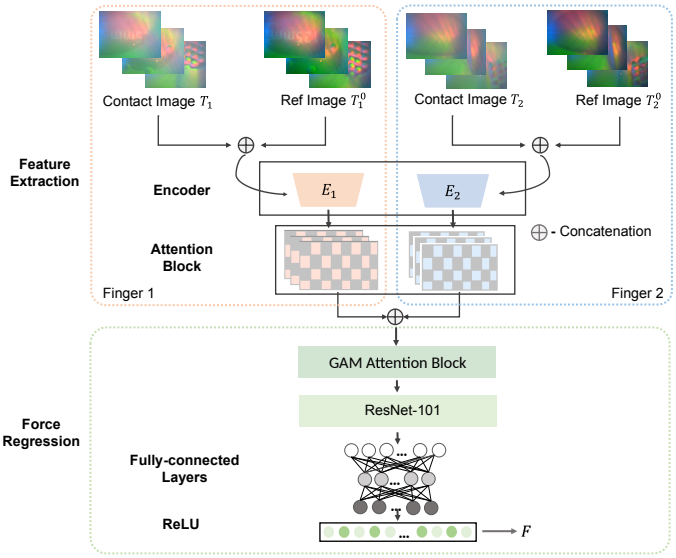

Fig. 2. Network Architecture. The network takes as input both the contact tactile images $\mathbf{T}_1$, $\mathbf{T}_2$ and the corresponding reference images under zero-force conditions $\mathbf{T}_1^0$, $\mathbf{T}_2^0$. Features are extracted using CNN-based encoders, followed by a Global Attention Mechanism to refine spatial and channel-wise representations.

been applied to estimate contact forces and torques from marker displacement features [30], while CNN-based U-Net architectures achieve high-accuracy three-dimensional force distribution estimation and generalize effectively across real-world VBTS scenarios [31]. ResNet architectures, particularly multiscale variants (MS ResNet), have addressed challenges such as gradient vanishing and improved multi-axis force estimation by capturing both local and global deformation patterns. These models have demonstrated superior performance compared to AlexNet, VGG, and DenseNet, achieving accurate predictions over a large force range of up to 30 N [12]. Incorporating reference images captured under zero-force conditions further enhances precision by providing a baseline for extracting contact-free displacement features [32]. Despite these advances, the application of VBTS for indirect force estimation, such as in tool-use scenarios, remains largely unexplored, representing a promising direction for future research.

III. Methodology

A. Problem formulation

Our objective is to estimate the force applied at the tool tip during tool-mediated interactions, specifically the normal force F_z and the shear force magnitude $F_s = \sqrt{F_x^2 + F_y^2}$, using only information from vision-based tactile sensors mounted on a two-fingered end-effector that grasps the tool. To formalize the problem, we assume the tool is rigid to ensure no delay in force transmission. Additionally, we assume the tool is securely grasped only by the tactile sensor, with no significant slip between the tool and the end-effector.

At any time instant t, as the tool is grasped to apply force F on the external object, the contact features are captured by the two tactile sensors as tactile image frames $\mathbf{T}_1^t$ and $\mathbf{T}_2^t$. These contact features reflect the interaction-specific deformation patterns for each combination of grasp pose and tool, encoding information relevant to the applied force. Corresponding reference image frames $\mathbf{T}_1^0$ and $\mathbf{T}_2^0$ captured under zero-force conditions are utilized to provide a baseline that provides information about the grasp pose and tool type. The objective is to learn a regression model that implements the desired mapping $I \in \mathbb{R}^{h\times w\times d} \rightarrow F \in \mathbb{R}$:

$$\hat{F}^t = f\left(\mathbf{I}^t; \theta\right) \tag{1}$$

$$\mathbf{I}^t = \left[\mathbf{T}_1^t, \mathbf{T}_2^t, \mathbf{T}_1^0, \mathbf{T}_2^0\right] \tag{2}$$

where $f(\cdot)$ denotes the mapping function, and θ represents the parameters of the proposed estimation model. During optimization, the Mean Squared Error (MSE) between the predicted forces and the ground truth is minimized to enhance estimation accuracy:

$$\mathcal{L} = \arg\min_{\theta} \frac{1}{n} \sum_{t=1}^{n} \left\| \hat{F}^t - F^t \right\|_2^2 \tag{3}$$

where F^t denotes the ground truth force value, and n represents the total sample numbers.

