Vi2TaP: A Cross-Polarization Based Mechanism for Perception Transition in Tactile-Proximity Sensing with Applications to Soft Grippers

Nhan Huu Nguyen[1], Nhat Minh Dinh Le[2], Quan Khanh Luu[3], Tuan Tai Nguyen[4], and Van Anh Ho[4], *Senior Member, IEEE*

***Abstract*—Vision-based soft sensors have emerged as a promising solution for multi-modal sensory systems. Rather than relying on complex integrations of numerous specialized sensors, these devices are advantageous in achieving multiple perceptual capabilities with a single device. However, the unsolved bottleneck for existing systems lies in preventing perceptual interference among visual fields and establishing a reliable mechanism for switching between perception domains. In this study, we present *Vi2TaP*, a novel mechanism leveraging the cross-polarization phenomenon to shift one perception domain to another in a tactile-proximity multimodal sensing paradigm. The core concept involves two polarizer films placed back-to-back. By adjusting the Plane of Polarization (PoP) between 0 to 90 deg, the camera can either fully open its Field-of-View (FoV) to the external environment for proximity sensing or restrict it to the internal space between the polarizers, tailored for tactile sensing. First implementation of *Vi2TaP* is showcased on a soft sensorized gripper. Additionally, we introduce efficient learning pipelines for both proximity and tactile perception, along with effective strategies for extracting valuable information. The experiment results have demonstrated the advantages of the proposed multi-modal sensing scheme in grasping and manipulating tasks. This mechanism is anticipated to accelerate the development and adoption of multimodal vision-based soft sensors across a wide range of practical applications.**

***Index Terms*—Force and Tactile Sensing, Perception for Grasping and Manipulation, Soft Sensors and Actuators, Sensor Fusion**

I. Introduction

Manuscript received: November 20, 2021; Revised February 20, 2025; Accepted: April 17, 2025. This paper was recommended for publication by Editor Yong-Lae Park upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by JSPS Grant-in-Aid for Scientific Research (KAKENHI) grant number 23K19096 and 25K17569. *Corresponding author: Nhan Huu Nguyen.*

[1]Nhan Huu Nguyen is with School of Information Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan. nhnhan@jaist.ac.jp

[2]Nhat Minh Dinh Le was with School of Materials Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan. Now he is with Faculty of Mechanical Engineering, The University of Danang–University of Science and Technology, Danang, 54 Nguyen Luong Bang, 50000 Da Nang, Vietnam. ldmnhat@dut.udn.vn

[3]Quan Khanh Luu was with School of Materials Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan. Now he is with Purdue University, West Lafayette, IN 47907, United States. luu15@purdue.edu

[4]Tuan Tai Nguyen and Van Anh Ho are with School of Materials Science, Japan Advanced Institute of Science and Technology, 1-1 Asahidai, Nomi, Ishikawa, 923-1292, Japan. tuan-nguyen@jaist.ac.jp, van-ho@jaist.ac.jp

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

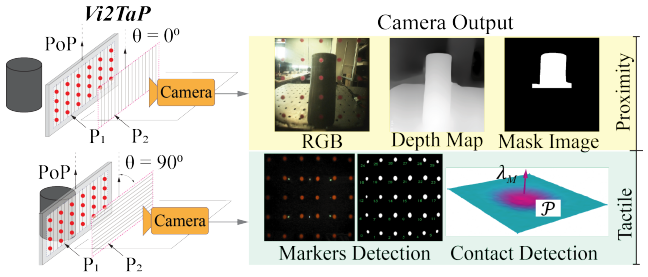


Fig. 1. General scheme for *Vi2TaP* mechanism which is based on the cross-polarization configuration to enable multi-modal visual-tactile sensing modalities.

VISION and touch offer noticeable advantage for robotics perception. Although the former modality has gained much benefit from the advances in vision systems and image-processing techniques, the rapid emergence of tactile sensors has recently been observed. Perceived tactile feedback (*e.g.,* contact forces, contours) unlocks various intelligent robotic behaviors [1] that potentially bring profit to areas requiring rich contacts such as the medical field [2], and agriculture [3]. However, an inevitable perception gap exists due to occlusion issues in the vision system and the lack of pre-touch information. To tackle the above issue, proximity perception is expected to fill this shortage with a pre-contact supervising channel to prevent unintentional collisions and occlusions during operation [4]. Efficient integration of proximity sensing modality as a complementary for tactile perception is still in the infancy stage.

