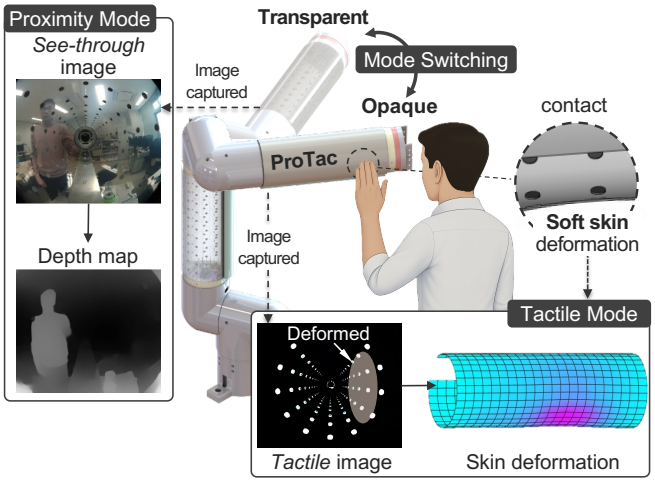
Vision-based Proximity and Tactile Sensing for Robot Arms: Design, Perception, and Control

Quan Khanh Luu[1], Dinh Quang Nguyen[2], Nhan Huu Nguyen[1], Nam Phuong Dam[1], and Van Anh Ho[1*], *Senior Member, IEEE*

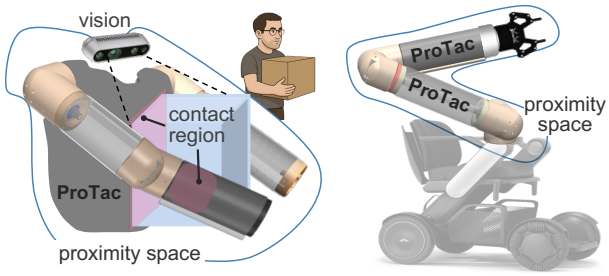
***Abstract*—Soft-bodied robots with multimodal sensing capabilities hold promise for versatile and user-friendly robotics. However, seamlessly integrating multiple sensing functionalities into soft artificial skins remains a challenge due to compatibility issues between soft materials and conventional electronics. While vision-based tactile sensing has enabled simple and effective sensor designs for robotic touch, there has been limited exploration of this technique for intrinsic multimodal sensing in large-sized soft robot bodies. To address this gap, this paper introduces a novel vision-based soft sensing technique, named ProTac, capable of operating either in tactile or proximity sensing modes. This vision-based sensing technology relies on a soft functional skin that can actively switch its optical properties between opaque and transparent states. Furthermore, the paper develops efficient learning pipelines for proximity and tactile perceptions, as well as sensing strategies enabled through the timing activation of the two sensing modes. The effectiveness of the soft sensing technology is demonstrated through a soft ProTac link, which can be integrated into newly constructed or existing commercial robot arms. Results suggest that robots integrated with the ProTac link, along with rigorous control formulation can perform safe and purposeful control actions, which enhances human-robot interaction scenarios and facilitates motion control tasks that are challenging to achieve with conventional rigid links.**
Supplementary video: https://youtu.be/5DhAhlTVxzg

I. Introduction

Nowadays, human-friendly robots are required to operate out of the safety fence zone, collaborating with humans in complex or physically demanding work. Moreover, they will soon become versatile and flexible service assistants that are integral parts of our daily lives. Concerning the essential nature of these foreseen applications, the robots are expected to operate in dynamic, unknown environments, sharing the same workspace, and even physically interacting with humans, as in the so-called human-robot interaction (HRI) scenario. In the HRI, the robots should be able to not only react to possible collisions but also handle both unavoidable and intentional physical contacts in a safe and purposive way [1], [2]. In order to accomplish this, efforts have been made on the invention of novel mechanical designs of robot links and actuators, aimed at passively reducing physical impacts based on compliant and lightweight structures. In addition, perception coming from external sensors or intrinsic sensing mechanisms permits reliable recognition of surroundings or detection of contact events, which benefits the development of motion planning and control regimes possible to react purposefully in the HRI scenarios.



(a) Illustration of ProTac multi-modal *soft* sensing technology



(b) ProTac's potential applications

Fig. 1. **Research overview.** (a) Conceptual overview of the vision-based ProTac sensing technology. ProTac can actively switch between proximity and tactile sensing modes, relying on input images captured by inner cameras and a *soft* functional skin with controllable transparency. (b) Illustration of ProTac's prospective applications.

Among the robotics technologies, the design of robots featuring large-area sensitive skins that can provide multi-modal perception is considered to be one of the crucial factors [11], [12]. Of sensing modalities, the sense of touch at large-area could bring in a variety of tactile information from interactions involving physical contacts, such as multiple applied forces, contact locations, complex touch patterns, or contact classification, which delivers an effective means for handling contact events, especially in physical HRI (pHRI) [13], [14]. Furthermore, proximity perception has recently

[1]Soft Haptics Lab., School of Materials Science, Japan Advanced Institute of Science and Technology, Asahidai 1-1, Nomi, Ishikawa, 923-1292 Japan.
[2]VNU University of Engineering and Technology, E3 Building, 144 Xuan Thuy Street, Cau Giay District, Ha Noi, Vietnam.
Data, source code for this article are published at: https://github.com/Ho-lab-jaist/protac.git
*Corresponding author. Email: van-ho@jaist.ac.jp

TABLE I
HARDWARE COMPARISON OF PROTAC AND OTHER REPRESENTATIVE SOFT BIMODAL VISION-BASED SENSORS

Sensor Design	Sensing Skin			Imaging System	Modality Switch Mechanism
	Material	Shape	Scale		
FingerVision [3]	Transparent gel	2D	Small	RGB	No switch (sync)
Jessica [4]	Translucent elastomer	2D	Small	RGB + Infrared ToF	Light switch
SpecTac [5]	Transparent gel	2D	Small	RGB	Light switch
Finger-STS [6]	Unidirectional perspective gel	2D	Small	RGB	Light switch
VisTac [7]	Transparent gel	2D	Small	RGB	Light switch
TIRgel [8]	Transparent gel	2D	Small	RGB	Focus adjustment
StereoTac [9]	Translucent elastomer	2D	Small	Stereo RGB	Light switch
M^3Tac [10]	Unidirectional perspective latex	2.5D	Small	RGB + Near-IR + Mid-IR	No switch (sync)
ProTac (Ours)	Bilayer PDLC skin	3D	Large	Stereo RGB	Skin transparency switch

attracted attention since its capability of closing perception gaps left by occlusions and blind spots in vision, and inaccessible pre-touch information in tactile modality [15]. Previous studies attempted to set up numerous sensors on the robot body, or introduced sophisticated sensor designs to satisfy the aforementioned requirements. However, such designs faced complexity in integration and data processing or were merely impractical for whole-body sensing and interaction. Therefore, to both reduce complexity in fabrication and increase sensing capabilities in HRI, it necessitates a novel, yet simple design solution for a *soft* sensing device that could deliver *multimodal* tactile and proximity perception on a large-scale, whole-arm robot's body.

This paper tackles this challenge by developing a vision-based proximity-tactile sensing technique with a *soft* functional skin named **ProTac**. This technique enables sensing operations in either *tactile* or *proximity* modes. Specifically, the technology is enabled through a soft PDLC[1] skin, which can actively switch its optical properties between opaque and transparent states. In the *opaque* state, tactile/contact sensing is facilitated by processing tactile images captured from an inner camera without interference from external light conditions. In contrast, the *transparent* state enables proximity perception by allowing the inner cameras to see through the skin and observe external obstacles. In this paper, we demonstrate the application of ProTac technology in a soft multi-modal sensing robot link, referred to as ProTac link. This study also showcases the utilization of such links for robot arms, facilitating safe and multi-modal robot tasks, particularly for enhancing challenging motion control and purposeful human-robot interaction scenarios.

A. Related work

1) Electronic skin: Tactile robotic (electronic) skins, exploited arrays of distributed sensing elements with various mechanotransduction principles (*e.g.*, resistance, capacitance, inductance, electromagnetic field strength, and light density [13]), have recently received considerable interest, since their surface conformability and scalability, which enables them to integrate on a wide range of robot parts, ranging from small-scale robotic fingers to larger body areas, such as whole-limb, or torso [16]–[19]. Notably, a large-scale electronic skin established by a network of rigid hexagonal printed circuit boards could provide sensory feedback with multiple modalities, including proximity, vibration, temperature, and light touch [20], that was useful in a variety of control frameworks and applications [21]–[23]. However, due to the fact that such sensor networks are formed by spatially distributed sensing modules, integrating various sensors and electronic components, which may cause difficulty in fabrication. In addition, the data acquisition and processing of such sensor networks were highly complex, and existing proximity sensations achieved through magnetic or capacity transduction often behaved differently according to the material properties of target objects, which may cause difficulties in calibration and perception.

2) Vision-based sensing skin: Vision-based sensing technology has recently emerged as an effective method for enabling robotic touch perception with minimal wiring and electronics, offering high spatial resolution at low cost [24]. For instance, a family of GelSight sensors could measure the high-fidelity surface texture of a touched object, which is reconstructed from images rendered from RGB illumination using photometric stereo algorithms [25]–[29]. GelForce sensor and its successors are made of elastomeric layers embedded with reflective markers tracked for inference of traction fields or force distribution, which requires a learning or calibration process to map the relationship between the applied force and the movements of markers [30], [31]. TacTip family in the form of a hemispherical shape could be learned to perform edge-aware contour following, manipulation, and exploration, mostly based on changes in positions of printed markers [32]. Recently, [33] introduced TacLink, a vision-based soft tactile link with a large sensing area, where tactile perception is efficiently learned using the recently developed sim-to-real learning platform [34]. Notably, a wide range of bimodal vision-based sensors has been recently developed to provide both tactile and visual information through internal imaging systems and transparent membranes [3], [10]. Specifically, [6]–[9] used contrasts between external and internal lighting to switch between visual and tactile sensing modes, achieved through transparent and opaque skin states, respectively. Another technique employed internal UV light to illuminate markers, creating a visually ambient membrane for tactile sensing [4], [5]. However, due to their reliance on light-dependent switching mechanisms, these existing designs are susceptible to external light conditions, potentially limiting their perceptual capabilities. Moreover, bimodal vision-based

[1]PDLC stands for Polymer Dispersed Liquid Crystal film

sensing for large-area devices with complex skin structures, such as robotic links, remains underexplored.

In response to this gap, we introduce ProTac link, a soft bimodal robot link that can actively switch between proximity and tactile sensing modes through controllable skin transparency. In ProTac, this transparency is achieved through the intrinsic optical properties of PDLC skin rather than lighting conditions, thereby reducing the impact of external visual background on tactile sensing performance. Table I summarizes and contrasts key features of the ProTac link with other representative bimodal vision-based sensors, highlighting its focus on large-area, 3D skin structures – an underexplored direction in visuotactile sensing for soft robotic skins.

B. Contribution

Building upon our previous work on single-mode vision-based tactile sensing [34], this paper presents a bimodal vision-based soft link (ProTac) that enables both proximity and tactile sensing, along with new multimodal perception algorithms using stereo vision. The preliminary evaluation of ProTac link was demonstrated in our previous work [35]; however, the design and fabrication methods necessary to ensure structural robustness and effective operational conditions were not thoroughly addressed. Also, the data-driven sensing algorithms developed in [35] lacked generalizability due to reliance on sensor-specific real-world training data and did not consider strategies for fusing stereo vision inputs. Moreover, the ProTac's capability of multi-point contact detection and proximal awareness across various nearby obstacles has not been investigated. Lastly, specific sensing strategies and potential use cases of the ProTac link for control tasks remain underexplored. Thus, the main contributions of this paper are summarized as follows:

1) Development of a learning platform to enable ProTac perceptions, where proximity and tactile perceptions are enabled through monocular depth-map estimation and zero-shot sim-to-real learning of soft skin deformation, respectively.
2) Integration of ProTac sensing for two safety control strategies, including creating reflex behavior and adaptive proximity-based speed regulation. The effectiveness is demonstrated using a ProTac-integrated robot arm.
3) Showcase of ProTac-specific sensing strategies with two multimodal tasks, which aim to enhance motion control in cluttered environments and facilitate seamless human-robot interaction scenarios.
4) Release of open-source design files, models, and code to support further research.[2]

II. Hardware Design

This paper showcases the implementation of ProTac with a large-scale sensor design of cylindrical skin shape (referred to as ProTac link, see Fig. 2). The choice of this design is made due to its resemblance to the links of lightweight industrial robot arms, offering practicality for demonstrating the effectiveness of sensing algorithms and control strategies.

[2]https://github.com/Ho-lab-jaist/protac.git

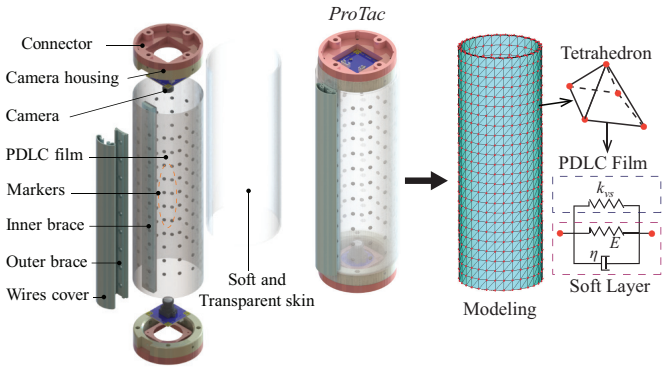


Fig. 2. **Design of ProTac link**. Left: Exploded view showing detailed parts and complete assembly of ProTac. Right: FE modeling of the ProTac skin, with the soft layer represented by the Kelvin-Voigt model characterized by Young's modulus E and viscosity η, while the mechanical properties of the PDLC film are simulated using a virtual spring with stiffness k_{vs}.

A. Design and Basic Working Principle

The design of a ProTac link is illustrated in Figure 2. Proximity-tactile sensing with mode-switching capability is enabled through internal cameras attached at the two ends and a soft functional skin that can switch its optical properties between opaque and transparent states. To achieve this, the soft skin is constructed with a layered structure consisting of an outer transparent silicone layer and an inner thin flexible polymer-dispersed liquid crystal (PDLC) film on which arrays of reflective markers are attached (see Fig. 2). The outer layer is designed to be soft and transparent to enhance the pleasantness of physical interaction and enable the see-through function of the ProTac skin. The inner PDLC film's opaque and transparent states can be actively switched by applying an external voltage. The transition time of state switching is as fast as $0.3\,\mathrm{s}$. Given that, the basic working principle of ProTac is as follows:

- *Proximity* mode: When the PDLC skin is in the *transparent* state, the internal cameras can see through the skin so that the proximity information of nearby obstacles can be extracted from image views (see Section III, Fig. 4).
- *Tactile* mode: As the soft PDLC skin switches to the *opaque* state, the camera captures marker-featured tactile images to infer tactile information, enabling robust operation without interference from external backgrounds. (see Section IV, Fig. IV).

Note that internal LED lights are used to illuminate the reflective markers during tactile operation but are turned off during proximity sensing to minimize reflectance and the markers' visual impact on the external scene.