B. Tool Force Estimation Model

Fig. 2 illustrates the network architecture of the proposed tool force estimation model, which consists of two main components: Feature Extraction and Force Regression. The input of the network contains both the contact tactile images $\mathbf{T}_1$, $\mathbf{T}_2$ and the reference images $\mathbf{T}_1^0$, $\mathbf{T}_2^0$ for each sensor. Unlike the non-contact images with no deformation typically used as references in direct contact force estimation, the reference images in this work capture the initial deformations caused by securely grasping the tool while applying zero force to the external object, which provides a baseline that enables the model to effectively isolate force-induced deformations from those caused by the initial grasp.

During the feature extraction stage, the concatenated pairs of tactile and reference images from both sensors are processed by the corresponding CNN-based encoders E_1 and E_2 to extract latent features. Each encoder consists of three convolutional layers with $64, 128,$ and 256 kernels, respectively, followed by batch normalization to stabilize training and mitigate the effects of internal covariate shift. The Global Attention Mechanism (GAM) is applied both to the extracted representations of each sensor individually and to the comprehensive feature vector that integrates information from both fingers, serving distinct functions at these two levels [33]. At the level of individual sensor representations, the GAM focuses on enhancing salient spatial and channel-specific features. This allows the model to capture fine-grained deformation patterns that reflect the local contact conditions at each grasping point, prioritizing informative features indicative of force transmission. Additionally, the GAM refines the joint representation by focusing on cross-sensor correlations and global deformation patterns that arise from the integration of tactile information from both sensors, which enables the model to capture coordinated force transmission characteristics

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

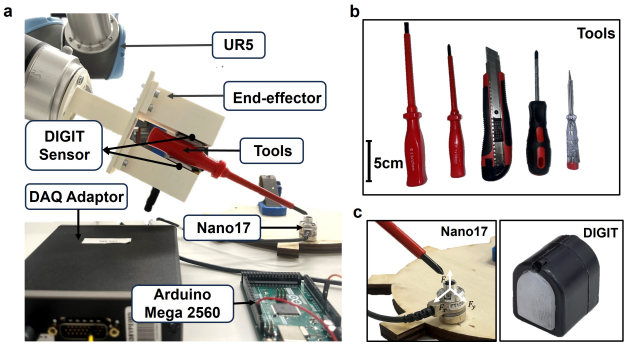

Fig. 3. Experiment setup. (a) A two-fingered end-effector equipped with DIGIT vision-based tactile sensors grasps the tool and presses it against a Nano17 force-torque sensor to collect data. Force readings are synchronized with tactile images using an Arduino Mega 2560 and a DAQ adaptor. (b) A variety of tools with different sizes, weights, and surface textures are used to evaluate model generalization. (c) left: Nano17 6-axis force-torque sensor. right: DIGIT tactile sensor.

during tool-mediated interactions that cannot be inferred from individual sensors alone.

The combined representations are subsequently fed into a regressor built on a ResNet-101 backbone, which is able to capture deep hierarchical features efficiently and to address challenges such as gradient vanishing and exploding through its residual connections. In order to perform regression, the last softmax layer of the ResNet is removed and replaced with a series of fully connected layers with ReLU activation functions to obtain the predicted force.

IV. Data Collection and Implementation

A UR5 robot arm is equipped with a custom two-finger end-effector, each mounted with a DIGIT vision-based tactile sensor. The tool is pressed against a Nano17 force-torque sensor to provide ground-truth force measurements. Data acquisition is handled via an Arduino Mega 2560 and a DAQ adaptor.

A. Experiment Settings and data collection

As illustrated in Fig. 3, the setup in our experiments consists of a custom two-fingered end-effector, each finger of which is mounted with a flexible vision-based tactile sensor DIGIT. The DIGIT sensor features a high spatial resolution of 640×480 pixels with a sensing area of $19 \times 16mm^2$ and operates at a frame rate of 30 frames per second (FPS), allowing it to capture fine-grained contact deformations in real-time [34]. During data collection, the end-effector is utilized to grasp the tool securely to apply vertical pressure onto a Nano17 six-axis force sensor, which provides the ground-truth force measurement. The raw force signals are acquired using an Arduino Mega 2560 microcontroller connected to the PC via a DAQ adaptor, which are synchronously sampled with the tactile images at a frame rate of 15 FPS. A set of daily life tools, as displayed in Fig. 3 (b) are used to improve the model's generalizability, which differ in weight, size, and surface texture. In human life, effective tool use often involves dynamically adjusting the contact pose between the end-effector and the tool to accommodate different tasks and force requirements. Therefore, multiple grasping poses are implemented for each tool, which are manually configured by varying the orientation and insertion depth of the tool within the gripper, as shown in Fig. 4 (b). This variation introduces diversity in the grasp configuration and contact geometry, aiming to cover a diverse range of interaction scenarios.