Existing literature explores a great deal of multi-modal sensing devices that include both proximity and tactile modes [5], [6]. The most straightforward approach involves assigning each mode to a specialized measurement element [7]. Despite offering several advantages, these tactile sensors also come with a range of limitations. Firstly, the distributed network of sensing components covering the touch surface presents significant wiring and fabrication issues in addition to the complexity of data acquisition and processing [8], [9]. Secondly, with the growing interest in flexible touch interfaces, the integration of sensing units (mostly rigid) within a soft body can hinder inherent compliance [10].

From another perspective, the coexistence of several sensing abilities using only one type of sensing component has become preferable. Vision-based tactile sensory systems (hereafter

shortened as *ViTac*) have stood out as a dominant method for developing tactile sensors thanks to numerous advantages, including high resolution, ease of fabrication, reliability, cost-effectiveness, and non-invasive integration [11]–[13]. To highlight the contrast of visual cues, the touch surface of most existing ViTac systems is typically opaque to isolate the internal space from the outer. The authors in [14] established the first proof-of-concept for soft transparent skin with printed markers, named *FingerVision*, to acquire both visual and tactile perception. This sensor exhibited the effectiveness of proximity information in complementing tactile sensing to perform grasping tasks. However, due to the transparent skin, perception fields of tactile and proximity modes are overlapped causing degradation in sensing capability, particularly in unstable lighting conditions. Following this pioneering work, several attempts have tried to tackle such problems by utilizing image-processing techniques [12] or generative adversarial learning algorithms [15] to selectively remove either markers for visual features or the background scene for tactile features. However, these approaches assign significant computing burdens to processors. More recently, a growing number of works focus on mechanisms that can control the transparency of the skin to activate selectively desired sensing mode. For example, in regards to reflective-based ViTac systems, the authors from [13], [16] have modulated the internal light to shift back and forth between the transparent and semi-transparent state. Nevertheless, they all faced difficulty in finding the proper setting for the illumination system adaptable to various object's colors, approaching angle, and lighting conditions of the environment.

Previously, we introduced a novel vision-based sensing system featuring the whole body with both the sense of touch and the proximity [17]. The key distinguishable attribute lies on the incorporation of PDLC (Polymer Dispersed Liquid Crystal) film within the soft skin. This film is able to actively modify the transparency of the skin between opaque and transparent states, corresponding to tactile and proximity modes, respectively. Although this paradigm is promising, there remain several drawbacks. First, customized PDLD film's price per area unit is still high, roughly $100\,\text{USD/cm}^2$ at the time of purchase. Moreover, it is required a specified control toolbox to trigger the film making this infeasible for small-scale robotic devices.

In this article, we establish the first insight of *Vi2TaP* - a novel, cost-effective, and scalable design scheme for visual-tactile sensors that leverage the polarization phenomenon to actively manipulate the monitoring range of the camera (see Fig. 1). Polarization, in physics, is the process in which unpolarized light, consisting of electromagnetic waves oscillating in multiple planes, is confined to oscillating in a single plane. The polarization can be done by the *Polaroid filters*, or simply polarizers. The polarizer serves as a filter to block incident light waves but those in a specific plane of oscillation, known as the *plane of polarization* (*PoP*). This mechanism further reveals a configuration where a pair of polarizers is placed back-to-back. By orienting one filter at 0 or 90 degrees relative to the other to create $Cross - Polarization$ configuration, we can control whether the polarized beam produced by the first filter passes through the second filter, *i.e.,* the visibility of the outer background to the camera. This paper also showcases the successful adoption of *Vi2TaP* in a soft, multi-modal sensorized gripper, emphasizing the reliability of the perception activation method. Furthermore, it highlights the effectiveness of proximity and tactile perception, particularly in enhancing grasping and manipulation performance. To the best of our knowledge, this is the *first use* of the polarization phenomenon in the development of a soft robotic mechanism.