B. Fabrication

The proposed fabrication process is guided by the desired durability and payload capacity of the soft ProTac link. Here, structural analysis is conducted in simulation to ensure its robustness under loads below $15\,\mathrm{N}$. Furthermore, the transparency, uniformity of the soft skin, and marker reflectivity are crucial specifications affecting the see-through ability and

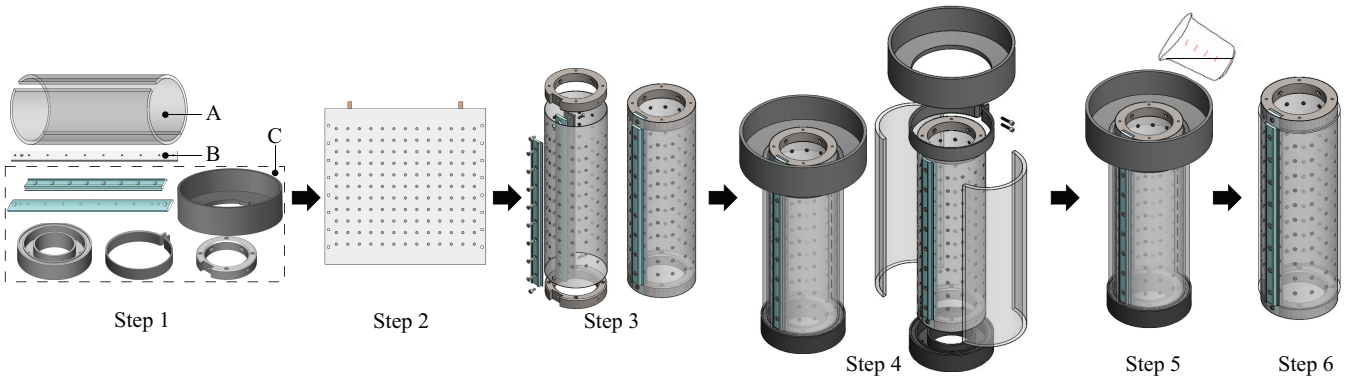


Fig. 3. **Fabrication process of the ProTac link**. Step 1 - Preparing parts (A - Part was fabricated by laser cutting. B - Part was fabricated by machining cutting. C - Part was fabricated by 3D printing technique). Step 2 - Reflective markers arrangement onto a PDLC film. Step 3 - Shaping the PDLC film. Step 4 - Molding assembly. Step 5 - Pouring deformable and transparent silicone. Step 6 - Releasing mold for a finished ProTac sensor.

contact detection in proximity and tactile modes, respectively. In this regard, a commercial acrylic tube with a smooth surface finish is utilized for the outer mold component to guarantee high transparency for the outer soft skin, thereby improving the efficiency of the see-through effect in proximity sensing. Additionally, to ensure the ProTac's performance in tactile mode, markers with a 3 mm diameter are created using reflexive tape (R25 WHI, 3M Company). With these specifications in mind, we devise six main steps for the fabrication process, categorized into preparation, molding, and releasing stages (see Fig. 3). In the first step, the outer mold (part A) is laser-cut into two halves, facilitating easier release upon separation. In step two, we fabricate reinforcing braces (part B) from steel using a machining process, while other parts (C) are 3D-printed with PLA material. Step three involves adhering a marker matrix to the PDLC film (LC Magic, TOPAN Inc., Japan) and shaping it into a cylindrical skeleton with camera housing at both ends, bolted through outer and inner braces. Subsequently, the skeleton is covered with all molded parts in step four. The outer soft skin is constructed by filling the mold with silicone liquid made from transparent silicone rubber (Zoukei-mura, Japan) and curing it for a minimum of 24 hours in step five. The final step involves removing all mold components to obtain the complete ProTac skin. Lastly, to ensure the seamless integration of the ProTac link with custom-built and commercial robot arms, all connected power and signal cables for the motors and cameras can be routed along the braces. Once arranged, 3D-printed covers will be installed over the braces to secure and conceal the cables.

III. Proximity Perception

The proximity perception is activated when the ProTac's skin is at transparent mode. While methods for distance measurement obtained from off-the-shelf RGB-D cameras or binocular/multi-view vision have been intensively examined in the past [36], [37], the inference from a critical arrangement of opposite cameras, as in the case of the ProTac link, has been barely investigated. This section describes a method to either estimate the distance from ProTac skin to the closest obstacle or evaluate the risk of collisions with surroundings, by processing the ProTac's internal camera view when the PDLC skin is in the *transparent* state (see Fig. 4). Specifically, we employ a data-driven monocular depth estimation based on a DNN [38] to generate a depth map of the external space from the ProTac transparent view (Section III-A), which forms the basis for distance measurement (Section III-B) and risk evaluation (Section III-C). This scheme enables the independent observation of obstacles from any direction using each of the two opposite cameras, thereby expanding sensing coverage and enhancing applicability to other sensor designs. The fusion of sensing information from multiple camera views to enhance the sensing performance is discussed in Section III-D.

A. Monocular Depth Estimation

In this paper, the projection between ProTac images and estimated depth maps is achieved by a monocular depth estimation network (DepthNet) trained through supervised learning. To learn this projection, we fine-tuned the pre-trained MiDas model [38] on MannequinChallenge dataset[3] [39]. Specifically, we employed a well-established multi-view stereo pipeline (COLMAP-based MVS) proposed in [39] to generate ground-truth depth maps $\mathbf{Z}^{\text{gt}}$ from the image sequences extracted from the *MannequinChallenge* dataset. Simultaneously, the image dataset was synthetically augmented using the alpha blending technique to replicate the see-through views $\mathbf{I}^{\text{prox}}$ (see Fig. 4). Consequently, the DepthNet, initialized with the MiDas model's weights, was trained to map the augmented images $\mathbf{I}^{\text{prox}}$ to the ground-truth depth images $\mathbf{Z}^{\text{gt}}$.

Loss Function: For the training process, the computed depth might yield an arbitrary scale. To address this, we adopt a scale-invariant depth regression loss, as proposed in [38]. We also ignore uncertain depth pixels in the ground truth, particularly in occluded regions caused by ProTac mechanical parts, to improve learning performance. Let $\mathbf{I}^m = (\mathbf{I}^m_j \in \{0,1\}, \forall j \in \{1,2,\cdots,a\times b\}) \in \mathbb{Z}_2^{a\times b}$ represent the mask of the mechanical structures covering the transparent view $\mathbf{I}^{\text{prox}} \in \mathbb{R}^{a\times b\times 3}$ of ProTac, where $\mathbf{I}^m_j = 0$ denotes an occluded pixel. We define a list of indexes for occluded pixels as $\mathcal{K} := \{j \mid \mathbf{I}^m_j = 0, \forall j \in \{1,2,\cdots,a\times b\}\}$. Given a

[3]MannequinChallenge is a compilation of video clips of frozen people imitating mannequins, publicly opened on YouTube by Google AI.

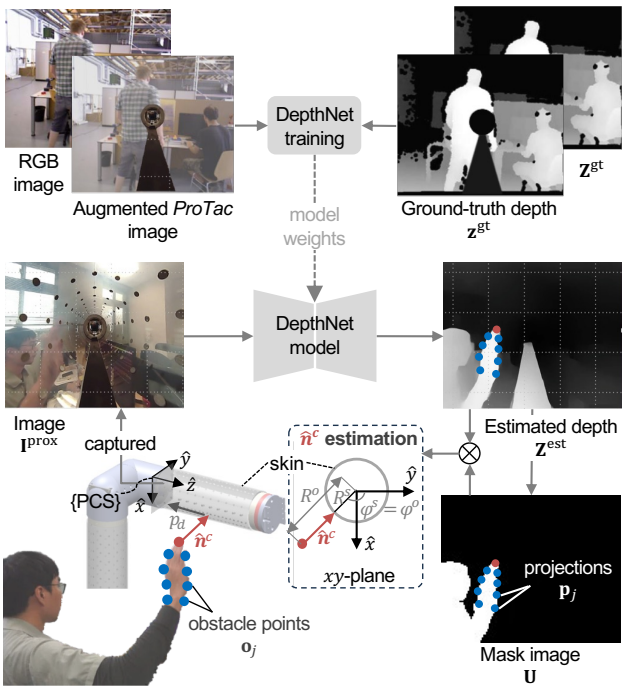


Fig. 4. **Illustration of proximity processing pipeline**. The DepthNet model is fine-tuned on augmented ProTac images using open-source datasets. The distance to the ProTac skin $\hat{\boldsymbol{n}}^c$ is estimated based on the depth-map estimation $\mathbf{Z}^{\text{est}}$ and the extracted mask image $\mathbf{U}$ through an image processing technique. Note that the obstacle points $\mathbf{o}_j$ and their projections $\mathbf{p}_j$ in the illustration may not correspond to real data points, and not all points are presented.

raw estimated and ground-truth depth map $\mathbf{Z}^{\text{est}}, \mathbf{Z}^{\text{gt}} \in \mathbb{R}^{a\times b}$, the valid depth estimation and ground truth can be defined respectively as $\mathbf{z}^{\text{est}} := (\mathbf{Z}^{\text{est}}_k, \forall k \in \mathcal{K})$, $\mathbf{z}^{\text{gt}} := (\mathbf{Z}^{\text{gt}}_k, \forall k \in \mathcal{K})$. Thus, the scale-invariant regression loss can be derived as:

$$\mathcal{L}_{\text{DepthNet}} = \mathcal{L}_{\text{ssitrim}}(\mathbf{z}^{\text{est}}, \mathbf{z}^{\text{gt}}) + \alpha \mathcal{L}_{\text{grad}}(\mathbf{z}^{\text{est}}, \mathbf{z}^{\text{gt}}). \tag{1}$$

where, the first term $\mathcal{L}_{\text{ssitrim}}$ penalizes the absolute difference in depth values between $\mathbf{z}^{\text{est}}$ and $\mathbf{z}^{\text{gt}}$, and the second multi-scale gradient term $\mathcal{L}_{\text{grad}}$ encourages sharp depth discontinuities and smooth gradient changes. The detailed derivation of these loss terms is presented in Appendix A.

Network Architecture and Training: DepthNet model for monocular depth estimation is designed upon the ResNet multi-scale architecture [40]. We initialize the DepthNet with the model weights as mentioned in [38]. For the fine-tuning process, we use Adam optimizer with the learning rate initialized at 10^{-4}, then linearly decaying at the 50$^{\text{th}}$ iteration out of a total of 100 training steps. The hyperparameter α in the combined loss function (1) is experimentally set to 0.1. Detailed network architecture can be found in [40].

B. Distance Estimation

Here, we describe a procedure to extrapolate the distance between an external obstacle and the ProTac link based on the estimated depth map $\mathbf{Z}^{\text{est}}$ obtained from the fine-tuned DepthNet model.

To achieve this goal, the mask image $\mathbf{U} \in \mathbb{Z}_2^{a\times b}$ of nearby obstacles is initially extracted from $\mathbf{Z}^{\text{est}}$ using binary thresholding. This assumes that obstacles in proximity would exhibit distinguishable, brighter pixel intensities. Given a set of obstacle points representing the nearby obstacles on the mask image $\mathbf{U}$, indexed by

$$\mathcal{O} = \{j \,|\, \mathbf{U}_j \wedge \mathbf{I}^m_j = 1, \forall j \in \{1, 2, \cdots, a \times b\}\}, \tag{2}$$

the obstacle points in 3D space $\mathbf{O} = [\mathbf{o}^\intercal_j, \forall j \in \mathcal{O}] \in \mathbb{R}^{|\mathcal{O}|\times 3}$ can be computed from their projections $\mathbf{P} = [\mathbf{p}^\intercal_j, \forall j \in \mathcal{O}] \in \mathbb{R}^{|\mathcal{O}|\times 3}$ on the depth image $\mathbf{Z}^{\text{est}}$ using the pinhole model of a ProTac camera (see Appendix B).

The problem is to determine the normal distance from all obstacle points $\mathbf{o}$ to the skin surface (subscript j is temporarily omitted for brevity). For ease of calculation, the Cartesian coordinates of the obstacle point were converted to cylindrical coordinates $[R^o, \varphi^o, p_d]^\intercal$ in 3D space, where $R^o \in \mathbb{R}_{>0}$ and $\varphi^o \in (-\pi, \pi]$ represent radial and angular coordinates of PCS (ProTac Coordinate System). Mathematically, the conversion can be expressed as:

$$R^o = \sqrt{o_x^2 + o_y^2}, \quad \varphi^o = \arctan 2(o_y, o_x). \tag{3}$$

From this, consider a given point $[R^s, \varphi^s, p_d]^\intercal$ on the skin surface having the same angular and axial coordinates with the obstacle point $\mathbf{o}$ ($\varphi^s = \varphi^o$). The normal distance vector $\hat{\boldsymbol{n}} \in \mathbb{R}^3$ between $\mathbf{o}$ and the skin surface can be estimated as:

$$\hat{\boldsymbol{n}} = (R^o - R^s)\frac{\mathbf{r}}{\|\mathbf{r}\|}. \tag{4}$$

where $\mathbf{r}$ is the directional vector of the obstacle point $\mathbf{o}$, perpendicular to the cylindrical axis of ProTac, defined as $\mathbf{r} := \mathbf{o}^\intercal - [0, 0, p_d]^\intercal$. Here, the radial coordinate R^s is constant for all control points on the skin surface. This is because the current ProTac design features a cylindrical skin shape with a radius R, ensuring $R^s = R$ for all (φ^s, p_d).

Finally, given $[\hat{\boldsymbol{n}}_j, \forall j \in \mathcal{O}]$ determined for every obstacle point $[\mathbf{o}_j, \forall j \in \mathcal{O}]$ (based on Eq. 4), the distance vector $\hat{\boldsymbol{n}}^c$ from an obstacle to the ProTac skin can be defined as the closest obstacle points. Thus, we have:

$$\hat{\boldsymbol{n}}^c := \arg\min_{\hat{\boldsymbol{n}}_j} \|\hat{\boldsymbol{n}}_j\| \ , \ \forall j \in \mathcal{O}. \tag{5}$$

From this, the distance estimation can be determined as the magnitude of the distance vector $\|\hat{\boldsymbol{n}}^c\|$.

C. Risk Score

We propose a *risk score* that can serve as an alternative to the estimated distance $\|\hat{\boldsymbol{n}}^c\|$. While the risk score does not directly measure the distance, it offers a more intuitive metric that increases as obstacles approach the ProTac link. Furthermore, the risk score is inherently more sensitive and consistently provides the same measurement range for different obstacles. This is in contrast to distance measurements, which require thorough calibration for each specific obstacle (as demonstrated in Sec. V-A). Based on the observation that an object's area increases as it gets closer to the ProTac link, we derive the risk score metric by combining an object's pixel area $A \in \mathbb{R}$ and corresponding estimated distance $\|\hat{\boldsymbol{n}}^c\|$.

Thus, while sharing the same direction with $\hat{\boldsymbol{n}}^c$, the risk-score magnitude can be calculated as:

$$r = \frac{A - \|\hat{\boldsymbol{n}}^c\| A_0^2}{A_0 \|\hat{\boldsymbol{n}}^c\| (\eta - A_0)}, \tag{6}$$

where A_0 represents the pixel area when the obstacle is initially detected. Equation (6) provides the raw risk score value $A/A_0\|\hat{\boldsymbol{n}}^c\|$, which is then normalized within the range $[A_0, \eta]$. We have set $\eta = 5$ for all the obstacles tested. The evaluation of the risk score, along with the distance estimate $\|\hat{\boldsymbol{n}}^c\|$, is presented in Section V-A.

D. Multi-camera Fusion

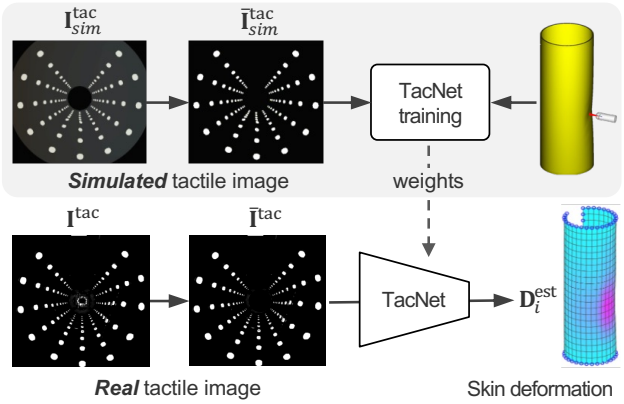
While ProTac proximity information can be extracted from a single camera, we note that the ProTac module is equipped with two opposing cameras positioned at its ends (see Fig. 2), designed to enhance its sensing coverage and performance. To fully exploit this unique setup, this section introduces a strategy for combining proximity sensing information, either the risk score r or the direct distance estimate $\|\hat{\boldsymbol{n}}^c\|$, obtained from the two opposing cameras via their respective monocular depth estimation networks. Given s_1 and s_2 denote the proximity information obtained from Camera-1 and Camera-2, respectively, the combined sensing signal s at every point in time can be computed as:

$$s = \max(s_1, s_2) \tag{7}$$

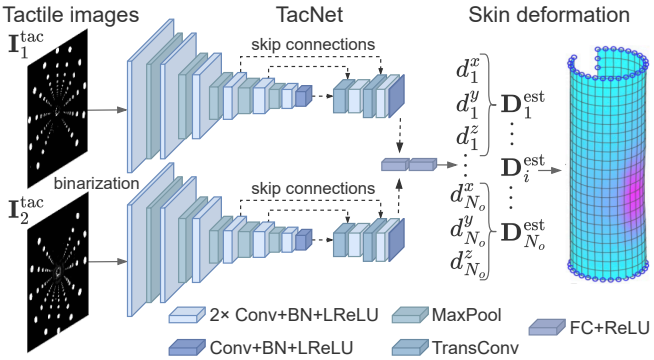
where the sensing signal s can represent either the risk score ($s := r$) or direct distance measurement ($s := \|\hat{\boldsymbol{n}}^c\|$). The performance of this fusion technique to enhance ProTac's proximity sensing is demonstrated in Section V-A.