B. Dataset Overview

The collected dataset comprises a total of 39,197 pairs of tactile frames captured by the two sensors, along with the corresponding calibrated force values ranging from 0 N to 30 N. The characteristics of the dataset are illustrated in Fig. 4.

Fig. 4 (b) shows the time series data of the applied force values during multiple pressing actions, alongside the root mean square difference (RMSE) between the corresponding tactile images and the reference image obtained under zero-force conditions for both sensors. The consistent correlation between the tactile information and the force variations suggests that the deformations captured by the tactile sensors effectively reflect the tool force during interactions, which establishes a rough foundation for the proposed approach to indirectly estimate the tool-tip force based on the observed tactile feedback patterns. Additionally, the progression of tactile images is displayed in Fig. 4, covering the phases of initial contact, pressing, peak force, lifting and restoration.

In Fig. 4 (d), the force profiles for various tools and grasp poses are displayed during a single pressing action. The horizontal gray lines represent cross-sections of specific force values across different interaction scenarios, as visualized in Fig. 4 (c), which illustrates that the same force value can correspond to different tool types and grasping poses, resulting in a variety of tactile patterns. This finding highlights the non-uniqueness of the tactile responses to the applied force, posing the real challenge in estimating tool force estimation for arbitrary grasp poses and tools.

C. Implementation Details

In the proposed method, the reference images are selected as the tactile images from both sensors that correspond to the zero-force condition, providing a baseline for capturing force-induced deformations. This baseline is consistent across tactile frames obtained within the same trial. All tactile images are converted to greyscale and resized to 224×224 pixels before being fed into the model. Additionally, the calibrated force values are normalized to ensure numerical stability during the regression process and later denormalized for error calculation. Unless otherwise specified, mean absolute error (MAE) is used as the primary evaluation metric in Section V. Additionally, root mean square error (RMSE) and the coefficient of determination (R^2) are employed to further assess the model's performance. The input dataset, consisting of paired reference and tactile images along with force labels, is split into training, validation, and test sets with a ratio of 0.7:0.15:0.15. The model is trained using the Adam optimizer with a learning rate of 0.0002 and the L1 loss criterion. A batch size of 64 is used for training, and all experiments are conducted on an NVIDIA RTX 4090 GPU.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