The primary contributions of this work are as follows:

1) Proposal of a unified multi-modal (proximity and tactile) sensing infrastructure based on polarization phenomenon.
2) Introduction of the first implementation of this system on a soft fingertip, detailing the mechanical design, fabrication process, and methods for proximity and tactile data extraction enabled through monocular depth-map estimation and zero-shot sim2real learning.
3) Evaluation on the significance of visual-tactile fusion perception for enhancing grasping efficiency.

II. Idea, Design and Fabrication

A. Conceptual Paradigm for Vi2TaP

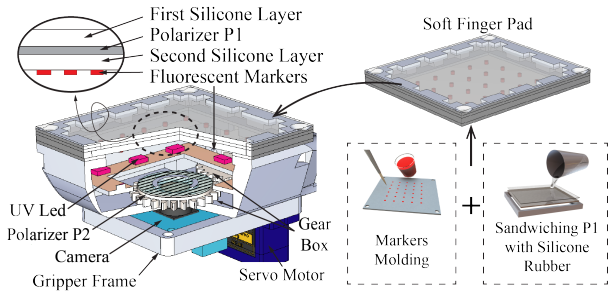
The conceptual scheme of *Vi2TaP* can be reviewed in Fig. 1. The underlying principle lies in the ability to actively manipulate the visibility of the external environment to the internal camera. This function is featured by two parallel-located polarizers, separating the exterior space from the camera. Specifically, the first polarizer (referred to as P_1) is embedded inside a transparent, soft skin with a matrix of markers printed on the inner layer to create a touch interface. Whereas, the second polarizer (P_2) is positioned in front of the camera lens, ensuring all light sources (or objects) within its field of view (FoV), including those polarized by the P_1.

Given this setup, multi-modal sensing modes are discovered under the following conditions:

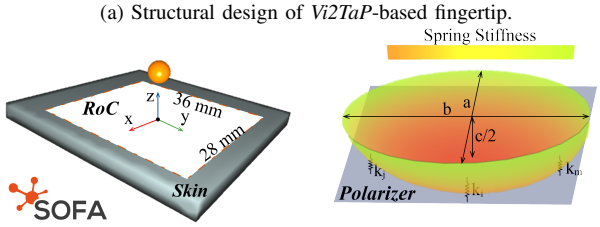
- **Proximity mode**: When the PoPs of P_1 and P_2 are aligned at the same angle ($\theta = 0\,\text{deg}$), the light from the exterior space will be able to travel through P_1 and P_2. This enables the camera to scan objects from the external environment (see Fig. 1), allowing the extraction of proximity information from nearby obstacles.
- **Tactile mode**: When the PoPs are orthogonal to each other ($\theta = 90\,\text{deg}$), the light sources polarized by P_2 will be completely intercepted by P_2. In this setting, the camera is only capable of tracking marker displacement to perceive tactile sensation while the exterior becomes invisible (Fig. 1). The intensity variation of a polarized light I_0 transmitted through another polarizer can be formulated using Malus's law [18]:

$$I = I_0 \cos^2\theta, \tag{1}$$

where I is the resulted light intensity corresponding to the angle θ between P_1 and P_2.



(a) Structural design of *Vi2TaP*-based fingertip.



(b) SOFA simulation. (c) Polarizer modeling in SOFA.

Fig. 2. (a) Structural design of the proposed *Vi2TaP*-based fingertip including the actuation system, illumination system, polarizer films, and the soft skin accompanied by its fabrication procedure. (b) presents the simulation environment in SOFA, in which, the mechanical contribution of the polarizer film is modeled as a series of virtual springs (c).

B. Design and Fabrication

In this section, we introduce the design and fabrication procedure for a soft robotic finger that integrates dual-sensing modalities using the paradigm from Section II-A. The design architecture accommodates a finger body serving as a core to mount three primary modules: the transparent soft skin and illumination system, the actuation system, and the vision system. The detailed breakdown of the proposed finger can be reviewed in Fig. 2a. The following subsections will provide detailed descriptions of the fabrication and assembly of these components, along with critical design considerations.