IV. Tactile Perception

The tactile perception is activated when the ProTac's skin is at opaque mode. Awareness of contact intensity and location on robot arms is valuable for HRI, particularly in collision handling frameworks [41]. Additionally, identifying multi-point contacts opens new possibilities for haptic interfaces, reflecting the unique capabilities of large-area vision-based sensors. This paper introduces a zero-shot sim-to-real learning method for inferring multi-point contact depths and contact locations from skin deformation, estimated using marker-featured tactile images. The mapping between tactile images and skin deformation is estimated through a deep learning model (called TacNet) trained with simulation datasets (see Fig. IV). These datasets comprise pairs of simulated tactile images and skin deformation states, collected using a simulation framework proposed in our previous work [34]. However, in contrast to [34] that necessitates a subset of real images to address the sim-to-real gap, this paper employs the domain randomization technique to enable zero-shot learning of the TacNet model based solely on simulation images. Additionally, this paper proposes a feature-level fusion scheme for stereo tactile images, which is demonstrated to achieve more effective sim-to-real sensing performance compared to the input-level tactile image concatenation scheme [34]. Lastly, the multilayered structure of the ProTac skin (see Fig. 2) has been successfully modeled in this paper.



(a) Overview of the data-driven pipeline for skin deformation sensing



(b) Feature-level fusion scheme for stereo tactile images

Fig. 5. **Illustration of tactile processing pipeline**. (a) The learning pipeline for skin deformation estimation, including the Unity-SOFA simulation platform for data collection and a domain randomization technique applied to binary tactile image inputs to reduce the sim-to-real gap. (b) TacNet's architecture, where TacNet models are jointly trained to estimate skin deformation from stereo tactile image inputs.

A. Compound ProTac Skin Modeling

In an effort to diversify the tactile dataset used to train TacNet without incurring significant costs, this work implements a Finite Element (FE) algorithm via the *SOFA* simulation framework[4]. The primary challenge lies in accurately simulating the mechanical behaviors of the ProTac's skin to physical stimuli. Specifically, the mechanical coupling between the PDLC film and the outer silicon layer is crucial. This section will detail the implementation of this relationship within the SOFA framework.

1) Silicon-made layer: Replicating accurately the mechanical behaviors of the outer elastomer layer presents a significant challenge due to its non-linearity and hysteresis. While hyper-elastic material models offer a viable solution, they demand substantial computational resources and pose difficulties in selecting model coefficients. In this work, we address this challenge by treating the soft layer as an elastic body governed by linear constitutive relations based on Hooke's laws, characterized by Young's modulus E and Poisson's ratio ρ. However, this approach may yield unrealistic results, especially for soft bodies experiencing large deformations. To mitigate this issue, the SOFA framework provides a co-rotational FEM formulation [34]. At a given simulation time, the current

[4]https://www.sofa-framework.org/

geometrical state of the soft layer can be obtained by solving the following dynamic equation:

$$\mathbf{M_s}(\mathbf{q})\ddot{\mathbf{q}} = \mathbf{F}^{ext}(t) - \mathbf{F}^{int}(\mathbf{q}, \dot{\mathbf{q}}) + \mathbf{J}^T\boldsymbol{\lambda}, \tag{8}$$

where $\mathbf{q} \in \mathbb{R}^{N\times 3}$ is the 3D position of element nodes (corresponding to N degrees of freedom), $\mathbf{M_s}(\mathbf{q}) = \text{diag}[..., \frac{M_s}{N}, ...] \in \mathbb{R}^{N\times N}$ is the uniform mass matrix, in which, M_s is total mass of the soft skin; while $\mathbf{F}^{ext}(t)$ denotes the external forces (*e.g,* gravity) at each time step t and $\mathbf{F}^{int}(\mathbf{q}, \dot{\mathbf{q}})$ represents internal forces (calculated by FE model) upon the system state. The Jacobian matrix $\mathbf{J}(\mathbf{q}) = \frac{\partial \boldsymbol{\xi}}{\partial \mathbf{q}}$ gathers the normal and tangential constraint directions to project the Lagrange multipliers $\boldsymbol{\lambda}$ - equivalent to the magnitude of contact forces, to the mapped DoFs. In our experimental scenarios (indentation tests), tangential forces are negligible compared to normal components, resulting in physical contact that is entirely normal and frictionless, without lateral slip.

Equation 8 is then rewritten as below (refer to [42] for the detailed conversion):

$$\underbrace{\left(\mathbf{M} + dt^2\mathbf{K} + dt\mathbf{C}\right)}_{\mathbf{A}}\underbrace{\ddot{\mathbf{q}}}_{\mathbf{x}} = \underbrace{-dt^2\mathbf{K}\dot{\mathbf{q}}_1 + dt\left(\mathbf{F}^{ext} - \mathbf{F}^{int}\right)}_{\mathbf{b}} + dt\mathbf{J}^T\boldsymbol{\lambda} \tag{9}$$

where $\mathbf{F}_2^{ext}$ is the external force at the next time step, $\mathbf{K} = \frac{\partial \mathbf{F}^{int}}{\partial \mathbf{q}}$ and $\mathbf{C} = \frac{\partial \mathbf{F}^{int}}{\partial \dot{\mathbf{q}}}$ are stiffness and damping matrices, respectively. The solution $\mathbf{x}$ is used to update the system state at the next step.

2) PDLC film: The mechanical coupling between the PDLC film and the outer elastomer layer must be achieved while maintaining computational efficiency. Here, the PDLC film, inheriting characteristics from PET films, is simplified as a stiffening substrate that constrains the deformation of the outer soft layer from its original representation. This contribution is modeled by integrating *virtual elastic springs*, which connect all paired nodes of the soft layer's mechanical model one by one as illustrated in Fig. 2, where the soft layer is characterized by Young's modulus E and viscosity η, and the mechanical effects of the PDLC film are represented by virtual springs with stiffness k_{vs}. This model operates under the assumption that the PDLC film will not exceed its elastic limit point (*i.e.*, undergo plastic deformation). Otherwise, the representation of ProTac skin would become inaccurate, resulting in unrealistic sets of simulated tactile images. At an equilibrium deforming state t, the virtual springs generate internal forces $\mathbf{f}^{spring}$, which are proportional to the nodal displacement $\boldsymbol{\delta} \coloneqq \mathbf{q}(t) - \mathbf{q}(0)$, where $\mathbf{q}(t)$ and $\mathbf{q}(0)$ represent the current and rest positions, respectively. Consequently, the motion equation (9) for the entire lumped system of ProTac skin is updated as follows:

$$\left(\mathbf{M_\Sigma} + dt^2\bar{\mathbf{K}} + dt\mathbf{C}\right)\ddot{\mathbf{q}} = -dt^2\bar{\mathbf{K}}\dot{\mathbf{q}}_1 + dt\left(\mathbf{F}^{ext} - \bar{\mathbf{F}}^{int}\right) + dt\mathbf{J}^T\boldsymbol{\lambda} \tag{10}$$

where $\mathbf{M_\Sigma}(\mathbf{q}) = \text{diag}[..., \frac{M_s+M_f}{N}, ...] \in \mathbb{R}^{N\times N}$ with M_f is total mass of the PDLC film, $\bar{\mathbf{K}} = \frac{\partial \bar{\mathbf{F}}^{int}}{\partial \mathbf{q}}$ and $\bar{\mathbf{F}}^{int} = \mathbf{F}^{int} - \mathbf{f}^{spring}$, in which $\mathbf{f}^{spring} = k_{vs} \times \boldsymbol{\delta}$.

3) Simulated skin deformation: We rely on the above ProTac skin model to acquire a dataset of ground-true skin deformation $\{\mathbf{D}^{\text{gt}}\}_{\text{sim}}$, which enables tactile sensing from the estimation of the global skin deformation (refer to next Sections IV-B-IV-C). To reduce computational costs, let us define the simplified skin shape in the non-deformed state as $\mathbf{X}_0 := [\mathbf{X}_{0,i} \in \mathbb{R}^3 \mid \mathbf{X}_{0,i} = \mathbf{q}_i(0), \forall i \in \mathcal{N}]$, where $\mathcal{N}$ denotes indices of nodes on the skin surface ($|\mathcal{N}| = N_o$). Upon physical contact, the soft skin is deformed, and the original skin state $\mathbf{X}_0$ is displaced to a new deformed state $\mathbf{X} \in \mathbb{R}^{N_o\times 3}$. This displacement complies with the soft skin dynamics derived in the section above. From this, the skin deformation $\mathbf{D}^{\text{gt}} = [\mathbf{D}_i^{\text{gt}} \in \mathbb{R}^3, \forall i \in \mathcal{N}]$ is defined as the nodal displacement vectors

$$\mathbf{D}_i^{\text{gt}} \coloneqq \mathbf{X}_i - \mathbf{X}_{0,i}\ ,\ \forall i \in \mathcal{N}. \tag{11}$$

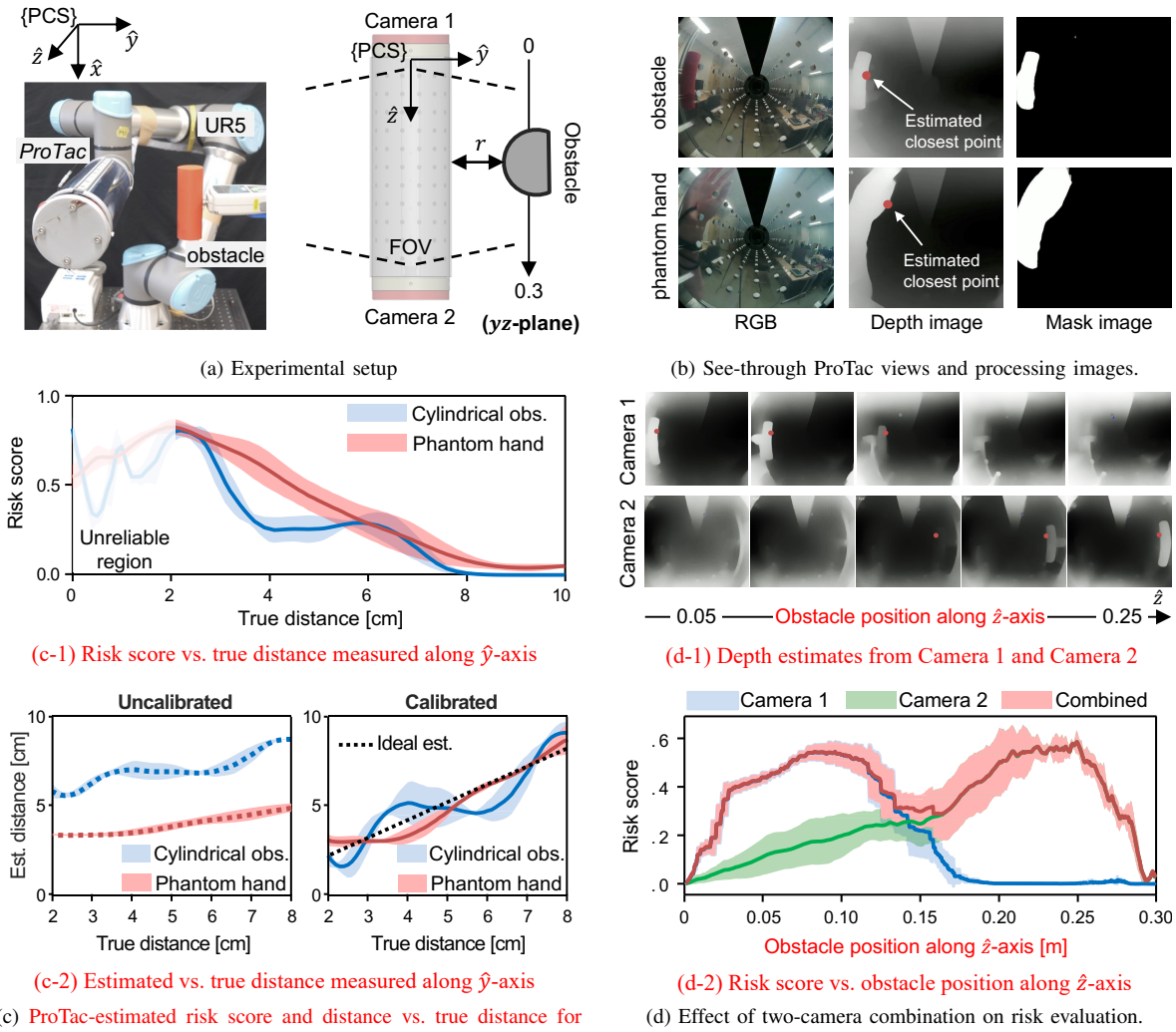
B. Zero-shot Learning of Deformation Sensing

Figure 5a illustrates the sim-to-real learning pipeline for deformation sensing. This process employs the DNN-based TacNet model ($\mathcal{T}$) to estimate ProTac skin deformation, represented as $\mathbf{D}^{\text{est}} \in \mathbb{R}^{N_o\times 3}$, from the opaque-state tactile image $\mathbf{I}^{\text{tac}} \in \mathbb{R}^{a\times b\times 3}$, where $a \times b$ denotes image resolution. To enable the fusion of two tactile images $\{\mathbf{I}_1^{\text{tac}}, \mathbf{I}_2^{\text{tac}}\}$ captured from ProTac, we leverage two independent TacNet models to extract the feature maps from the tactile images. Specifically, we compute the feature embeddings $\mathbf{e}_1 = \mathcal{T}_1(\bar{\mathbf{I}}_1^{\text{tac}})$ and $\mathbf{e}_2 = \mathcal{T}_2(\bar{\mathbf{I}}_2^{\text{tac}})$, respectively. The concatenation of these two feature vectors along the channel axis, $\mathbf{e} = [\mathbf{e}_1||\mathbf{e}_2]$, is then passed through two fully connected layers, resulting in the estimated output displacement vectors $\mathbf{D}^{\text{est}}$ (see Fig. 5b). Note that $\bar{\mathbf{I}}^{\text{tac}} \in \mathbb{Z}_2^{a\times b}$ represents the binary version of $\mathbf{I}^{\text{tac}}$. For the training dataset, skin deformation labels $\{\mathbf{D}^{\text{gt}}\}_{\text{sim}}$ are obtained from the SOFA environment, while the corresponding set of simulated input images $\{\bar{\mathbf{I}}^{\text{tac}}\}_{\text{sim}}$ is acquired from Unity.

Training: The training procedure, including the architecture of the TacNet models and the loss function, follows the settings adopted in [34]. It is important to note that the dimension of the network's output layer must be adapted to align with the size of ProTac's FEM mesh model, represented by N_o nodal displacement vectors. In this paper, the mesh resolution is empirically chosen as $N_o = 521$ to balance spatial sensing resolution with computational efficiency. To address the sim-to-real gap, we employ the domain randomization technique applied to the binary version of tactile images $\{\bar{\mathbf{I}}^{\text{tac}}\}_{\text{sim}}$. The domain randomization involves performing affine transformations during the training process to diversify the perspective of tactile binary images, including translation, rotation, and scaling. This technique, along with the high-fidelity physical modeling of the soft ProTac skin, facilitates zero-shot sim-to-real transfer, eliminating the need for real data or an additional network to mitigate the sim-to-real gap. Section V-B highlights the effectiveness of our learning approach in facilitating sim-to-real contact sensing.

C. Multi-point Contact Depth Estimation and Localization

Based on the estimated skin deformation $\mathbf{D}^{\text{est}}$ provided by the TacNet model, this section outlines the process for identi-



(a) Experimental setup

(b) See-through ProTac views and processing images.

(c-1) Risk score vs. true distance measured along $\hat{y}$-axis

(d-1) Depth estimates from Camera 1 and Camera 2

(c-2) Estimated vs. true distance measured along $\hat{y}$-axis

(d-2) Risk score vs. obstacle position along $\hat{z}$-axis

(c) ProTac-estimated risk score and distance vs. true distance for two obstacles using single Camera 1.

(d) Effect of two-camera combination on risk evaluation.