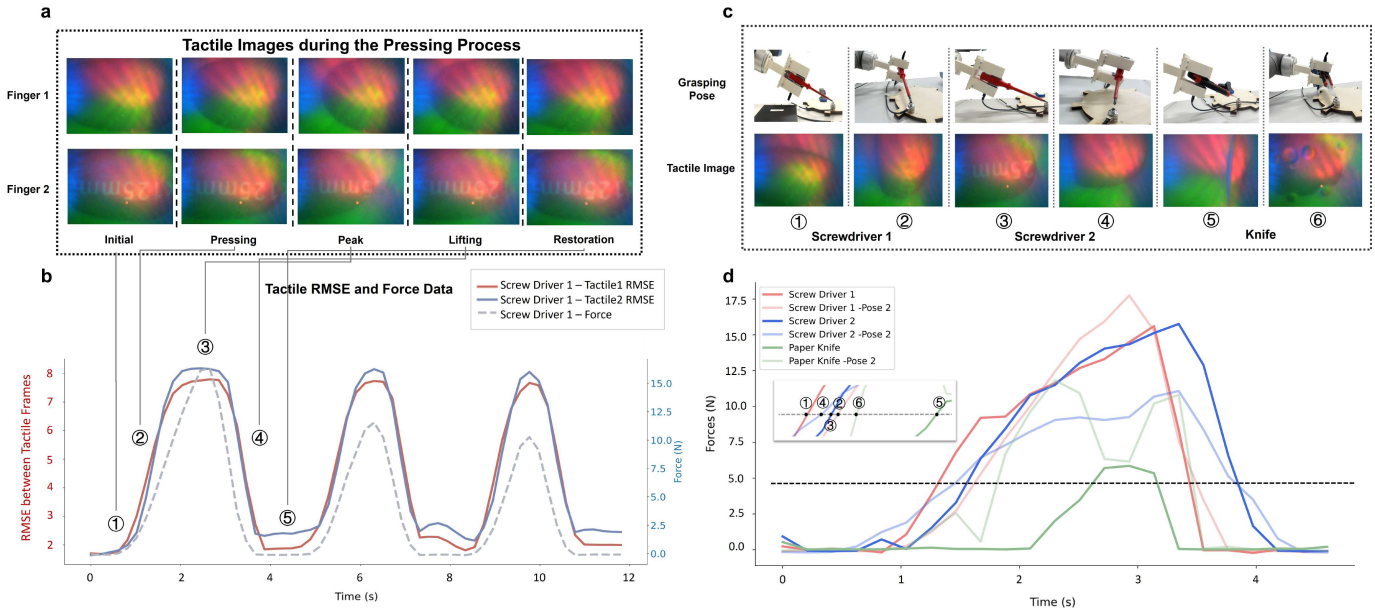

Fig. 4. Overview of the raw data. (a) Example tactile image sequences from both fingers during a pressing cycle, showing key phases: initial contact, pressing, peak force, lifting, and restoration. (b) Tactile RMSE compared with ground-truth force over time, demonstrating consistency and responsiveness of tactile signals to applied force (c) Tactile images and corresponding grasp poses for different tools. (d) Force signals for multiple tools and grasp poses. The same force magnitude (e.g., at the horizontal line) corresponds to different tactile responses, highlighting the non-uniqueness of the tactile signal and the challenge in generalizing force estimation across tool types and poses.

V. Results

In this section, we assess the effectiveness of the proposed indirect tool force estimation method by analyzing its performance across various model configurations and interaction scenarios. Specifically, we present the predictions in both data and feature space to qualitatively evaluate the alignment between the estimated forces and the ground truth. Additionally, a series of baseline models is introduced to examine the impact of incorporating reference images, dual sensor information, and the attention mechanism on force prediction accuracy across different force ranges and tool types. Furthermore, a linear regression analysis is conducted to quantify the correlation between the predicted and actual forces for each model, providing a comprehensive evaluation of the model's ability to accurately capture the underlying force transmission dynamics. Finally, we include an extension study on shear force estimation to further assess the generalizability of the proposed framework.

A. Qualitative Assessment of Tool Force Estimation

Fig. 5 presents the predicted tool force values plotted against the real forces, sorted in ascending order, on the test set. Overall, the model effectively captures the underlying force transmissions across different magnitudes, demonstrating a strong alignment between the predictions and ground truth forces. Given the inherent skewness in the dataset's force distribution, the model exhibits slightly greater deviation at force values exceeding 10 N while maintaining overall consistency across most of the force range. This behavior is expected as the imbalance in the dataset limits the model's ability to generalize to underrepresented force ranges.

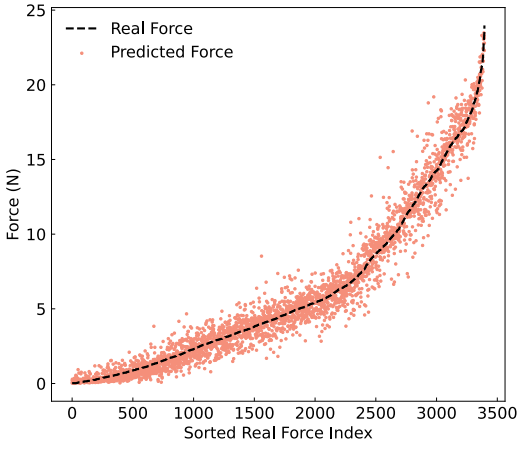

Fig. 5. Raw Prediction Result: Predicted normal forces (scatter plot) are shown against the sorted ground-truth values (black dashed line).

The model's performance is further examined through a time-series analysis across multiple pressing actions, as illustrated in Figure 6. This comparison across different model architectures provides general insights into the contribution of various model components. The proposed model, which integrates tactile information from both sensors along with a reference image capturing the initial contact state, demonstrates better stability and accuracy over time compared to the baseline models. This enhancement is particularly significant for downstream force control tasks in robotic tool use, where precise and consistent force estimation is crucial for reliable manipulation. Notably, when the model is deprived of either dual-sensor information or the initial reference frame, higher

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

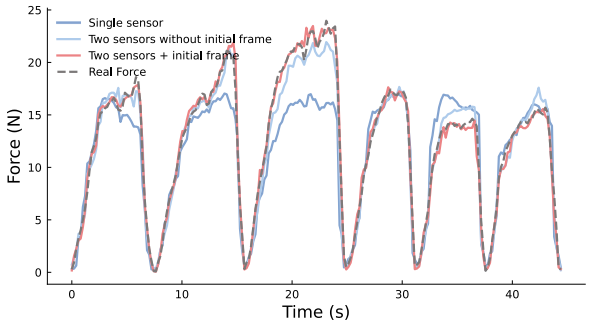

Fig. 6. Comparison of prediction performance across model variants: Force predictions over time are shown for three model configurations: Model with single sensor information (darker blue), Model with dual sensors without initial frame (lighter blue), and the proposed model combining dual sensor information with initial frame (red).