1) Transparent Soft Skin and Illumination System: First, we utilize Solaris™silicone rubber (Smooth-On, USA) to fabricate a soft, transparent skin with dimensions of 48×43×5 mm (length×width×thickness). This material is chosen for its high transparency and more importantly, notable flexibility which assures a gentle touch even with the presence of the polarizer film. The first polarizer P_1 is sandwiched in between for stable positioning under external stimulation. Markers are distributed in a 9.5×8.5 mm grid pattern creating a 5×4 matrix (see Fig. 2a). Inspired by [19], we color the markers with UV-fluorescent pigment (Silc Pig™, Smooth-on, USA).

Regarding the illumination system, we integrate an array of UVC LEDs inside the body frame to generate Ultraviolet type C (UVC) radiation, which triggers the light-emitting ability of fluorescent-based markers. It's worth noting that UVC light is typically undetectable by standard RGB cameras. Therefore, visual noises are significantly diminished.

The fabrication process is illustrated in Fig. 2a. First, the marker array is created. Each dot is manually placed into a pre-etched hole with the dimension of 1.5 mm×1 mm (radius×depth) on an acrylic plate. This plate is then assembled with a 5 mm-thick frame to form the mold which also serves as the touchpad afterward. To adopt the polarizer film (P_1), it will be gently submerged in the raw silicone rubber and situated at the desired position. After 4 hours, the touchpad is then assembled with the finger frame.

2) Actuation System: The rotation of P_2 is supplied by a 3D-printed gear set with a ratio of 1:1 and a servo motor, as illustrated in Fig. 2a. The driven gear, which carries the filter, is positioned coaxially with the principal axis of the image frame. The negligible load allows us to use a cost-effective servo motor (MG90S). The motor and the camera are affixed to the rear carrier as shown in Fig. 2a.

3) Vision System: The vision system consists of two important elements: camera and polarizer filters. For the camera, the mini USB camera ELP-USB100W07M-MHV120 with the FoV of approximately 120° is chosen. Moreover, we also turned off the auto-focus and auto-light-balancing functionalities to reduce visual noise for both proximity and tactile sensing modes as reported in [11].

In this work, we process commercial polarizer films (approximately 3.5 USD/cm^2 from Selens, USA) with a theoretical polarization ratio up to 99%, into pieces of the desired dimensions using a laser-cutting machine. Specifically, P_1 is a rectangular film with a size of 49 x 44 mm, in length and width, to fully cover the open view of the soft skin. Meanwhile, P_2 is cut into a circular shape (diameter of 20 mm) for installation onto the driven gear.

III. Multi-Modal Perception

A. Tactile Perception

Tactile perception is an indispensable feature for robotic fingers to effectively manipulate the environment. In this paper, we rely on the spatial transition of visual cues (markers) in the tactile mode (*i.e.,* $\theta = 90°$) to infer desired information including contact forces and location during the grasping phase. We build upon and refine the framework presented in [11]. The simulation module generates training datasets extracted from simulated tactile images and corresponding tactile knowledge. Then, the correlation between these two domains is established by training a deep neural network.

1) Elastomer-Polarizer Composite Skin: Modeling, Simulation and Visual Recording: Firstly, to accurately replicate the mechanical behavior of the soft skin with an implanted polarizer, we leveraged *SOFA*[1] - a multi-physics engine based on Finite Element Method (FEM) (Fig. 2b). We adopt an elastic material model (via Young's modulus (E) and Poisson's ratio (ν)) and co-rotational FEM formulation [11] for the characterization of the soft layer. While E is determined as 0.04 N/mm^2 by tensile test, ν is set as 0.49 [20]. The generic dynamic equation for a deformable body is as below [21] :

$$\begin{aligned}\left(\mathbf{M}+dt^2\mathbf{K}+dt\mathbf{C}\right)\ddot{\mathbf{q}} = -dt^2\mathbf{K}\dot{\mathbf{q}}_1 + dt\left(\mathbf{F}^{ext}-\mathbf{F}^{int}\right)\\ + dt\mathbf{J}^T\boldsymbol{\lambda}\,,\end{aligned} \tag{2}$$

where $\mathbf{q} \in \mathbb{R}^{N\times 3}$ is the 3D position of element nodes, $\mathbf{M}$ is total mass of the soft skin. $\mathbf{F}^{\text{ext}}$ denotes the external forces (*e.g,* gravity) at each time step t and $\mathbf{F}^{\text{int}}$ represents internal