Fig. 6. **Proximity mode evaluation**. (a) Experimental setups. (b) Sample images from the transparent ProTac skin along with their processed outputs. (c) Performance of risk score estimation and estimated distance compared to the true ProTac-obstacle distance for two different obstacles. (d) Demonstration of the effectiveness of two-camera fusion for enhanced proximity sensing.

fying *multi-point* contact depths $\{\hat{\boldsymbol{d}}_l^c\}$ and their corresponding contact locations $\{\hat{\boldsymbol{x}}_l^c\}$. To achieve this, we adopt the graph-based contact region labeling (CRL) method proposed in [34] for skin-based contact sensing. Thus, with L contact regions $\{\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_L\}$ identified through the CRL procedure, represented by distinct groups of node indexes i, multi-point contact locations $\{\hat{\boldsymbol{x}}_1^c, \hat{\boldsymbol{x}}_2^c, \cdots, \hat{\boldsymbol{x}}_L^c\}$ can be defined for all contact regions as follows:

$$\hat{\boldsymbol{x}}_l^c := \mathbf{X}_{0,i_l^*}, \forall l \in \{1, \ldots, L\}, \tag{12}$$

where i_l^* represents the index of the mesh node where the contact is exerted. We assume the contact location to be the position at which the node experiences the most significant deformation within a contact region $\boldsymbol{R}_l$; thus

$$i_l^* = \arg\max_{i \in \boldsymbol{R}_l} \|\mathbf{D}_i^{\text{est}}\|, \forall l \in \{1, \ldots, L\}. \tag{13}$$

From this, the multi-point contact depth vectors $\{\hat{\boldsymbol{d}}_1^c, \hat{\boldsymbol{d}}_2^c, \cdots, \hat{\boldsymbol{d}}_L^c\}$ can be extracted as:

$$\hat{\boldsymbol{d}}_l^c = \mathbf{D}_{i_l^*}^{\text{est}}, \forall l \in \{1, \ldots, L\}. \tag{14}$$

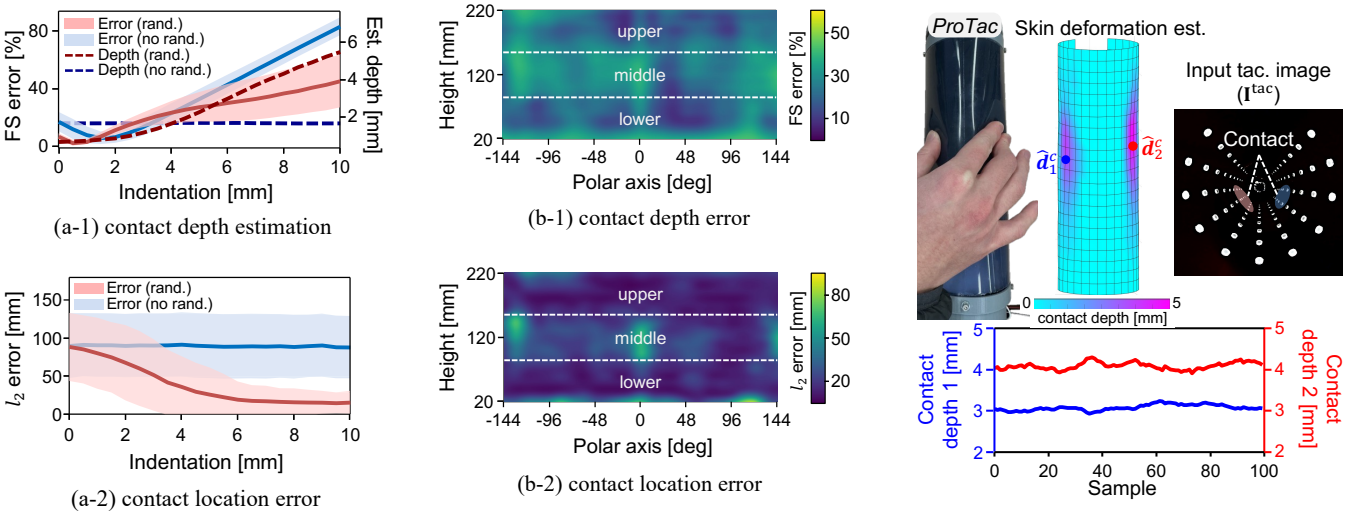
Thus, the magnitude of the contact depth vector is referred to as the contact depth, that is $\|\hat{\boldsymbol{d}}_l^c\|$. For brevity, we utilize $\|\hat{\boldsymbol{d}}^c\|$ and $\hat{\boldsymbol{x}}^c$ to denote a single-point contact depth and location without the subscript index throughout the paper.

V. Sensing Performance

In this section, we evaluate the sensing performance of ProTac in both the *proximity* mode (Sec. V-A) and the *tactile* mode (Sec. V-B), respectively. ProTac was implemented with fish-eye cameras (ELP, 180° lens, 30 Hz), and a PC (64GB RAM, NVIDIA RTX 3090 GPU). The control for switching the skin transparency is regulated by the PC, which connects to the power control unit (LP1, TOPAN Inc., Japan) of the PDLC film through an RS232 serial port.

A. Proximity Mode

This section discusses the performance of the proposed method in estimating the distance between the closest external obstacles and the ProTac skin. Furthermore, we evaluate the



(a-1) contact depth estimation

(a-2) contact location error

(a) Tac. sensing accuracy w.r.t indentation

(b-1) contact depth error

(b-2) contact location error

(b) Tac. sensing accuracy on the whole skin

(c) Demonstration of two-point contact detection

Fig. 7. **Tactile mode evaluation**. (a) Errors in contact depth estimation and contact localization by contact intensity. (b) Errors in contact depth estimation and contact localization across the entire skin surface (note that contact localization error is averaged for indentations exceeding 3 mm, excluding outliers from minor contact intensities). (c) Demonstration of the ProTac link's ability to identify multi-point contacts.

risk-score metric, demonstrating its advantage over direct distance measurement.

1) Setups: The experimental setup is illustrated in Fig. 6a. The ProTac link is attached to the end-effector of a UR5 robot arm, controlled linearly toward a fixed obstacle along the $\hat{y}$-axis of the ProTac Coordinate System (PCS). During this movement, measurements of the distance $\|\hat{\boldsymbol{n}}^c\|$ and equivalent risk score r were recorded. The true distance from the obstacle to the ProTac skin was deduced from UR5's predefined movements using position feedback. To assess the repeatability of ProTac, the measurement process was repeated several times with different UR5 moving speeds within the range from 0.01 m/s to 0.02 m/s. The evaluation was performed with two obstacles of different shapes, including a cylinder-shaped obstacle and a phantom arm (see Fig. 6b).

Moreover, we conducted an additional experiment to examine how ProTac with the setup of two opposing cameras can enhance the sensing performance while varying the obstacle's position along the $\hat{z}$-axis of the PCS (see schematic in Fig. 6a). This performance was compared to that measured by single cameras. During this experiment, the UR5 robot moved the ProTac link along the $\hat{z}$-axis relative to a fixed cylindrical obstacle, while maintaining a constant distance between the obstacle and the ProTac skin. This motion resulted in an apparent obstacle displacement from 0 to 0.3 m along the $\hat{z}$-axis from the perspective of the ProTac skin. Risk score values were recorded throughout the motion, and the measurement was repeated several times at varying UR5 speeds between 0.05 m/s and 0.1 m/s along the $\hat{z}$-axis.

2) Results: Figure 6b illustrates the extraction of depth and mask images for nearby obstacles using the ProTac view in the transparent state. The estimated risk score r and distance $\|\hat{\boldsymbol{n}}^c\|$, relative to the ground-truth distance, are shown in Figures 6c-1 and 6c-2, respectively. The results emphasize the ProTac's reliable measurement range from 2 cm to 8 cm. Notably, within this range, the risk-score estimation r exhibited a linear trend, demonstrated a consistent measurement scale, and greater sensitivity, compared with the raw distance measurement $\|\hat{\boldsymbol{n}}^c\|$ (Fig. 6c-2, uncalibrated) for the two different obstacles. However, the ProTac's direct distance measurement $\|\hat{\boldsymbol{n}}^c\|$ remains usable through the calibration for every specific obstacle, as reported in Figure 6c-2.

Additionally, we demonstrate how leveraging the complementary views of ProTac's binocular vision system enhances proximity sensing performance. Ideally, as the obstacle is positioned along the $\hat{z}$-axis of the PCS, the risk score r should maintain a consistent value within the field of view (FOV) of the ProTac link. However, when only one camera is used, the risk score measurements gradually degrade as the obstacle approaches the opposite end of the link, away from the observing camera, whether it is Camera-1 or Camera-2 (see Fig. 6d-2). By fusing measurements from both cameras, the system can compensate for these blind spots, resulting in more robust sensing across the observable range of the link (Fig. 6d-2), which indicates the advantages of ProTac's design in utilizing stereo camera views to improve sensing performance. Nevertheless, even with two-camera fusion, a noticeable drop in risk score, up to approximately 40%, is observed when the obstacle is positioned near the center of the link (*i.e.*, the region farthest from both cameras). This degradation may result from the compounded effects of optical distortion from the skin material and variable lighting conditions, which reduce depth estimation quality, especially in central areas that are optically disadvantaged from both camera viewpoints, as seen in Fig. 6d-1. These results highlight both the advantages and limitations of ProTac's binocular vision system in mitigating directional blind spots influenced by the physical properties of the ProTac skin.

TABLE II
FULL-SCALE CONTACT DEPTH ERROR AND INFERENCE TIME FOR DIFFERENT INPUT SETTINGS AND FUSION SCHEMES

Setting	Contact Depth FS Error by Skin Regions [%]			Inference Time [ms]*
	Lower	Middle	Upper	
Single Camera-1	38.9±24.8	27.0±17.5	21.2±14.2	3.6 ± 0.2
Single Camera-2	25.2±17.4	23.6±14.8	32.2±21.7	3.6 ± 0.2
Input-level Fusion	25.0±16.6	25.9±16.4	35.5±22.6	9.7 ± 2.7
Feature-level Fusion	23.8±17.1	28.5±18.9	22.4±15.2	10.2 ± 3.0

Gray-shaded cells indicate inferior performance (error $> 30\%$).
*Inference times are averaged over 500 runs.

B. Tactile Mode

This section characterizes the estimation accuracy of contact depth $\|\hat{\boldsymbol{d}}^c\|$ and contact location $\hat{\boldsymbol{x}}^c$ across different regions of the large-area skin. In addition, we demonstrate the effectiveness of the feature-level fusion scheme for stereo tactile images by comparing it to single tactile image inputs and the input-level concatenation scheme used in [34], [43].

1) Setups: For the tactile sensing evaluation across the entire large-area skin, we collected a set of *unseen* real images by pressing a finger-shaped object onto the ProTac skin at various locations corresponding to the nodes of a cylindrical grid defined on the skin surface. The grid spans polar axis coordinates from $-144°$ to $144°$ in $12°$ increments and height axis coordinates from 20 mm to 220 mm in 10 mm increments. At each location, data were recorded as the contact indentation was increased from 0 to 10 mm in 0.5 mm increments.

2) Results: Figure 7a illustrates the estimation errors in contact depth and contact location with increasing contact indentation. The results indicate that ProTac delivers sensitive sensing signals when the indentation exceeds 3 mm across the full 10 mm range. At this point, the estimated contact depth exhibits an approximately linear relationship with indentation (Fig.7a-1), while the contact location error decreases as contact indentation increases (Fig.7a-2). We note that the contact location error (Fig. 7a-2) is averaged only for contact indentations exceeding 3 mm, excluding outliers caused by insensitive depth estimation at subtle contact intensities. Moreover, Figure 7a shows that data augmentation, by randomizing the perspectives of input images, is effective for sim-to-real transfer, enabling zero-shot learning without the need for additional real-world adaptation. In addition, Figure 7b presents the estimation errors of contact depth and location across the entire skin. Variations in sensing accuracy across contact regions can be attributed to intricate soft skin morphologies that are not fully captured by the proposed skin model. However, the accuracy remains acceptable for general robotics applications [34] and can serve as a performance baseline for further development of vision-based soft, whole-arm sensing skins. Figure 7c demonstrates a two-point contact detection scenario, where the estimation of skin deformation $\mathbf{D}^{\text{est}}$ allows for the measurement of contact depths $\{\hat{\boldsymbol{d}}_1^c, \hat{\boldsymbol{d}}_2^c\}$ at two separate contact locations (Eq. 14).

Furthermore, we evaluate the effectiveness of the feature-level fusion scheme for stereo tactile images by comparing it with single tactile image inputs and the input-level concatenation scheme [34], [43]. To this end, we examine the average full-scale (FS) error in contact depth estimation across three regions along the height ($\hat{z}$-axis) of the ProTac skin (see Fig. 7b): the lower region ($[20, 80]$ mm, near Camera-2), the middle region ($[90, 150]$ mm), and the upper region ($[160, 220]$ mm, near Camera-1). The results, as shown in Table II, indicate that the single image input settings lead to inferior sensing performance at the far end of the respective cameras. Specifically, compared to the other regions, using a single input from Camera-1 results in reduced performance in the lower region, while a single input from Camera-2 leads to reduced performance in the upper region. Moreover, the input-level concatenation scheme demonstrates inferior sim-to-real performance, with substantially high estimation errors in the upper region. In contrast, the proposed feature-level fusion scheme (Fig. 5b) achieves consistent sensing accuracy across the three regions, with full-scale (FS) errors averaging around 25%. Despite the slightly elevated error, this level is comparable to prior works on large-area vision-based soft sensing skins [34], [44], and remains sufficient for dynamic interactive tasks, as demonstrated in Section VII. In terms of inference speed, although the feature-level fusion scheme exhibits a higher latency compared to single-camera inputs (Table II), its average inference time of approximately 10 ms (or 100 Hz) remains well within the requirements for real-time applications. These results confirm that stereo vision, combined with an appropriate fusion strategy, can provide consistent tactile sensing accuracy across large-area skins without compromising real-time performance.

VI. PROTAC-DRIVEN SAFETY CONTROL

While previous work [34] has explored larege-area tactile sensing for interaction-based tasks, such as non-prehensile manipulation, this section focuses on the application of ProTac, particularly in proximity mode, for safety-critical scenarios. We begin by providing a brief overview of an admittance control framework that can be combined with distance estimation in proximity for obstacle avoidance. Subsequently, we introduce a simple strategy to adjust the robot's speed according to the distance estimated by the ProTac link, which relies on the adaptive time-scaling of a trajectory.

A. ProTac-driven Admittance-based Reactive Control

This subsection showcases a ProTac-driven robot arm system with a reflex behavior. This behavior enables the robot to respond to close-contact obstacles, which can be utilized for safety applications, such as spontaneous collision avoidance or collision reaction. To this end, we employ an admittance controller [45] that treats the robot as a mass-spring-damper system and can be formulated as follows

$$\boldsymbol{M}_v \ddot{\boldsymbol{x}}_d + \boldsymbol{D}_v \dot{\boldsymbol{x}} + \boldsymbol{K}_v \boldsymbol{x} = \boldsymbol{f}_{\text{ext}}, \tag{15}$$

where $\boldsymbol{M}_v \in \mathbb{R}^{3\times3}$ is the virtual positive-definite inertia matrix, $\boldsymbol{D}_v \in \mathbb{R}^{3\times3}$ is the virtual positive-definite diagonal

damping matrix, and $\boldsymbol{K}_v \in \mathbb{R}^{3\times3}$ is the virtual positive-definite diagonal stiffness matrix. Here, $\boldsymbol{x} = [x_x, x_y, x_z]^\intercal \in \mathbb{R}^3$ and $\dot{\boldsymbol{x}} = [\dot{x}_x, \dot{x}_y, \dot{x}_z]^\intercal \in \mathbb{R}^3$ are the position and velocity states of the robot end-effector, while $\ddot{\boldsymbol{x}}_d$ is the desired end-effector acceleration. Thus, the Cartesian-space admittance control law can be derived as

$$\ddot{\boldsymbol{x}}_d = \boldsymbol{M}_v^{-1}(\boldsymbol{f}_{\text{ext}} - \boldsymbol{D}_v\dot{\boldsymbol{x}} - \boldsymbol{K}_v\boldsymbol{x}). \tag{16}$$

Here, $\boldsymbol{f}_{\text{ext}} := \boldsymbol{f}_v$ can be considered as the virtual repulsive force $\boldsymbol{f}_v$, which is associated with the ProTac-estimated distance vector from the obstacle $\hat{\boldsymbol{n}}$. The mapping function can be defined as:

$$\boldsymbol{f}_v = f_v \frac{\hat{\boldsymbol{n}}^c}{\|\hat{\boldsymbol{n}}^c\|}, \tag{17}$$

which has the same direction as $\hat{\boldsymbol{n}}$ but with magnitude:

$$f_v = \frac{f_v^{\max}}{1 + e^{(\|\hat{\boldsymbol{n}}^c\|(2/\rho)-1)\gamma}}. \tag{18}$$

This mapping function is chosen to enhance the smoothness of the robot's reactive action [36], where $f_v^{\max}$ is the maximum magnitude of the resulting virtual force, and γ is a shape factor. As an obstacle approaches the ProTac skin, the magnitude of the repulsive vector f_v increases, reaching $f_v^{\max}$ when the estimated distance $\|\hat{\boldsymbol{n}}^c\|$ closes to 0. The virtual force f_v gradually diminishes as the distance to the obstacle is near to or extends beyond the value ρ (*i.e.*, $\|\hat{\boldsymbol{n}}^c\| \geq \rho$).