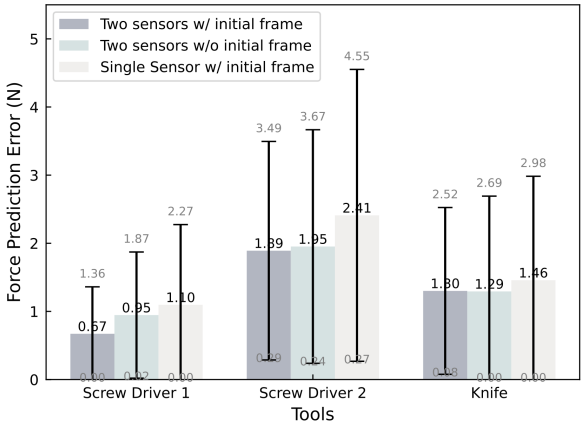

Fig. 7. Prediction Performance across different tools: The bar plot shows the tool force prediction error for three model variants evaluated on three daily household tools: screwdrivers of two sizes and the paper knife. The error bars indicate the standard deviation. The proposed model (dark gray) consistently achieves the lowest error and variance.

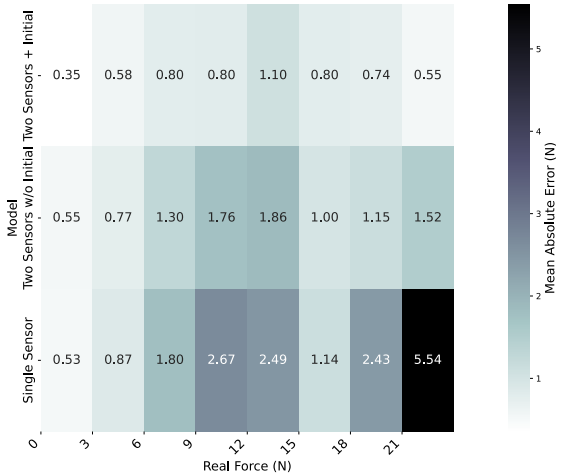

Fig. 8. Evaluation of prediction error across force ranges: The heatmap shows the error of predicted forces across different real force intervals (x-axis) for three model configurations (y-axis). The proposed model consistently achieves the best performance across all force ranges, maintaining MAE below 1.5 N.

fluctuations and deviations are observed in the predicted forces, especially at peak values, which suggests that both components are essential for providing a more comprehensive representation of interaction dynamics.

B. Generalization across Tool type

As discussed in Section IV-B, although different tools may stimulate distinct tactile responses, they can correspond to equivalent force values. Consequently, a robust tool force estimation model is supposed to generalize effectively across diverse tool types to emulate the adaptability observed in human tool use. As shown in Figure. 7, the proposed model demonstrates consistent performance across different tools, achieving relatively low mean prediction errors of 0.57 N, 1.89 N, and 1.30 N for Screwdriver 1, Screwdriver 2, and the paper knife, respectively. The differences in error magnitudes suggest that prediction accuracy is influenced by tool-specific characteristics, such as size, weight, and surface texture. Notably, Screwdriver 2, the heaviest and largest tool, exhibits the highest variability in prediction error, with a standard deviation of 3.49 N, which contrasts with Screwdriver 1, which features a lighter weight and thinner handle size. In terms of model architecture, the absence of dual-sensor combination drastically deteriorates the prediction performance across all three tools, producing a peak mean of 2.41N and a standard deviation of 4.55 N for screw driver 2.

C. Quantitative Assessment of Tool Force Estimation

The prediction error is broken down into various force range windows to investigate the error distribution. As illustrated in Fig. 8, the full model consistently yields the lowest error across all force intervals, remaining MAE below 1.5 N. Notably, the model without the initial frame exhibits increased error in the mid-to-high force ranges(above 9 N), reaching an error of 1.86 N in the 12-15 N window. The single-sensor model performs the worst, with the performance deteriorating significantly as force magnitude increases, peaking at 5.54 N for forces above 21 N.

Fig. 9 further examines the prediction accuracy through a linear regression analysis. The proposed model achieves the highest coefficient of determination R^2 of 0.9743 and MAE of 0.5964 N, indicating a strong correlation between the estimation and the ground truth. The high R^2 close to 1 value indicates that 97.43% of the variance in real force values is effectively captured by the model, highlighting its reliability and robustness across different force magnitudes. Additionally, removing the GAM attention mechanism degrades the performance, suggesting that the employment of the attention mechanism enhances the representation of force-related patterns as well as facilitates the feature fusion of the two sensors.