[1]Simulation Open Framework Architecture: www.sofa-framework.org

forces upon the system state. $\mathbf{K} = \frac{\partial \mathbf{F}^{\text{int}}}{\partial \mathbf{q}}$ and $\mathbf{C} = \frac{\partial \mathbf{F}^{\text{int}}}{\partial \dot{\mathbf{q}}}$ are stiffness and damping matrices, respectively.

In the context of interaction with the surroundings, SOFA treats all physical contacts as constraints through the term $\mathbf{J}^T\boldsymbol{\lambda}$, governed by contact mechanics based on Signorini's law [21]. The Lagrange multipliers, $\boldsymbol{\lambda}_{\mathcal{Q}}$, where $\mathcal{Q}$ represents the subset of nodes under contact, denote the magnitude of the contact force, while the Jacobian matrix $\mathbf{J}(\mathbf{q})$ gathers the constraint directions projected on $\mathbf{q} \in \mathcal{Q}$. Summing $\mathbf{J}^T\boldsymbol{\lambda}_{\mathcal{Q}}$, hereafter shortened as $\boldsymbol{\lambda}$ yield the estimation of the resultant force vector (magnitude, direction).

On the other hand, we emulate the presence of the polarizer by incorporating a series of *virtual* elastic springs with stiffness k_i, where i is skin nodes within the polarizer plane. These springs act as a stiffening substrate that restrains the deformation of the outer soft layer from its original position. At an equilibrium state t, the virtual springs generate internal forces $\mathbf{f}^{spring}$, which are proportional to the nodal displacement. Given that the reinforcement effect of the polarizer film varies across the soft skin, determining the stiffness coefficient k_i is challenging. Considering the symmetrical geometry, the flat contact surface, and the fact that the stiffness k_i has a descending tendency toward the center, we hypothesize that the distribution of k_i can be represented by the lower half of a spheroid (see Fig. 2c). The general equation is as follows:

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1,\ s.t. -\frac{c}{2} < z < 0, \tag{3}$$

where a, b, and c are half the lengths of the principal axes defining the overall geometry as pictured in Fig. 2c. The stiffness $k_i = z_i$ of the virtual spring attached to any other point $p_i(x_i, y_i)$ can be estimated using the following equation:

$$z = -c\sqrt{1 - \frac{x^2}{a^2} - \frac{y^2}{b^2}}. \tag{4}$$

Based on this equation, it is crucial to seek an appropriate set of coefficients a, b, and c. For simplification, a and b are chosen so that the spheroid neatly overlays the whole contact surface. Next, an empirical study was conducted to find the optimal c so that the simulated contact force $\lambda_{\zeta,sim}$ is close to the actual values. Here, we collected actual forces exerted on four representative points located in a different quadrant of the skin. Then, c can be estimated as follows:

$$c = \arg\min_{c} \mathcal{L}_{\text{MSE}}(\lambda_{\zeta,\text{sim}}, \lambda_{\zeta,\text{real}}). \tag{5}$$

2) Training Data Collection: This section addresses the data collecting procedure for the image of the inner skin surface correlating with tactile information extracted from the simulation model described above. First, the SOFA simulation scene emulates the indentation of a spherical pointer on the soft skin portraying the gripping action. At each pushing position $\mathcal{P} = [\mathcal{P}_x, \mathcal{P}_y]$ [mm], the indentor is driven perpendicularly toward the top surface with a constant speed until the contact depth reaches 3 mm. This limitation is experimentally determined based on the deformation constraint at the edge region of the soft skin, which equivalently defines the sensor's operational range. At each 0.2 mm step, the relevant tactile data, including the contact force vector $\lambda = [\lambda_x, \lambda_y, \lambda_z]$ [N] and the nodal position $\mathbf{X} \in \mathbb{R}^{N\times 3}$ will be recorded. Skin deformation $\mathbf{D}_{\text{gt}}$ is calculated as follow:

$$\mathbf{D}_{\text{gt},i} \coloneqq \mathbf{X}_i - \mathbf{X}_{0,i}\ ,\ \forall i \in \mathcal{N} = \{1, 2, \cdots, N\}, \tag{6}$$

where $\mathbf{X}_0 \coloneqq [\mathbf{X}_{0,i} \in \mathbb{R}^3\ |, \mathbf{X}_{0,i} = \mathbf{q}_i(0), \forall i \in \mathcal{N}]$ are initial state of each node. For stable grasping, it is envisioned that the desired contact area is within the central region of the skin. Hence, the tactile data acquisition procedure is implemented on a subset of nodes (176 nodes) within the Region-of-Contact (RoC) as seen in Fig. 2b.

Next, we employed Unity to render simulated images $\mathbf{I}_{\text{sim}}$ (a resolution of 640×480 pixels) based on deformation inputs transferred from SOFA environment. The procedure can be reviewed in detail from [11]. Furthermore, while [11] used a subset of real images to address the sim2real gap, this paper employs the domain randomization technique to enable zero-shot learning of the tactile model based solely on simulation images. Specifically, the procedure involves performing affine transformations during the training process to diversify the perspective of tactile binary images via translation, rotation, and scaling. The database includes 2640 samples, in which 80% ($K = 2112$ samples) is used for the training process and the remainder is for validation.

3) Training: The training process is performed on the dataset $\mathcal{D} = \{\mathbf{I}^k_{\text{sim}}, \mathbf{D}^k_{\text{gt}}\}^K_{k=1}$, collected in both Unity and SOFA environments. However, rather than directly using the tactile images as input for the TacNet model, we leverage the vector displacements of the markers as the visual input representation, which proves more effective for the simulation-to-real process. The displacements of the markers are computed as $\Delta\mathbf{U}^k = \mathbf{U}^k - \mathbf{U}_0 \in \mathbb{R}^{m\times 2}$, where m is the number of markers, and $\mathbf{U}^k$ and $\mathbf{U}_0$ represent the positions of the tracked markers in the instance k and at the rest (no-contact) state, respectively. To detect the marker array in a single image $\mathbf{I}^k_{\text{sim}}$, we first convert the image to binary, then apply Canny edge detection to identify the round markers, locating the central positions of all markers in the coordinates of pixels. Given the TacNet model $\mathcal{T}_\varphi$ parameterized by $\boldsymbol{\varphi}$, the training loss function $\mathcal{L}_{\text{MSE}}(\cdot)$ is computed as:

$$\mathcal{L}_{\text{MSE}}(\mathbf{D}^k_{\text{gt}}, \mathbf{D}^k_{\text{est}}) = \frac{1}{3n}\sum_{i\in\mathcal{N}}\sum_{j\in\{x,y,z\}}(d^j_{\text{gt},i} - d^j_{\text{est},i})^2, \tag{7}$$

where $\mathbf{D}^k_{\text{est}} = \mathcal{T}_\varphi[\Delta\mathbf{U}^k]$, and $d^j_i\ ,\ \forall j \in \{x, y, z\}$ are the components of displacement vector $\mathbf{D}^k_{:,i} = [d^x_i, d^y_i, d^z_i]$ at the respective skin node $i \in \mathcal{N}$ along the x, y and z axes. For optimizing $\boldsymbol{\varphi}$, we use Stochastic Gradient Descent (SGD) with an experimentally tuned learning rate of 0.015 over 40 epochs. Additionally, unlike the network proposed in [11], we employ a simple chain of three fully connected (MLP) layers, making it well-suited for handling small, structured inputs and ideal for small-scale applications thanks to low computational cost. The input layer is adapted with two input channels representing the marker displacements $\Delta\mathbf{U}^k$ in the x- and y-directions, respectively. The model is trained by using a desktop PC (AMD Ryzen™ Threadripper™ 3970X Processor) with GPU acceleration (RTX 8000, NVIDIA).