Given the resulting virtual force $\boldsymbol{f_v}$ and the control law (16), the desired joint accelerations $\ddot{\boldsymbol{\theta}}_d \in \mathbb{R}^n$ (n is the number of robot joints) can be solved as [46]

$$\ddot{\boldsymbol{\theta}}_d = \boldsymbol{J}_e^\dagger(\ddot{\boldsymbol{x}}_d - \dot{\boldsymbol{J}}_e\dot{\boldsymbol{\theta}}) \tag{19}$$

where $\boldsymbol{J}_e^\dagger \in \mathbb{R}^{n\times3}$ is the Moore-Penrose pseudoinverse of the end-effector Jacobian defined as $\dot{\boldsymbol{x}}_d = \boldsymbol{J}_e\dot{\boldsymbol{\theta}}_d$. From this, the commanded joint velocities can be computed as $\dot{\boldsymbol{\theta}}_d = \int \ddot{\boldsymbol{\theta}}_d$.

In this paper, our primary examination focuses on the reflex behavior within 3D linear space, without considerations for rotational components to ease implementation processes. Also, it is worth noting that while this paper specifically evaluates the framework for proximity-based sensing signals, the control law (16) can also be incorporated with the contact depth vector $\hat{\boldsymbol{d}}^c$ by defining $\boldsymbol{f}_{\text{ext}} := \hat{\boldsymbol{d}}^c$ to generate robot-safe behavior in response to contacts, as introduced in [47].

B. ProTac-driven Speed Regulation

This subsection describes a time-scaling strategy to dynamically adjust the robot speed in real-time based on the magnitude of the ProTac-estimated distance $\|\hat{\boldsymbol{n}}^c\|$, enabling fast and efficient adaptation without the need to re-plan a pre-defined trajectory from scratch.

For a given desired trajectory in Cartesian space $\boldsymbol{x}_d(s)$ parameterized by the time-scaling function $s(\tau) : [0,1] \mapsto [0,1]$ with $\tau = t/T$; where $t \in [0,T]$ is the time instance and T is the base motion time of the given trajectory, we can compute the robot velocity profile $\dot{\boldsymbol{x}}_d = [\dot{x}_x^d, \dot{x}_y^d, \dot{x}_z^d]^\intercal \in \mathbb{R}^3$ that linearly scales with the base motion time T

$$\dot{\boldsymbol{x}}_d = \frac{d\boldsymbol{x}_d}{ds} \cdot \frac{ds}{d\tau} \cdot \frac{1}{kT}. \tag{20}$$

Here, by introducing a time-scaling factor $k \geq 1 \in \mathbb{R}$, the robot velocity $\dot{\boldsymbol{x}}_d$ can be adjusted (*i.e.*, faster or slower) by scaling the motion time kT by the factor k. To dynamically adjust the robot speed according to the estimated distance $\|\hat{\boldsymbol{n}}^c\|$, we, therefore, compute the factor k from $\|\hat{\boldsymbol{n}}^c\|$ as

$$k = \frac{k^{\max} - 1}{1 + e^{(\|\hat{\boldsymbol{n}}^c\|(2/\rho)-1)\gamma'}} + 1. \tag{21}$$

The mapping function is similar to (17) to encourage motion smoothness in the face of sensing noises but with different shape factor γ' and interval $[1, k^{\max}]$. This means that the robot moves on the given trajectory with the base speed over duration T if the estimated distance to an obstacle $\|\hat{\boldsymbol{n}}^c\|$ goes beyond ρ (*i.e.*, $\|\hat{\boldsymbol{n}}^c\| \geq \rho$) at which $k = 1$. However, as the obstacle approaches, the speed gradually reduces with the increasing scale factor k which reaches $k^{\max}$ when $\|\hat{\boldsymbol{n}}^c\|$ closes to 0. Finally, given the desired Cartesian-space velocity $\dot{\boldsymbol{x}}_d$ computed with the resulted scale factor k in Eq. (20), the commanded joint velocities can be computed through the pseudoinverse of the robot Jacobian as $\dot{\boldsymbol{\theta}}_d = \boldsymbol{J}_e^\dagger\dot{\boldsymbol{x}}_d$.

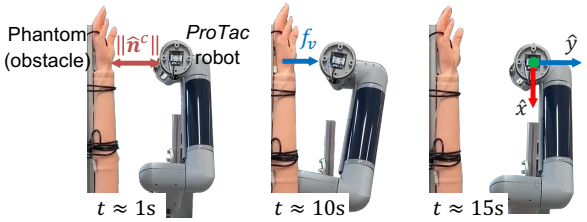
We note that while the estimated distance $\|\hat{\boldsymbol{n}}^c\|$ is utilized for derivations in this section to maintain intuitive formulations and demonstrate its usefulness, it is reasonable to employ the risk score metric r in those formulas. This can be achieved by simply replacing $\|\hat{\boldsymbol{n}}^c\|$ with $1/r$ and adjusting pre-defined parameters (such as γ, γ', ρ), which ensures the preservation of the overall functionality of proposed controllers.

TABLE III
CONTROL PARAMETERS FOR PROTAC-DRIVEN SAFETY CONTROLLERS

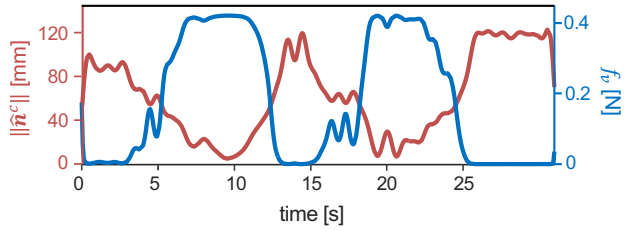
Parameter		Value	Unit
Virtual inertia matrix	$\boldsymbol{M}_v$	$\text{diag}(1.0, 1.0, 1.0)$	-
Virtual damping matrix	$\boldsymbol{D}_v$	$\text{diag}(1.5, 1.5, 1.5)$	-
Virtual stiffness matrix	$\boldsymbol{K}_v$	$\text{diag}(2.0, 2.0, 2.0)$	-
Max. virtual force	$f_v^{\max}$	0.45	N
Shape factor	γ	6.0	-
Proximal threshold	ρ	0.065	m
Max. time-scaling value	$k^{\max}$	2.0	-
Time-scaling shape factor	γ'	20.0	-

C. Experiment

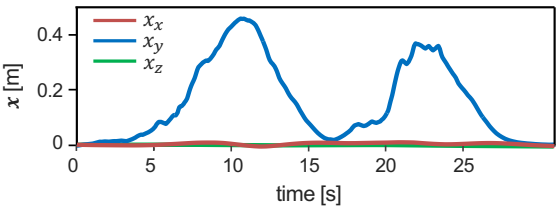
Setups: To validate the usefulness of the ProTac link in triggering safe responses of robot arms with environmental awareness, we utilized a custom-built 4-DOF (degree-of-freedom) robot arm, comprising two ProTac links employed as the upper arm and forearm (referred to as ProTac-integrated robot, Fig. 8a). Note that, in this paper, only the forearm was active for the demonstrations of the ProTac-driven control strategies. The ProTac links were interconnected through revolute joints actuated by electric motors (Dynamixel-P series, Robotis). In this system, coordination among control strategies, ProTac sensing interface, and motor control was facilitated through ROS (robot operating system). The commanded joint velocities $\dot{\boldsymbol{\theta}}_d$, computed from the control laws, were regulated by the built-in motion controller of the embedded motors. The values of the used controllers' parameters are reported in Table III.



(a) Experimental setup and rollout of *ProTac*-integrated robot's motion in response to the approaching phantom arm.

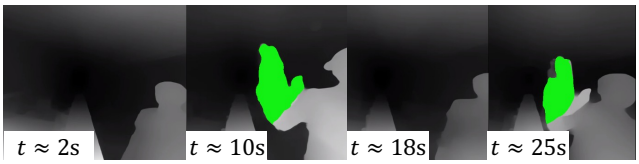


(b) Estimated distance (red line) and the resultant magnitude of virtual repulsive force (blue line).

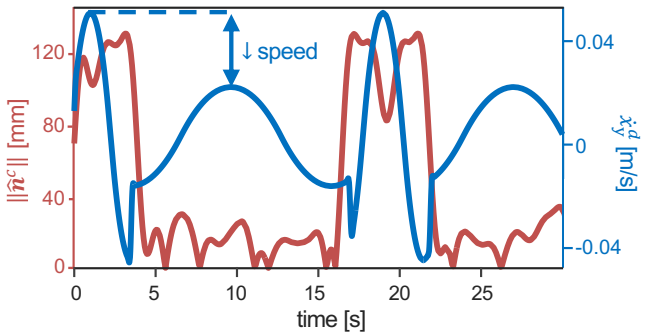


(c) Robot displacement in response to the virtual force.

Fig. 8. **Demonstration of ProTac-driven reactive control**. (a) Experimental setup and reactive behaviors of a custom-built robot arm integrated with ProTac links. (b) Logs of distance feedback and the corresponding virtual repulsive force. (c) Robot position records showing displacement along the $\hat{y}$-axis to avoid an approaching obstacle, driven by the virtual repulsive force derived from the ProTac-based distance estimation.



(a) Rollout of ProTac-based depth images. The green-shaded area indicates a group of obstacle points with a distance below ρ. (Here, $\rho = 65$ mm)



(b) Logs of ProTac-estimated distance (red line) and resultant Cartesian robot velocity along the $\hat{y}$-axis (blue line).

Fig. 9. **Demonstration of ProTac-driven speed regulation**. (a) Sequence of estimated depth images with nearby object mask labeling, used to trigger speed control. (b) Results demonstrating reduced robot speed triggered by the estimated distance between the ProTac link and an approaching human hand.

1) Admittance-based obstacle avoidance: This experiment demonstrates the effectiveness of the ProTac link in enabling the robot's reflex behavior for obstacle avoidance (see Algorithm VI-A). Here, a phantom arm was displaced back and forth relative to the forearm ProTac link, triggering the motion of the ProTac-integrated robot in response to the estimated distance between the link and the phantom arm $\|\hat{\boldsymbol{n}}^c\|$ (see Fig. 8a). Specifically, the variations in distance estimation $\|\hat{\boldsymbol{n}}^c\|$ resulted in changes in the virtual force f_v (as depicted in Fig. 8b) according to the mapping given in Eq. (18), which in turn influenced the robot's position along the $\hat{y}$-axis (as shown in Fig. 8c) based on the admittance control law in Eq. (16). For instance, as the phantom arm progressively approached the robot from $t \approx 4$ s to $t \approx 10$ s, the magnitude of the virtual force f_v gradually increased to around $f_v^{\max} = 0.45$ N. Consequently, the robot moved from the rest position $x_y = 0$ to a position approximately 0.4 m that is away from the obstacle, as seen in Figs. 8a and 8b-2. Conversely, at $t \approx 15$ s, when the phantom arm retreated from the robot, the virtual force diminished ($f_v \approx 0$), allowing the robot to return to its rest position.

2) Robot speed regulation: This experiment showcases the performance of robot speed regulation based on ProTac distance estimation $\|\hat{\boldsymbol{n}}^c\|$ (see Section VI-B). Here, a human approached the ProTac-integrated robot, which was periodically moving back and forth along a predefined trajectory linearly along the $\hat{y}$-axis. The depth images estimated by ProTac during the experiment are depicted in Fig. 9a, showing the human with their approaching hand. In these images, the green-shaded area indicates obstacle points with a distance falling below ρ (here, $\rho = 65$ mm). As presented in Fig. 9b, at $t \approx 10$ s and $t \approx 25$ s, the planned robot velocity $\dot{x}_y^d$ along the $\hat{y}$-axis was scaled approximately $k^{\max}$ times (here, $k^{\max} = 2$), peaking at around 0.025 m/s, while the human hand approached or the estimated distance $\|\hat{\boldsymbol{n}}^c\|$ became close to the forearm ProTac link. Conversely, when the human was not closer than the pre-defined distance ρ, the robot reverted to its initial speed (peaking at around 0.05 m/s), as depicted in Fig. 9b.

Last but not least, the system not only demonstrates the utility of the ProTac link in safety controls but also showcases the feasibility of a next-generation robot arm built with the *soft* proximity-tactile sensing. This integration has the potential to enhance safety in human-robot interaction scenarios.

VII. ProTac-Driven Control

This section demonstrates ProTac's bimodal capabilities through two robotic tasks: adaptive motion control in cluttered environments (Section VII-B) and multi-phase human-robot interaction (Section VII-C). To support these tasks, we introduce a unique sensing mode, *flickering* sensing, along with active sensing strategies (Section VII-A). In both tasks, the risk score r is used as the proximity feedback signal, as it has demonstrated better generalization to unseen obstacles.

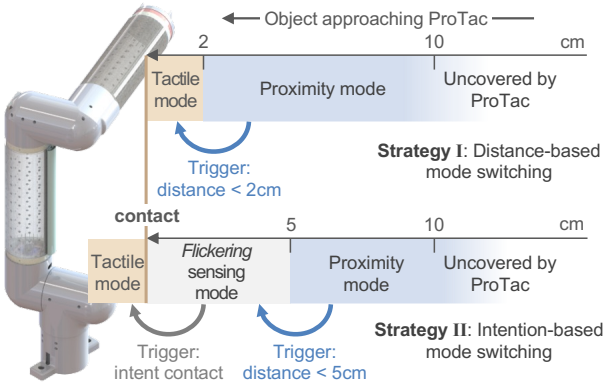


Fig. 10. **Illustration of strategies for ProTac mode switching**. **Strategy I**: When the distance is below 2 cm, ProTac switches from proximity to tactile mode for contact anticipation. **Strategy II**: ProTac switches from proximity to *flickering* mode as the distance is below 5 cm. During *flickering* sensing mode, ProTac switches to tactile mode if intentional contact is detected; otherwise, it returns to proximity mode.

A. ProTac flickering sensing mode and sensing strategies for multimodal tasks

To facilitate the utilization of ProTac for multimodal tasks, this subsection introduces an additional unique ProTac sensing mode named *flickering* sensing, along with two different sensing strategies for ProTac mode switching.

1) Flickering sensing mode: The "flickering" sensing mode refers to a ProTac operational mode where proximity and tactile sensing can be enabled nearly simultaneously. This mode is achieved by constantly switching between the proximity and tactile modes at a high frequency or at a certain switching period T_s. During the *flickering* mode, the last sample value estimated in one mode is retained for one switching period T_s when switching to the other mode, following the zero-order hold model.

2) Sensing strategies: Figure 10 illustrates the two strategies for switching among ProTac sensing modes, which are outlined as follows:

- Distance-based mode switching (Strategy I): ProTac begins in *proximity* mode. When obstacles are detected at a close distance or when high-risk level is observed, ProTac switches to *tactile* mode to anticipate contacts or collisions in a pre-contact phase. This mode switch occurs when the proximity sensing becomes unreliable, that is, below 2 cm.
- Intention-based mode switching (Strategy II): The ProTac initially operates in *proximity* mode. When a proximal distance is detected, typically below 5 cm, ProTac switches to *flickering* mode. If intentional contact is detected, ProTac switches to *tactile* mode; otherwise, it returns to *proximity* mode when minimal risk is observed.

B. Motion control with contact and obstacle awareness

Robot arms traversing in cluttered environments often burden external perception and navigation systems to avoid collisions and commonly necessitate kinematic redundancy to achieve target locations. However, collisions are inevitable in some specific environments (*e.g.*, a dense thicket of trees); thus mitigating damages to surroundings while achieving the goals is essential in these scenarios. Previous works primarily focused on the use of high-stiffness skin-based tactile sensing, without being aware of surroundings before contact, to achieve the task objective [48]. In contrast, this work, in addition to *soft* skin-based tactile sensing, leverages proximity perception to be better aware of obstacles in a pre-contact phase, which enhances the contact-constrained motion in mitigating impact forces (see results in Section VII-D1).