Given the nature of tool-mediated interactions in this study, where the tool is held at a downward angle during pressing, force transmission is inherently asymmetrical between the two tactile sensors, with this asymmetry becoming more pronounced as greater force is applied. Due to momentum effects and the non-uniform distribution of force through the tool,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

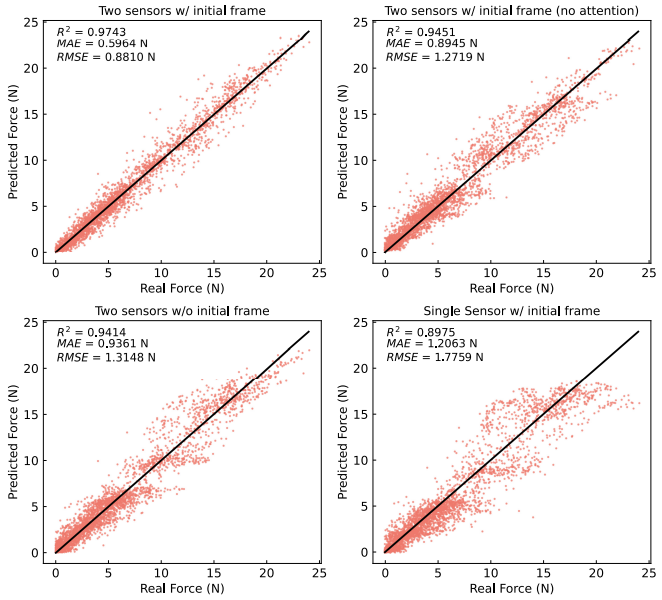

Fig. 9. Linear regression of prediction against real forces across model variants: Each subplot compares predicted normal force with ground-truth values, with the diagonal line indicating perfect prediction. The proposed model (top-left) achieves the highest accuracy, with an R^2 of 0.9743 and a MAE of 0.5964 N. Performance degrades in models using only a single sensor or lacking an initial reference frame.

each finger perceives distinct contact dynamics. Consequently, relying on a single sensor limits the model's ability to fully capture force-induced deformations, resulting in incomplete or less accurate estimations. This limitation is evident in the regression results, where the single-sensor model demonstrates the worst performance, with increasing deviations as the applied force magnitude rises.

In human tool use, tactile feedback provides the first cues about a tool's weight, stiffness, and texture before any external forces are applied. This initial haptic perception enables humans to understand a tool's physical characteristics and adapt their manipulation strategy accordingly. For instance, the flexible bristles of a paintbrush versus the rigid body of a crayon demand different grip adjustments and force applications. Before beginning a task, humans instinctively explore and refine their grasp using this early tactile information to optimize control and precision. Similarly, in robotic tool use, establishing an initial force-free reference state forms a crucial basis for accurate force estimation and adaptive manipulation. As shown in Fig. 9, omitting this reference frame leads to a greater drop in precision than removing the attention mechanism, indicating that the absence of a force-free baseline more strongly impairs prediction accuracy. While attention enhances feature extraction, it cannot compensate for the lack of a reference state, emphasizing the importance of contextual information in effective tool-use behavior.

D. Shear Force Estimation

Since real-world tool use often involves diverse force conditions such as shear and torque, we further evaluate the generalizability of the proposed approach through shear force

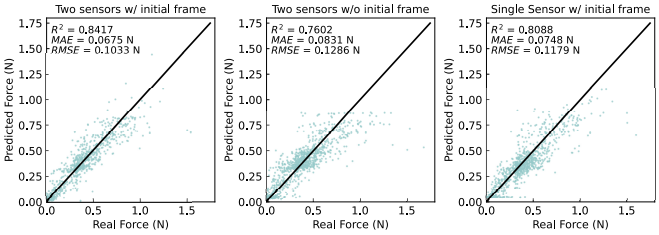

Fig. 10. Linear regression of prediction against real forces in shear force estimation.

estimation. To this end, a new dataset was collected by sliding tools across the force sensor surface to induce shear-dominant interactions. Using the same model architecture and training protocol, the model was retrained to predict the shear force magnitude $F_s = \sqrt{F_x^2 + F_y^2}$.