B. Proximity Perception

1) Data Preparation and Training: Self-consciousness about the object before approaching contact is necessary to enable safe and optimized grasping strategy. Particularly, we employ a data-driven monocular depth estimation based on a DNN [22] to generate a depth map that records the presence of obstacles surrounding the touchpad and serves as the basis for distance measurement. Here, we fine-tune the proven MiDas model [22] on *MannequinChallenge* dataset [23]. The image dataset was, at first, synthetically augmented using the alpha blending technique to replicate the see-through views. The monocular depth estimation network (DepthNet) is trained to regress the augmented images to corresponding depth images $\mathbf{Z}^{\text{gt}}$ generated by the MVS pipeline proposed in [24]; which shows as ground-truths for model training.

For the loss function, we adopt a scale-invariant depth regression loss, as proposed in [22]. Given a raw estimated and ground-truth depth map $\mathbf{Z}^{\text{est}}, \mathbf{Z}^{\text{gt}} \in \mathbb{R}^{a\times b}$, the scale-invariant regression loss can be derived as:

$$\mathcal{L}_{\text{DepthNet}} = \mathcal{L}_{\text{ssitrim}}(\mathbf{Z}^{\text{est}}, \mathbf{Z}^{\text{gt}}) + \alpha\mathcal{L}_{\text{grad}}(\mathbf{Z}^{\text{est}}, \mathbf{Z}^{\text{gt}}), \quad (8)$$

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.

For training DepthNet model, we initialize the DepthNet based on the ResNet multi-scale architecture [25] with the model weights as mentioned in [22]. For the fine-tuning process, we use Adam optimizer with the learning rate initialized at 10^{-4}, then linearly decaying at the 50^{th} iteration out of a total of 100 training steps. The hyperparameter α in the combined loss function (8) is experimentally set to 0.1. Detailed network architecture can be found in [25].

2) Distance Estimation: This section presents a procedure to estimate the distance between the object and the top surface of *Vi2TaP* using the depth map $\mathbf{Z}^{\text{est}}$ produced by DepthNet model. With the assumption that the closer object would exhibit a larger brightness value in the grey scale, the target object will be recognized by masking other obstacles using binary thresholding. However, since the pixel value substantially varies as the background or ambient light condition changes, it is challenging to solely infer the distance based on the brightness value. Alternatively, we make use of a more intuitive metric based on object's pixel area $\mathcal{A}$ detected in the depth map. This metric is based on the observation that an object's area increases as it gets closer to the touch skin. Dimensionless metric r can be calculated as:

$$r = \mathcal{A}\frac{\mathcal{A}_0}{\mathcal{A}'_0}, \quad (9)$$

where $\mathcal{A}$ represents the pixel area when the obstacle is initially detected and $\mathcal{A}'_0$ is a constant representing for initial area of the reference object. Equation 9 actually calibrates the pixel area of an arbitrary object by multiplying this value with the ratio of the initial area, $\mathcal{A}_0$, to that of the reference object $\mathcal{A}'_0$. This approach is particularly effective with gripper applications thanks to the fact that the object is typically centered between the jaws with constant stroke before the gripping action. Hence, the initial pixel area $\mathcal{A}_0$ is expected to be determined at the same distance. The evaluation of the metric is presented in the next section.

TABLE I
LIST OF EXAMINED POINTS.

Index	1	2	3	4	5	6	7	8	9	10	11	12	13
x [mm]	0	0	0	0	5	5	5	5	10	10	10	10	15
y [mm]	0	5	10	15	0	5	10	15	0	5	10	15	17