Specifically, we propose a strategy to search for robot motion to reach a target location in Cartesian space $\boldsymbol{x}_{\mathrm{G}} \in \mathbb{R}^3$ which is close to or obstructed by obstacles, while mitigating physical impacts on such obstacles. The latter condition can be achieved by restricting the estimated contact depth $\|\hat{\boldsymbol{d}}^c\|$ below a certain threshold $d_{\max} \in \mathbb{R}_{>0}$. Overall, this task is formulated as a constrained Quadratic Programming (QP) problem that optimizes commanded joint velocities $\dot{\boldsymbol{\theta}}_d \in \mathbb{R}^n$ to minimize an objective function $\mathcal{J}(\dot{\boldsymbol{\theta}}_d)$. This function $\mathcal{J}(\dot{\boldsymbol{\theta}}_d)$ is defined in a quadratic form such that it captures the target location error, that is

$$\mathcal{J}(\dot{\boldsymbol{\theta}}_d) := \frac{1}{2}\left(\frac{\boldsymbol{K}_p}{k}\Delta\boldsymbol{x} - \boldsymbol{J}_e\dot{\boldsymbol{\theta}}_d\right)^{\intercal}\left(\frac{\boldsymbol{K}_p}{k}\Delta\boldsymbol{x} - \boldsymbol{J}_e\dot{\boldsymbol{\theta}}_d\right), \tag{22}$$

where $\Delta\boldsymbol{x} \in \mathbb{R}^3$ is the directional vector toward the target location defined as $\Delta\boldsymbol{x} := \boldsymbol{x}_{\mathrm{G}} - \boldsymbol{x}$, $\boldsymbol{J}_e \in \mathbb{R}^{3\times n}$ is the Jacobian matrix, and $\boldsymbol{K}_p \in \mathbb{R}^{3\times 3}$ is positive-definite diagonal proportional matrix. Importantly, we directly impose the proximity effect, accounting for unknown obstacles prior to the contact, into the optimization problem by scaling the proportional matrix $\boldsymbol{K}_p$ with the time-scaling factor k (refer to Eq. 21). This means that, at any given moment, the velocity $\dot{\boldsymbol{\theta}}_d$ optimized from (22) is smaller when the robot is close to an obstacle (where $k > 1$), rather than when the obstacle is far away (*i.e.*, $k = 1$).

Additionally, as the robot comes into contact with the obstacle, we impose a motion constraint $\mathcal{C}(\dot{\boldsymbol{\theta}}_d)$ on the contact measured by $\hat{\boldsymbol{d}}^c$ as follows

$$\mathcal{C}(\dot{\boldsymbol{\theta}}_d) := (\hat{\boldsymbol{d}}^c + \beta\hat{\boldsymbol{e}}_c^{\intercal}\boldsymbol{J}_c\dot{\boldsymbol{\theta}}_d\hat{\boldsymbol{e}}_c)^{\intercal}(\hat{\boldsymbol{d}}^c + \beta\hat{\boldsymbol{e}}_c^{\intercal}\boldsymbol{J}_c\dot{\boldsymbol{\theta}}_d\hat{\boldsymbol{e}}_c) \leq d_{\max}^2, \tag{23}$$

where $\hat{\boldsymbol{e}}_c \in \mathbb{R}^3$ is the unit vector of estimated contact direction defined as $\hat{\boldsymbol{e}}_c := \hat{\boldsymbol{d}}^c/\|\hat{\boldsymbol{d}}^c\|$, β is a factor to regulate the smoothness of the constrained motion, and $\boldsymbol{J}_c$ is the Jacobian matrix at the contact location $\boldsymbol{x}_c$. The derivation of $\boldsymbol{J}_c$ from the end-effector Jacobian $\boldsymbol{J}_e$ can be found in [46]. Therefore, the commanded velocity $\dot{\boldsymbol{\theta}}_d$ can be solved as

$$\dot{\boldsymbol{\theta}}_d = \begin{cases} \arg\min \mathcal{J}(\dot{\boldsymbol{\theta}}_d), & \text{if } \|\hat{\boldsymbol{d}}^c\| \geq \epsilon_d \\ \arg\min \mathcal{J}(\dot{\boldsymbol{\theta}}_d), \text{ s.t. } \mathcal{C}(\dot{\boldsymbol{\theta}}_d) \leq d_{\max}^2, & \text{otherwise} \end{cases}, \tag{24}$$

where the constraint is imposed only when the robot has come into contact, detected as $\|\hat{\boldsymbol{d}}^c\|$ exceeding a threshold ϵ_d.

Lastly, Algorithm 1 outlines the procedure in which the robot is instructed to sequentially reach multiple target locations, with the respective sensing modes activated at different phases based on the assessment of risk level (Strategy I). In this procedure, ProTac switches from *proximity* to *tactile* mode when the risk score $r \geq \epsilon_p$ and the obstacle is on the way to

Algorithm 1 Motion control with contact and obs. awareness

Input: $\mathbf{X}_\mathrm{G} := [\boldsymbol{x}_\mathrm{G}^1, \boldsymbol{x}_\mathrm{G}^2, \cdots]$: a sequence of target locations
Output: $\dot{\boldsymbol{\theta}}_d$: commanded joint velocities

mode $\leftarrow$ *proximity* ▷ activate proximity mode
for $\boldsymbol{x}_\mathrm{G}$ in $\mathbf{X}_\mathrm{G}$ **do**
  while $\|\Delta\boldsymbol{x}\| > 10^{-3}$ **do**
    r, $\hat{\boldsymbol{n}}^c$, $\hat{\boldsymbol{d}}^c \leftarrow$ obtain sensing signals from ProTac
    if $r \geq \epsilon_d$ and $c_\mathrm{sim}(\Delta\boldsymbol{x}, \hat{\boldsymbol{n}}^c) > 0$ **then**
      mode $\leftarrow$ *tactile* ▷ switch to tactile mode
    end if
    if $\|\hat{\boldsymbol{d}}^c\| \geq \epsilon_d$ **then**
      $\dot{\boldsymbol{\theta}}_d \leftarrow \arg\min \mathcal{J}(\dot{\boldsymbol{\theta}}_d)$, s.t. $\mathcal{C}(\dot{\boldsymbol{\theta}}_d) \leq d_\mathrm{max}^2$
    else
      $\dot{\boldsymbol{\theta}}_d \leftarrow \arg\min \mathcal{J}(\dot{\boldsymbol{\theta}}_d)$
    end if
  end while
  mode $\leftarrow$ *proximity* ▷ get back to proximity mode
end for

Algorithm 2 Human-robot interaction with *flickering* sensing

Input: $\dot{\boldsymbol{\theta}}_0$: base joint velocities, T_e: execution time
Output: $\dot{\boldsymbol{\theta}}_d$: commanded joint velocities

mode $\leftarrow$ *proximity* ▷ get in *coexistence* state
$\dot{\boldsymbol{\theta}}_d \leftarrow \dot{\boldsymbol{\theta}}_0$ ▷ initialize normal operation
while $t < T_e$ **do**
  r, $\hat{\boldsymbol{d}}^c \leftarrow$ obtain sensing signals from ProTac
  if *human* detected **then**
    mode $\leftarrow$ *flickering*
    $\dot{\boldsymbol{\theta}}_d \leftarrow \mathbf{0}$ **if** $r \geq \epsilon_p$ **else** $\dot{\boldsymbol{\theta}}_d \leftarrow \dot{\boldsymbol{\theta}}_0/k$ ▷ k, see (21)
    if $\|\hat{\boldsymbol{d}}^c\| \geq \epsilon_d$ **then**
      mode $\leftarrow$ *tactile* ▷ get into *interaction* phase
      while $L < 2$ **do** ▷ L denotes #contacts
        $\dot{\boldsymbol{\theta}}_d \leftarrow$ obtain from $\hat{\boldsymbol{d}}_c$ (refer to [34])
      end while
    end if
  else
    mode $\leftarrow$ *proximity* ▷ return to *coexistence* phase
    $\dot{\boldsymbol{\theta}}_d \leftarrow \dot{\boldsymbol{\theta}}_0$
  end if
end while

the target location which is determined by the cosine similarity $c_\mathrm{sim}(\Delta\boldsymbol{x}, \hat{\boldsymbol{n}}^c)$ between $\Delta\boldsymbol{x}$ and $\hat{\boldsymbol{n}}^c$, that is

$$c_\mathrm{sim}(\Delta\boldsymbol{x}, \hat{\boldsymbol{n}}^c) := \frac{\Delta\boldsymbol{x} \cdot \hat{\boldsymbol{n}}^c}{\|\Delta\boldsymbol{x}\|\|\hat{\boldsymbol{n}}^c\|} > 0. \tag{25}$$

C. Human-robot interaction with ProTac flickering sensing

Human-robot interaction commonly involves two phases: a coexistence phase, where robots work alongside humans with safe behaviors such as collision avoidance or speed adjustments, and a physical interaction phase, where robots engage in physical interaction with humans [49]. Nonetheless, seamlessly transitioning from the coexistence to the physical interaction phase or recognizing human intention for physical interaction remains challenging and requires sophisticated perception and learning techniques [50]. To address this challenge, we leverage the proposed intention-based mode switching (Strategy II), incorporating the ProTac's flickering sensing. This facilitates the detection of human-intended contacts, allowing the robot to seamlessly transition from a coexistence to a physical interaction state.

Consider a scenario, where a robot integrated with ProTac, initially in a coexistence phase, operates under normal conditions with a base velocity $\dot{\boldsymbol{\theta}}_d := \dot{\boldsymbol{\theta}}_0$, while ProTac is in *proximity* mode. In this phase, if a human is detected, the *flickering* mode is activated. In the *flickering* mode, the controller triggers the speed regulation process to update the reduced speed $\dot{\boldsymbol{\theta}}_d$ based on the factor k (refer to Sec. VI-B), or stop the robot motion ($\dot{\boldsymbol{\theta}}_d = \mathbf{0}$) when the risk score r exceeds a certain threshold ϵ_p. In short, the commanded joint velocity $\dot{\boldsymbol{\theta}}_d$ can be computed with corresponding conditions as (considering the ProTac being in the *flickering* mode)

$$\dot{\boldsymbol{\theta}}_d = \begin{cases} \dot{\boldsymbol{\theta}}_0/k, & \text{if human detected and } r < \epsilon_p \\ \mathbf{0}, & \text{else human detected and } r \geq \epsilon_p \\ \dot{\boldsymbol{\theta}}_0, & \text{otherwise } \textit{not} \text{ human detected} \end{cases}, \tag{26}$$

Additionally, if the human is no longer observed, the robot speed reverts to the base profile $\dot{\boldsymbol{\theta}}_0$. However, the transition to the physical interaction phase occurs upon detecting human-intended contact, signaled by the estimated contact depth $\|\hat{\boldsymbol{d}}^c\|$ exceeding a contact threshold ϵ_d.

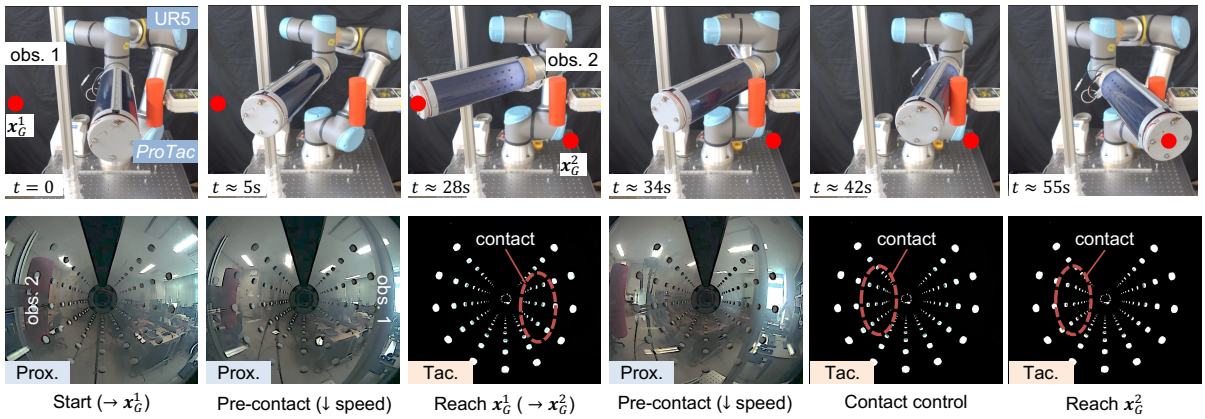
Upon the transition, ProTac switches to the *tactile* mode to enable the interaction phase. In this phase, the commanded velocity $\dot{\boldsymbol{\theta}}_d$ can be determined from physical interactions with the human. Here, we employ the strategy proposed in [34] to guide the robot motion based on the contact depth vector $\hat{\boldsymbol{d}}^c$. Last but not least, the recognition of human-intended two-point contact can be considered as a condition to terminate the interaction phase, allowing the robot to return to normal operation. The overview of this scenario is summarized in Algorithm 2.

TABLE IV
CONTROL PARAMETERS FOR PROTAC-DRIVEN MULTIMODAL TASKS

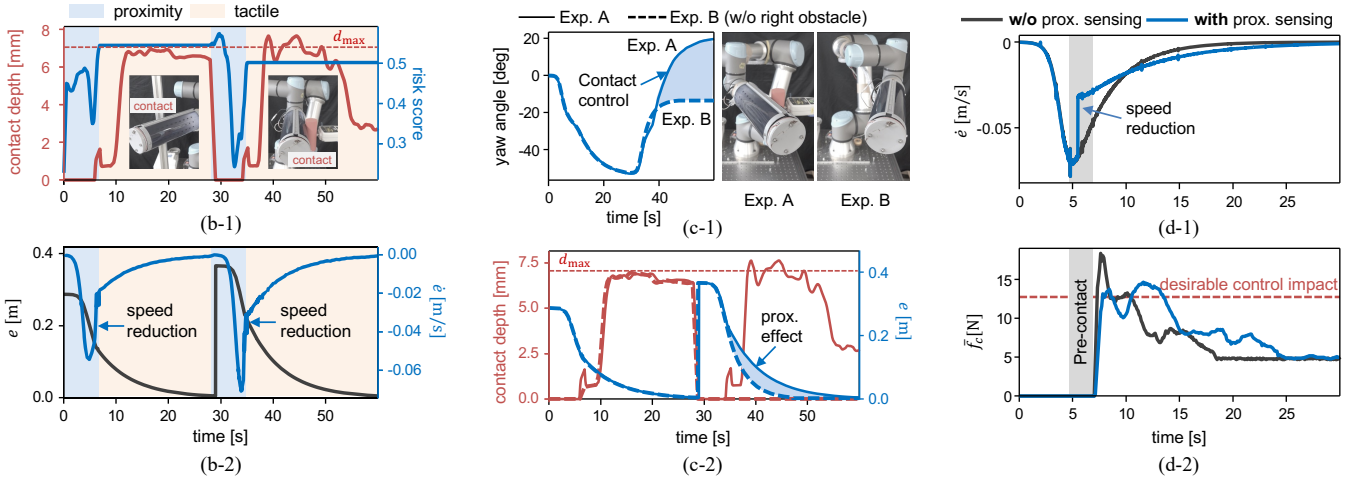
Parameter		Value	Unit
Diagonal proportional matrix	$\boldsymbol{K}_p$	$\mathrm{diag}(0.3, 0.3, 0.3)$	-
Regularization factor	β	0.5	-
Max. admissible contact depth	d_max	7.0	mm
Contact threshold	ϵ_d	2.0	mm
Critical risk threshold	ϵ_p	0.45	-

D. Experiment

Setups: This section validates the efficacy of ProTac for multimodal tasks, particularly in enhancing motion control and human-robot interaction scenarios. To this end, we employed the 6-DoF UR5e robot arm, equipped with ProTac as an extended link attaching to the robot's end-effector (see Fig. 11a). This setup aims to showcase an additional use case of ProTac for existing commercial robot arms, complementing the custom-built ProTac-integrated robot demonstrated in the previous section. Here, the commanded joint velocities $\dot{\boldsymbol{\theta}}_d$, derived from the proposed control strategies based on ProTac



(a) Video stills of motion rollout and corresponding images of ProTac views (obs. stands for obstacle). The red dots in the upper row's pictures indicate the target position for the end-effector.