Fig.10 demonstrates that the proposed model maintains strong effectiveness under shear force conditions, with the full configuration achieving the best results across all three metrics. The decline in accuracy when omitting the initial reference frame further emphasizes the importance of context in force interpretation. Notably, in shear-based tool interactions, deformation tends to concentrate on the finger opposite the direction of motion. Despite this asymmetry, the inclusion of both tactile sensors improves estimation, suggesting that even the less-deformed sensor contributes contextual information, such as relative motion cues or grasp stability. This reflects a parallel in human tool use, where two fingers jointly gather complementary tactile information that helps disambiguate contact conditions and guide force modulation. These findings demonstrate the adaptability of our framework to alternative force directions and interaction modalities, further supporting its applicability to a wider range of tool-use scenarios.

VI. Conclusion

Humans exhibit a remarkable ability to seamlessly integrate tools into their actions, effectively extending their sensory and motor capabilities. This skill enables precise force control in complex, tool-mediated tasks, where tactile feedback is essential for interpreting force transmission through the tool. Inspired by this capability, our work demonstrates indirect tool force estimation for robotics, leveraging vision-based tactile sensing (VTS) and deep learning to infer applied forces without direct measurement at the tool tip. Through both qualitative and quantitative evaluations, our model effectively captures force transmission dynamics, achieving promising results with a mean absolute error (MAE) of 0.5964 N across a force range of 0 N to 24 N. Moreover, the proposed model generalizes across tools with varying sizes, weights, and surface properties, maintaining consistent performance despite differences in tool characteristics and grasping poses. This demonstrates robustness to diverse force transmission patterns.

Through a comparative study of model components, we identify key factors influencing tool force estimation accuracy, grounded in the nature of tool-use behavior and its biological inspiration. Results show that dual-sensor input is critical

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

for capturing asymmetric force perception, while an initial zero-force reference frame provides essential context to distinguish grasp-induced from externally applied deformations. The attention mechanism further enhances feature extraction and estimation accuracy. Finally, successful extension to shear force estimation confirms the generalizability and flexibility of our framework across diverse force modalities encountered in real-world tool use.

While this study confirms the feasibility of indirect tool force estimation through VTS, several areas remain open for future research. One key challenge is slippage detection, as unintended tool motion relative to the sensor or external object can affect estimation accuracy. In addition, force transmission in tool use involving torsional forces remains insufficiently explored. Future work will therefore address more complex interaction conditions, including slip, torsion, and combined loading scenarios. The proposed model also provides a promising perception module that can be integrated with advanced control strategies, enabling robots to perform a broader range of tool-mediated manipulation tasks. Overall, this work represents a significant step toward robotic systems with dexterity and adaptability comparable to human tool users.

REFERENCES

[1] C. Nabeshima, Y. Kuniyoshi, and M. Lungarella, "Adaptive body schema for robotic tool-use," *Advanced Robotics*, vol. 20, no. 10, pp. 1105–1126, 2006.

[2] P. Lanillos and G. Cheng, "Adaptive robot body learning and estimation through predictive coding," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4083–4090.

[3] T. Heed, "Tool use: Two mechanisms but one experience," *Current Biology*, vol. 29, no. 24, pp. R1301–R1303, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982219314344

[4] N. Kumar, N. Hack, K. Dörfler, A. Walzer, G. Rey, F. Gramazio, M. Kohler, and J. Buchli, "Design, development and experimental assessment of a robotic end-effector for non-standard concrete applications," 05 2017.

[5] M. Park, B.-G. Bok, J.-H. Ahn, and M.-S. Kim, "Recent advances in tactile sensing technology," *Micromachines*, vol. 9, no. 7, p. 321, 2018.

[6] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

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

[8] B. Fang, J. Zhao, N. Liu, Y. Sun, S. Zhang, F. Sun, J. Shan, and Y. Yang, "Force measurement technology of vision-based tactile sensor," *Advanced Intelligent Systems*, vol. 7, no. 1, p. 2400290, 2025.

[9] M. Li, T. Li, and Y. Jiang, "Marker displacement method used in vision-based tactile sensors—from 2-d to 3-d: A review," *IEEE Sensors Journal*, vol. 23, no. 8, pp. 8042–8059, 2023.

[10] L. Van Duong, "Bitac: A soft vision-based tactile sensor with bidirectional force perception for robots," *IEEE Sensors Journal*, vol. 23, no. 9, pp. 9158–9167, 2023.

[11] B. Fang, F. Sun, C. Yang, H. Xue, W. Chen, C. Zhang, D. Guo, and H. Liu, "A dual-modal vision-based tactile sensor for robotic hand grasping," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 4740–4745.

[12] Z. Lu, T. Yang, Z. Cao, D. Luo, Q. Zhang, Y. Liang, and Y. Dong, "Optical soft tactile sensor algorithm based on multiscale resnet," *IEEE Sensors Journal*, vol. 23, no. 10, pp. 10 731–10 738, 2023.