(b) The ProTac measurements and position-controlled error dynamics

(c) The motion resulted from contact constraint (Exp. A) and contact-free motion (Exp. B)

(d) The effect of proximity sensing on the impact reduction for contact-constrained motion control

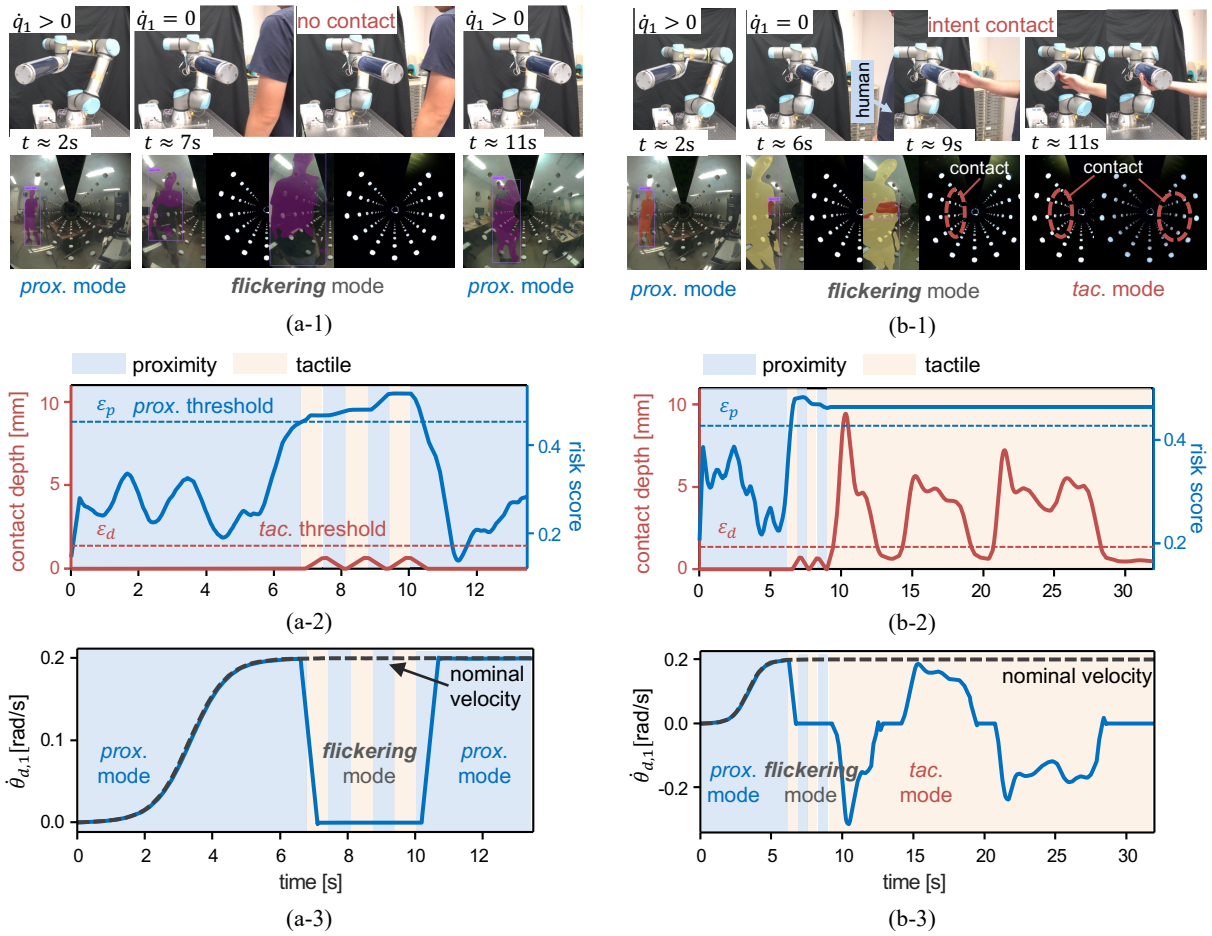
Fig. 11. **Demonstration of adaptive motion control with contact and obstacle awareness.** (a) Experimental setup and rollout sequence of ProTac-enabled motion control with contact and obstacle awareness. (b) Logs of ProTac feedback and error dynamics, showcasing the system's effectiveness in achieving the task objective, which may be challenging to attain with a conventional rigid link. (c) Results demonstrating the system's behavior and effectiveness in contact-constrained motion. (d) Ablation results comparing the system's behavior with and without ProTac-enabled proximity sensing, demonstrating its impact in reducing peak impact force during the contact-constrained motion phase.

feedback, were regulated by the UR5's low-level controller and coordinated via ROS. The parameter values of the controllers are summarized in Table IV.

1) ProTac-enhanced motion control: The experiment demonstrates how ProTac with proximity-tactile sensing modalities integrated into the optimization controller (see Section VII-B) can enhance the robot's motion in a cluttered environment. The QP optimization problem (24) was solved numerically using the CVXPY optimization library[5]. Here, the ProTac-integrated robot was tasked to sequentially reach two target locations $\boldsymbol{x}_{\mathrm{G}}^1$ and $\boldsymbol{x}_{\mathrm{G}}^2$ (marked by red circles in Fig. 11a), while attempting to maintain the magnitude of contact depth $\|\hat{\boldsymbol{d}}^c\|$ below $d_{\max}$ to mitigate the impacts of possible contacts. The two obstacles were positioned close to each other and target locations as well, reproducing a cluttered environment (see Fig. 11a). Furthermore, we attached obstacle-2 (the right one) to a force gauge (ZTS50N, IMADA Inc., Japan) to assess the true impact force exerted on the ProTac link upon contact.

Figure 11a illustrates the robot's task performance and corresponding ProTac views for different sensing modes. The video stills at $t \approx 42\,\mathrm{s}$ and $t \approx 55\,\mathrm{s}$ showcase how the controller harnessed the deformation of *soft* ProTac skin to compensate for the goal error, in which the soft skin can accommodate and gently conform to the obstacle. This behavior is difficult, or not possible to achieve by a rigid link. Quantitatively, Figure 11b illustrates the ProTac measurements (Fig. 11b-1) and the dynamics of the positional error $e := \|\Delta \boldsymbol{x}\|$ with relative to the target locations (Fig. 11b-2), evolving in the experimental scenario. Figure 11b shows that while the robot was approaching the target location, the robot slowed down the speed based on the increased risk score r and switched to the *tactile* mode as $r \geq \epsilon_p$. Once contact occurred, the robot attempted to maintain the contact depth around the predefined admissible threshold $d_{\max}$. Notably, when the location

[5]CVXPY is a Python tool for solving convex optimization problems.



(a) Experiment A: a human passerby without any interaction intention. (b) Experiment B: a human intending physical touch interaction.

Fig. 12. **Demonstration of multi-phase human-robot interaction scenarios enabled by *flickering* sensing mode.** (a) Scenario where the robot stops upon detecting a nearby human, and resumes its planned nominal velocity profile when the human moves away. (b) Scenario where ProTac detects human contact, enabling tactile-based interaction where the human guides the robot's motion. The *flickering* mode facilitates the recognition of intentional contact and seamless switching between robot operation modes.

error e was sufficiently close to 0, the contact depth remained observable on the soft ProTac link (see Fig. 11b), which demonstrates the benefit of ProTac's soft skin in facilitating target-driven motion control in unstructured environments, a capability that would be infeasible with a rigid link.

Figure 11c showcases the effect of contact-based control and obstacle awareness in scenarios with and without an obstacle on the right side (*i.e.*, obs. 2), denoted as Exp-A and Exp-B, respectively. Specifically, in Exp-A, the robot did incur a significant rotation in the yaw angle to accommodate the contact with the obstacle (Fig. 11c-1), as well as converged to the target location slower than Exp-B (Fig. 11c-2) since the robot slowed down the speed as a result of obstacle awareness enabled by the proximity sensing. Moreover, the incorporation of proximity sensing into the motion control, indeed, mitigates the impact of the ProTac link with an obstacle, as validated in Figure 11d. With proximity sensing, Fig. 11d-2 shows that the actual impact force measured by the force gauge $\bar{f}_c$ was suppressed to approximately the desirable threshold, while a significant peak impact force was observed in the scenario without proximity integration.

The obtained results confirm that the soft ProTac skin, combining multimodal sensing with optimization-based control, enhances motion control in cluttered environments where contacts are unavoidable.

2) ProTac-enhanced human-robot interaction: The experiment demonstrates the utility of distinct ProTac sensing modes in facilitating seamless multi-phase human-robot interaction scenarios (see Section VII-C). In this experiment, a switching period of $T_s = 0.5\,s$ is set for the *flickering* mode, which was empirically chosen to balance the quality of the contact sensing signal with its responsiveness to human contacts. Human detection and tracking were performed using the open-source DEVA pipeline [51]. The robot arm was programmed to follow a nominal joint-space velocity profile $\dot{\boldsymbol{\theta}}_0 = [\dot{\theta}_1, \mathbf{0}_5]_0^\intercal$. Two experimental scenarios were tested as follows:

- Experiment A: a human without interaction intent approaches and then moves away from the robot (Fig. 12a).
- Experiment B: a human with interaction intent approaches and touches the robot to initiate physical interaction (Fig. 12b).

Figures 12a-1 and 12b-1 illustrate the experimental sce-

narios and ProTac views for different sensing modes in Experiment A and Experiment B, respectively. Additionally, the robot's behavior is characterized in terms of the commanded velocity profile of the base joint, $\dot{\theta}_{d,1}$, for Experiment A and Experiment B in Figures 12a-3 and 12b-3, respectively. Specifically, as depicted in Figures 12a-2 and 12b-2, the approaching human led to an increasing risk score r observed in the *proximity* mode. Once r surpassed the critical risk threshold ϵ_p, the robot came to a halt ($\dot{\theta}_{d,1} = 0$), activating the *flickering* mode, indicated by the alternating orange and blue-shaded strips in Figs. 12a-2 and 12b-2.

During the *flickering* mode, in Experiment A, contact was not detected ($\|\hat{\boldsymbol{d}}^c\| < \epsilon_p$), and the human moved away ($r < \epsilon_p$), as shown in Fig. 12a-2. Consequently, the robot reverted to the nominal velocity profile $\dot{\theta}_{d,1} := \dot{\theta}_{0,1}$ (Fig. 12a-3). In contrast, in Experiment B, contact occurred, determined by $\|\hat{\boldsymbol{d}}^c\| \geq \epsilon_p$, immediately triggering the *tactile* mode and interaction phase (as depicted in Fig. 12b-2). As a result, the robot moved in response to the human's touch-based interaction (Fig. 12b-3). Therefore, the latter experiment demonstrates the effectiveness of the *flickering* mode in facilitating the seamless switch between the proximity and tactile modalities for multi-phase human-robot interaction.

VIII. Discussion

A. Bimodal Perception Performance

The proposed tactile sensing pipeline, incorporating a feature-level fusion method for stereo tactile inputs, achieves zero-shot learning and improves performance over the previous approach [34]. For proximity sensing, the system achieves accuracy comparable to [4] without requiring an additional infrared time-of-flight (ToF) camera, though performance is slightly lower than [6]. However, it is important to note that [6] relies on extensive manual data collection and labeling. In contrast, our approach leverages a large-scale, open-source simulation dataset with fine-tuning adaptation, significantly reducing the overhead associated with data preparation. As a result, the proposed method demonstrates strong potential for scalable, zero-shot transfer of bimodal sensing capabilities across devices, particularly for large-area skin systems, mitigating the challenges of manual data collection.

In terms of runtime performance, both the tactile and proximity pipelines operate within the range required for real-time applications. The tactile perception pipeline using feature-level fusion achieves an average inference time of approximately 10 ms (or 100 Hz), despite incurring slightly higher latency than single-camera baselines (Table II). For proximity sensing, the average inference time is approximately 20 ms (50 Hz) with a single-camera configuration and 40 ms (25 Hz) with two-camera fusion. Although the fusion schemes introduce additional latency, this can potentially be reduced through parallel processing. It should be noted that while the tactile and proximity pipelines can achieve inference rates of approximately 100 Hz and 25 Hz, respectively, their effective processing speeds are limited by the 30 Hz frame rate of the RGB cameras, resulting in operational rates of approximately 30 Hz for both pipelines. The performance confirms that ProTac's binocular configuration, combined with appropriate fusion strategies, enables effective bimodal perception without compromising runtime efficiency, highlighting its potential for practical real-time applications.

Beyond low-level sensing, our system also shows promise for high-level perception tasks. As illustrated in Fig. 12, human detection and segmentation can be performed using an off-the-shelf DEVA model applied to ProTac's transparent views, indicating its potential for broader high-level perception applications. Lastly, *flickering* sensing highlights the potential for enhanced simultaneous proximity-tactile sensing, possibly through advances in signal processing or data-driven methods.

B. Limitations

1) Hardware: To achieve the cylindrical form and enhance the structural integrity of the ProTac link, a tri-layered arrangement of braces composed of steel and polylactic acid (PLA) is employed. While this configuration is essential for mechanical stability and payload support, it introduces 'blind' regions in the tactile sensing field, particularly at the brace locations. Another limitation stems from the construction of the PDLC film, which includes two cover layers made of polyethylene terephthalate (PET), a relatively stiff material. This rigidity reduces the skin's compliance and, in turn, diminishes its sensitivity to fine contact, making it less responsive than softer tactile systems, such as TacLINK [34]. Additionally, the use of commercial flat PDLC film limits the design's adaptability to curved or custom geometries. To address this, custom-fabricated PDLC film with built-in curvature and softer cover layers would enable more flexible integration while maintaining optical functionality. Lastly, design factors such as the arrangement, density, and size of visual markers significantly influence sensing performance. The current marker configuration was empirically determined to balance tactile and proximity sensing while minimizing occlusion of the external scene during proximity mode. Future work will focus on systematically optimizing the marker design and other skin parameters to enhance bimodal perception performance.

2) Sensing and Control System Performance: While the proposed ProTac system demonstrates effective real-time performance for tactile and proximity sensing, several limitations remain regarding perception accuracy, sensing coverage, and evaluation methodology. First, although the feature-level fusion scheme achieves consistent performance across sensing regions, with full-scale (FS) errors averaging around 25%, this level of error, while comparable to prior works on soft large-area tactile sensing [34], [43], may not suffice for applications requiring fine-grained contact precision. The elevated error observed in our system can be attributed to the increased complexity of the multilayered ProTac skin. Nonetheless, this accuracy has proven sufficient for interactive control tasks, as shown in Section VII. Future work could explore the integration of temporal modeling to enhance the continuity and robustness of contact estimation. Another sensing limitation is the system's inability to estimate tangential forces. Due to the PDLC skin's high in-plane stiffness, it primarily responds to normal deformation. Extending to full 3D force reconstruction

will require modeling tangential components, which our simulation framework can support via non-frictionless interaction settings and Jacobian-based force modeling. In addition, the tactile perception shows limited ability to distinguish closely spaced multi-point contacts (*e.g.*, within 10 cm), likely due to the absence of such examples in the training data. Improving this capability remains an open challenge and may be addressed through targeted data augmentation or model refinement [34].

For proximity sensing, the system still suffers from performance degradation in regions distant from both camera ends. Even with two-camera fusion, a drop of approximately 40% in risk score estimation occurs when the obstacle is positioned near the center of the link. We attribute this to optical distortion introduced by the skin's material properties and variations in external lighting, which affect depth estimation quality. Addressing this issue may require mitigating optical distortions, along with developing improved depth reconstruction models and fusion strategies designed to compensate for region-specific signal degradation, ultimately aiming to expand ProTac's perceptual coverage.

Lastly, although this study demonstrates the practical integration of ProTac into control tasks, a more rigorous evaluation using standardized quantitative metrics would strengthen our understanding of the system's applicability and enable more meaningful comparisons with other approaches. We regard this as an important future direction to support systematic benchmarking and comprehensive validation.

IX. Conclusion

This paper presents a novel vision-based proximity-tactile sensing technology for soft robotic devices, enabled by actively switching the skin's transparency. The proposed system, implemented in the ProTac link, offers a compact design without complex electronics or wiring, supporting operation in proximity, tactile, or combined flickering modes. Tactile and proximity perception are realized through zero-shot deformation sensing and monocular depth estimation, respectively. By leveraging two opposing cameras, ProTac can improve both the sensing coverage and perception performance of large-scale robot bodies. Lastly, we demonstrate the integration of the ProTac link with both a custom soft robot and a commercial robot arm to enable safe interaction within contact-rich environments, capabilities that are challenging to achieve with conventional rigid single-modal robotic systems.

Appendix A
DepthNet Loss function

The DepthNet loss function $\mathcal{L}_{\text{DepthNet}}$ is composed of $\mathcal{L}_{\text{ssitrim}}$ and $\mathcal{L}_{\text{grad}}$ loss terms (1). First, $\mathcal{L}_{\text{ssitrim}}$ penalizes the absolute difference in depth values between $\mathbf{z}^{\text{est}}$ and $\mathbf{z}^{\text{gt}}$. Thus, $\mathcal{L}_{\text{ssitrim}}$ can be expressed as:

$$\mathcal{L}_{\text{ssitrim}} = \frac{1}{2|\mathcal{K}|}\sum_{j=1}^{U_m} |\bar{\mathbf{z}}_j^{\text{est}} - \bar{\mathbf{z}}_j^{\text{gt}}|, \tag{27}$$

where $\bar{\mathbf{z}}^{\text{est}}$, $\bar{\mathbf{z}}^{\text{gt}}$ denote the normalized depth prediction and ground-truth (*i.e.*, zero mean and unit scale) and $|\mathcal{K}|$ is the number of valid pixels. In addition, to enhance the training robustness, the 20% largest residuals of the depth deviation are trimmed out such that $|\bar{\mathbf{z}}_j^{\text{est}} - \bar{\mathbf{z}}_j^{\text{gt}}| \leq |\bar{\mathbf{z}}_{j+1}^{\text{est}} - \bar{\mathbf{z}}_{j+1}^{\text{gt}}|$ and $U_m = 0.8|\mathcal{K}|$ [38]. Second, the multi-scale gradient term $\mathcal{L}_{\text{grad}}$ encourages sharp depth discontinuities and smooth gradient changes by computing the sum of absolute difference between the predicted depth derivatives and the ground-truth depth derivatives (along x and y direction) at multiple scale ($M = 4$), by which the image resolution is halved at each scale level [52]:

$$\mathcal{L}_{\text{grad}} = \frac{1}{|\mathcal{K}|}\sum_{m=1}^{M}\sum_{j=1}^{|\mathcal{K}|}(|\nabla_x R_j^m| + |\nabla_y R_j^m|), \tag{28}$$

where R^m represents the difference of depth maps at scale m with $R_j = \bar{\mathbf{z}}_j^{\text{est}} - \bar{\mathbf{z}}_j^{\text{gt}}$. Detailed derivations of the normalized depth values and the loss functions can be found in [38].

Appendix B
Mapping of obstacle points in 3D space

Given the fisheye-lens camera of the ProTac sensor is modeled as a classic pinhole with the assumption that pixel sensors are square in shape (*i.e.*, the focal lengths f along x- and y-axes are equal $f = f_x = f_y$), the geometrical relationship between the 3D position of an obstacle point $\mathbf{o} = [o_x, o_y, o_z]^\intercal \in \mathbb{R}^3$ in PCS (ProTac coordinate system) and its projection $\mathbf{p} = [p_x, p_y, p_d]^\intercal \in \mathbb{R}^3$ on the depth space with p_d obtained from the estimated map $\mathbf{Z}^{\text{est}}$ at a specific pixel location $(u, v)^\intercal$ can be derived as (here, we omit the subscript j for $\mathbf{o}$ and $\mathbf{p}$, for short)

$$\begin{bmatrix} p_x \\ p_y \end{bmatrix} = \frac{f}{b + o_z}\begin{bmatrix} o_x \\ o_y \end{bmatrix},\ p_d = o_z \tag{29}$$

with

$$p_x = u - c_x,\ p_y = v - c_y, \tag{30}$$

where b denotes the position of PCS origin in the camera frame; (c_x, c_y) is the pixel location of the image principal point on the pixel uv-coordinate. Thus, the obstacle location in Cartesian space could be algebraically calculated as:

$$o_x = \frac{(u - c_x)(b + o_z)}{f},\ o_y = \frac{(v - c_y)(b + o_z)}{f},\ o_z = p_d. \tag{31}$$

The calibration of model parameters $\{f, c_x, c_y, b\}$ and the fisheye-lens correction were conducted following the method proposed in [33].

References

[1] A. Bicchi, M. A. Peshkin, and J. E. Colgate, *Safety for Physical Human–Robot Interaction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 1335–1348.

[2] A. De Santis, B. Siciliano, A. De Luca, and A. Bicchi, "An atlas of physical human–robot interaction," *Mechanism and Machine Theory*, vol. 43, no. 3, pp. 253–270, 2008.

[3] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, 2016, pp. 1045–1051.

[4] J. Yin, G. M. Campbell, J. Pikul, and M. Yim, "Multimodal proximity and visuotactile sensing with a selectively transmissive soft membrane," in *2022 IEEE International Conference on Soft Robotics (RoboSoft)*, 2022, pp. 802–808.

[5] Q. Wang, Y. Du, and M. Y. Wang, "Spectac: A visual-tactile dual-modality sensor using uv illumination," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 10 844–10 850.

[6] F. R. Hogan, J.-F. Tremblay, B. H. Baghi, M. Jenkin, K. Siddiqi, and G. Dudek, "Finger-sts: Combined proximity and tactile sensing for robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 865–10 872, 2022.

[7] S. Athar, G. Patel, Z. Xu, Q. Qiu, and Y. She, "Vistac toward a unified multimodal sensing finger for robotic manipulation," *IEEE Sensors Journal*, vol. 23, no. 20, pp. 25 440–25 450, 2023.

[8] S. Zhang, Y. Sun, J. Shan, Z. Chen, F. Sun, Y. Yang, and B. Fang, "Tirgel: A visuo-tactile sensor with total internal reflection mechanism for external observation and contact detection," *IEEE Robotics and Automation Letters*, vol. 8, no. 10, pp. 6307–6314, 2023.

[9] E. Roberge, G. Fornes, and J.-P. Roberge, "Stereotac: A novel visuotactile sensor that combines tactile sensing with 3d vision," *IEEE Robotics and Automation Letters*, vol. 8, no. 10, pp. 6291–6298, 2023.

[10] S. Li, H. Yu, G. Pan, H. Tang, J. Zhang, L. Ye, X.-P. Zhang, and W. Ding, "M^3tac: A multispectral multimodal visuotactile sensor with beyond-human sensory capabilities," *IEEE Transactions on Robotics*, vol. 40, pp. 4484–4503, 2024.

[11] A. Schmitz, P. Maiolino, M. Maggiali, L. Natale, G. Cannata, and G. Metta, "Methods and technologies for the implementation of large-scale robot tactile sensors," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 389–400, 2011.

[12] T.-H.-L. Le, P. Maiolino, F. Mastrogiovanni, and G. Cannata, "Skinning a robot: Design methodologies for large-scale robot skin," *IEEE Robotics Automation Magazine*, vol. 23, no. 4, pp. 150–159, 2016.

[13] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile sensing—from humans to humanoids," *IEEE Transactions on Robotics*, vol. 26, no. 1, pp. 1–20, 2010.

[14] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[15] S. E. Navarro, S. Mühlbacher-Karrer, H. Alagi, H. Zangl, K. Koyama, B. Hein, C. Duriez, and J. R. Smith, "Proximity perception in human-centered robotics: A survey on sensing systems and applications," *IEEE Transactions on Robotics*, pp. 1–22, 2021.

[16] V. A. Ho, S. Hirai, and K. Naraki, "Fabric interface with proximity and tactile sensation for human-robot interaction," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 238–245.

[17] J. C. Yang, J. Mun, S. Y. Kwon, S. Park, Z. Bao, and S. Park, "Electronic skin: Recent progress and future prospects for skin-attachable devices for health monitoring, robotics, and prosthetics," *Advanced Materials*, vol. 31, no. 48, p. 1904765, 2019.

[18] G. Pang, G. Yang, and Z. Pang, "Review of robot skin: A potential enabler for safe collaboration, immersive teleoperation, and affective interaction of future collaborative robots," *IEEE Transactions on Medical Robotics and Bionics*, vol. 3, no. 3, pp. 681–700, 2021.

[19] H. Park, K. Park, S. Mo, and J. Kim, "Deep neural network based electrical impedance tomographic sensing methodology for large-area robotic tactile sensing," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1570–1583, 2021.

[20] P. Mittendorfer and G. Cheng, "Humanoid multimodal tactile-sensing modules," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 401–410, 2011.

[21] Q. Leboutet, E. Dean-Leon, F. Bergner, and G. Cheng, "Tactile-based whole-body compliance with force propagation for mobile manipulators," *IEEE Transactions on Robotics*, vol. 35, no. 2, pp. 330–342, 2019.

[22] G. Cheng, E. Dean-Leon, F. Bergner, J. Rogelio Guadarrama Olvera, Q. Leboutet, and P. Mittendorfer, "A comprehensive realization of robot skin: Sensors, sensing, control, and applications," *Proceedings of the IEEE*, vol. 107, no. 10, pp. 2034–2051, 2019.

[23] S. Armleder, E. Dean-Leon, F. Bergner, and G. Cheng, "Interactive force control based on multimodal robot skin for physical human-robot collaboration," *Advanced Intelligent Systems*, vol. 4, no. 2, p. 2100047, 2022.

[24] K. Kamiyama, H. Kajimoto, N. Kawakami, and S. Tachi, "Evaluation of a vision-based tactile sensor," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, vol. 2, 2004, pp. 1542–1547 Vol.2.

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

[26] 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.

[27] W. K. Do and M. Kennedy, "Densetact: Optical tactile sensor for dense shape reconstruction," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 6188–6194.

[28] Z. Lin, J. Zhuang, Y. Li, X. Wu, S. Luo, D. F. Gomes, F. Huang, and Z. Yang, "Gelfinger: A novel visual-tactile sensor with multi-angle tactile image stitching," *IEEE Robotics and Automation Letters*, vol. 8, no. 9, pp. 5982–5989, 2023.

[29] J. Zhao and E. H. Adelson, "Gelsight svelte: A human finger-shaped single-camera tactile robot finger with large sensing coverage and proprioceptive sensing," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 8979–8984.

[30] K. Sato, K. Kamiyama, N. Kawakami, and S. Tachi, "Finger-shaped gelforce: Sensor for measuring surface traction fields for robotic hand," *IEEE Transactions on Haptics*, vol. 3, no. 1, pp. 37–47, 2010.

[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*, 2021.

[32] 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.

[33] L. Van Duong and V. A. Ho, "Large-scale vision-based tactile sensing for robot links: Design, modeling, and evaluation," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 390–403, 2021.

[34] Q. K. Luu, N. H. Nguyen, and V. A. Ho, "Simulation, learning, and application of vision-based tactile sensing at large scale," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 2003–2019, 2023.

[35] Q. K. Luu, D. Q. Nguyen, N. H. Nguyen, and V. A. Ho, "Soft robotic link with controllable transparency for vision-based tactile and proximity sensing," in *2023 IEEE International Conference on Soft Robotics (RoboSoft)*, 2023, pp. 1–6.

[36] F. Flacco, T. Kröger, A. De Luca, and O. Khatib, "A depth space approach to human-robot collision avoidance," in *2012 IEEE International Conference on Robotics and Automation*, 2012, pp. 338–345.

[37] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2003.

[38] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 03, pp. 1623–1637, mar 2022.

[39] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman, "Learning the depths of moving people by watching frozen people," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4516–4525, 2019.

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

[41] S. Haddadin, A. De Luca, and A. Albu-Schäffer, "Robot collisions: A survey on detection, isolation, and identification," *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1292–1312, 2017.

[42] E. Coevoet, T. Morales-Bieze, F. Largilliere, Z. Zhang, M. Thieffry, M. Sanz-Lopez, B. Carrez, D. Marchal, O. Goury, J. Dequidt, and C. Duriez, "Software toolkit for modeling, simulation, and control of soft robots," *Advanced Robotics*, vol. 31, no. 22, pp. 1208–1224, 2017.

[43] V. A. Ho and S. Nakayama, "Iotouch: whole-body tactile sensing technology toward the tele-touch," *Advanced Robotics*, vol. 35, no. 11, pp. 685–696, 2021.

[44] T. T. Nguyen, Q. K. Luu, D. Q. Nguyen, and V. Ho, "ConTac: Continuum-Emulated Soft Skinned Arm with Vision-based Shape Sensing and Contact-aware Manipulation," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, July 2024.

[45] S. Haddadin, A. Albu-Schaffer, A. De Luca, and G. Hirzinger, "Collision detection and reaction: A contribution to safe physical human-robot interaction," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2008, pp. 3356–3363.

[46] K. M. Lynch and F. C. Park, *Modern Robotics: Mechanics, Planning, and Control*, 1st ed. USA: Cambridge University Press, 2017.

[47] Q. K. Luu, A. Albini, P. Maiolino, and V. A. Ho, "Taclink-integrated robot arm toward safe human-robot interaction," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 12 356–12 362.

[48] A. Jain, M. D. Killpack, A. Edsinger, and C. C. Kemp, "Reaching in clutter with whole-arm tactile sensing," *The International Journal of Robotics Research*, vol. 32, no. 4, pp. 458–482, 2013.

[49] A. De Luca and F. Flacco, "Integrated control for phri: Collision avoidance, detection, reaction and collaboration," in *2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 288–295.
[50] S. Golz, C. Osendorfer, and S. Haddadin, "Using tactile sensation for learning contact knowledge: Discriminate collision from physical interaction," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 3788–3794.
[51] H. K. Cheng, S. W. Oh, B. Price, A. Schwing, and J.-Y. Lee, "Tracking anything with decoupled video segmentation," in *ICCV*, 2023.
[52] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Quan Khanh Luu (Member, IEEE) received the B.Eng. degree in mechatronics from Ho Chi Minh City University of Technology, Vietnam, in 2018, and the M.S. and Ph.D. degrees in robotics from the Japan Advanced Institute of Science and Technology (JAIST), Japan, in 2021 and 2024, respectively.

He worked as a software engineer at Bosch Corp., Ho Chi Minh City, in 2019 and joined Carnegie Mellon University, USA, as a visiting researcher in 2023. He is currently a postdoctoral research associate at Purdue University, USA. His research interests include tactile sensing, robot learning and control, and their applications in safe interaction with contact-rich environments.

He was awarded the prestigious Japan Society for the Promotion of Science (JSPS) Research Fellowship for Young Scientists (DC2) and was selected as a member of the RSS Pioneers cohort (2024). He was also nominated for the Best Paper Finalist at IEEE/SICE SII (2024). He was honored as the Best Graduate Student for outstanding academic performance during his master's degree. He served on the Program Committee for RSS Pioneers (2025).

Dinh Quang Nguyen (Member, IEEE) received the degree of Engineer and Master of Science, in the field of Mechanical Engineering, from the Hanoi University of Science and Technology (HUST), Vietnam, in 2011 and 2013, respectively. From 2013 to 2018, he worked as a lecturer at Hanoi University of Industry (HaUI), Vietnam. He achieved a Ph.D. degree in School of Materials Science from Japan Institute of Science and Technology (JAIST), Ishikawa, Japan, in 2022 (supported by MEXT Scholarship). After graduating, Dr. Nguyen led a laboratory on simulation and high-performance computing (SHPC) at the Institute of Technology, HaUI. From 2024, he joined Department of Robotics Engineering, Faculty of Electronics and Telecommunications, University of Engineering and Technology - Vietnam National University (VNU-UET). His research interests are soft robotics, biomimetics, functional materials, and mechanical engineering.

Nhan Huu Nguyen (Member, IEEE) has been working as a postdoctoral researcher in Soft Haptics Labs, Japan Advanced Institute of Science and Technology (JAIST), Japan where he received the Ph.D. degree in robotics. He received his bachelor and master degrees in mechanical engineering at The University of Da Nang - University of Science and Technology (Viet Nam - 2015) and Ming Chi University of Science and Technology (Taiwan - 2017). His research focuses on exploiting complex physical interaction between deformable robots and the surrounding environment to enable novel functionalities such as tactile sensing, or facilitate learning and controlling tasks.

Nam Phuong Dam (Student Member, IEEE) received the B.Eng. degree in Robotics from the University of Engineering and Technology, Vietnam National University, Hanoi, Vietnam, in 2023. He is currently pursuing the Master's degree in Robotics at the Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan, where he conducts research in the Soft Haptics Lab.

He has also joined Seoul National University as a Visiting Researcher. His research interests focus on robot learning and control, tactile sensing and perception, and applications for human–robot interaction.

Van Anh Ho (Senior Member, IEEE) received the Ph.D. degree in robotics from Ritsumeikan Unviersity, Kyoto, Japan, in 2012. Before that, he obtained the Bachelor degree in Electrical Engineering at Hanoi University of Science and Technology, Vietnam, in 2007; and Master degree in Mechanical Engineering in 2009 at Ritsumeikan University. He completed the JSPS Postdoctoral Fellowship in 2013 before joining Advanced Research Center Mitsubishi Electric Corp., Japan as a research scientist. From 2015 to 2017, he worked as Assistant Professor with Ryukoku University, Kyoto, Japan where he led a laboratory on soft haptics, soft modeling. From 2017, he joined the Japan Advanced Institute of Science and Technology (JAIST) for setting up a laboratory on soft robotics. His current research interests are soft robotics, soft haptic interaction, tactile sensing, grasping and manipulation, bio-inspired robots.

He was recipient of the prestigious Japan Society for the Promotion of Science (JSPS) Research Fellowship for Young Scientist for his PhD course (DC2) and postdoctoral fellowship. Ho was the recipient of 2019 IEEE Nagoya Chapter Young Researcher Award, Best Paper Finalists at Robotics: Science and Systems (RSS 2023), IEEE SII (2024, 2016) and IEEE RoboSoft (2020). He is member of The Robotics Society of Japan (RSJ), and Senior Member of the IEEE. He is serving as Associate Editor for many international conferences, such as IROS, SII, RoboSoft; as well as for journals such as IEEE Transactions for Robotics (T-RO), IEEE Robotics and Automation Letters (RA-L), and Advanced Robotics. He is General Co-Chair of 2023 IEEE/SICE International Symposium on System Integration (SII 2023), and General Chair of SII 2024.