[13] L. Zhang, Y. Wang, and Y. Jiang, "Tac3d: A novel vision-based tactile sensor for measuring forces distribution and estimating friction coefficient distribution," *arXiv preprint arXiv:2202.06211*, 2022.

[14] P. Naisson, J. Rech, and H. Paris, "Analytical modeling of thrust force and torque in drilling," *Proceedings of the Institution of Mechanical Engineers, Part B: Journal of Engineering Manufacture*, vol. 227, no. 10, pp. 1430–1441, 2013.

[15] T. M. Williamson, B. J. Bell, N. Gerber, L. Salas, P. Zysset, M. Caversaccio, and S. Weber, "Estimation of tool pose based on force–density correlation during robotic drilling," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 4, pp. 969–976, 2012.

[16] C. Sferrazza, A. Wahlsten, C. Trueeb, and R. D'Andrea, "Ground truth force distribution for learning-based tactile sensing: A finite element approach," *IEEE Access*, vol. 7, pp. 173 438–173 449, 2019.

[17] Z. Kappassov, J.-A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands," *Robotics and Autonomous Systems*, vol. 74, pp. 195–220, 2015.

[18] S. Zhang, Z. Chen, Y. Gao, W. Wan, J. Shan, H. Xue, F. Sun, Y. Yang, and B. Fang, "Hardware technology of vision-based tactile sensor: A review," *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21 410–21 427, 2022.

[19] N. F. Lepora and B. Ward-Cherrier, "Superresolution with an optical tactile sensor," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 2686–2691.

[20] N. F. Lepora and J. Lloyd, "Pose-based tactile servoing: Controlled soft touch using deep learning," *IEEE Robotics & Automation Magazine*, vol. 28, no. 4, pp. 43–55, 2021.

[21] A. K. Gupta, L. Aitchison, and N. F. Lepora, "Tactile image-to-image disentanglement of contact geometry from motion-induced shear," in *Conference on Robot Learning*. PMLR, 2022, pp. 14–23.

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

[23] B. Wang, B. Li, L. Li, Z. Zhang, S. Qiu, H. Wang, and X. Wang, "Object recognition using shape and texture tactile information: A fusion network based on data augmentation and attention mechanism," *IEEE Transactions on Haptics*, 2024.

[24] Z. Lu, Z. Liu, X. Zhang, Y. Liang, Y. Dong, and T. Yang, "3d force identification and prediction using deep learning based on a gelsight-structured sensor," *Sensors and Actuators A: Physical*, vol. 367, p. 115036, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924424724000293

[25] Y. Gao, S. Matsuoka, W. Wan, T. Kiyokawa, K. Koyama, and K. Harada, "In-hand pose estimation using hand-mounted rgb cameras and visuotactile sensors," *IEEE Access*, vol. 11, pp. 17 218–17 232, 2023.

[26] W. Chen, Y. Yan, Z. Zhang, L. Yang, and J. Pan, "Polymer-based self-calibrated optical fiber tactile sensor," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 10 197–10 203.

[27] Y. Yang, M. Zhao, Y. Jia, L. Zhao, D. Piao, Y. Zheng, and L. Song, "A tactile sensor with slippage prediction by unequal-height dome array," *IEEE Sensors Journal*, vol. 23, no. 16, pp. 17 958–17 967, 2023.

[28] G. Zhang, Y. Du, H. Yu, and M. Y. Wang, "Deltact: A vision-based tactile sensor using a dense color pattern," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 778–10 785, 2022.

[29] J. Castaño-Amoros and P. Gil, "Grasping force estimation for markerless visuotactile sensors," *IEEE Sensors Journal*, 2024.

[30] Y. Zhang, Z. Kan, Y. Yang, Y. A. Tse, and M. Y. Wang, "Effective estimation of contact force and torque for vision-based tactile sensors with helmholtz–hodge decomposition," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4094–4101, 2019.

[31] C. Sferrazza and R. D'Andrea, "Sim-to-real for high-resolution optical tactile sensing: From images to three-dimensional contact force distributions," *Soft Robotics*, vol. 9, no. 5, pp. 926–937, 2022.

[32] H. Sun, K. J. Kuchenbecker, and G. Martius, "A soft thumb-sized vision-based sensor with accurate all-round force perception," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 135–145, 2022.

[33] Y. Liu, Z. Shao, and N. Hoffmann, "Global attention mechanism: Retain information to enhance channel-spatial interactions," *arXiv preprint arXiv:2112.05561*, 2021.

[34] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.