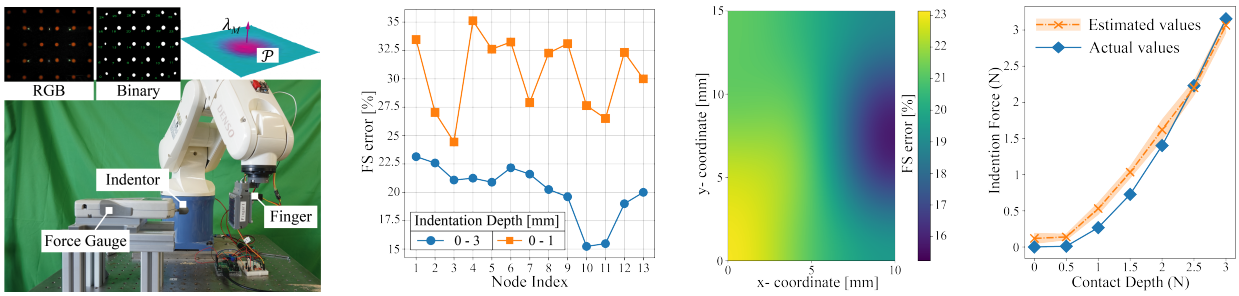
IV. EXPERIMENT RESULTS AND DISCUSSION

In this section, we evaluate the sensing performance of each mode as an individual ability. Then, a strategy to harmonically incorporate both perception fields for enhancing the grasping proficiency will be validated via a practical scenario. Here, *Vi2TaP* was implemented with the desktop PC (Intel(R) i5-9600K 3.70GHz, 32G RAM, NVIDIA GeForce GTX 1050).

A. Tactile sensing mode

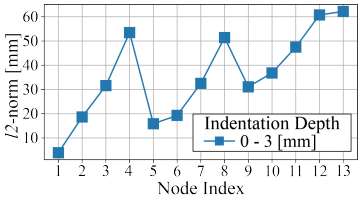
1) Contact force estimation: Tactile sensing ability will be evaluated through the estimation accuracy for contact location and contact force on different regions of the tactile skin. The evaluation session starts with a series of indentation attempts vertically conducted on the predefined locations (see Table I), in which Nodes 1 to 12 are within the RoC (see Fig. 2b), and Node 13 is outside the RoC. Due to the symmetrical and flat touch surface, indention locations are selected within a quarter of the RoC region. The experiment setup can be seen in Fig. 3a, where the robot arm (Denso VP-6242M) will carry the finger and drive toward the indentor with a constant speed (10 mm/s), until reaching the desired indentation depth of 3 mm, with sensor feedback collected at 0.5 mm intervals. During the indention, estimated normal contact force λ_z and pressing position $\mathcal{P}$ are recorded for post-processing. Note that only λ_z is comparable to the measured value due to its absolute dominance over other components. This process is repeated five times and the recorded data is synchronized with actual values from a Force Gauge (Imada ZTA-5N) and filtered out outliers before in-depth analysis. In this mode, the system achieves a sampling rate of 25 Hz when using 30 FPS cameras.

The contact force estimation results are summarized in Fig. 3b-3d. Figure 3b presents the average errors in estimated contact forces in all examined nodes. At first glance, the average error rate appears to be below 23%. However, it is important to note that inaccuracies during the initial indentation phase (0-1 mm) contribute significantly to these values as seen in Fig. 3b. A similar observation has already been reported in [11]. This phase is considered a threshold above which the sensing performance stabilizes. Moreover, the results indicate that the sensor is highly sensitive to the escalation of external stimulation. Additionally, the estimation accuracy varies with the contact location. Figure 3c provides strong evidence for this observation by showing the error distribution (Full-Scale (FS) error) across the regions tested. Interestingly, the accuracy

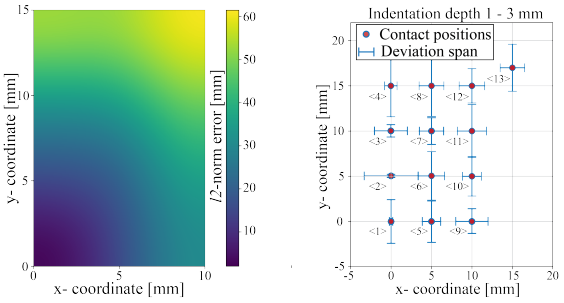


(a) Experiment setup. (b) Force estimation error. (c) Error distribution across the skin. (d) A force measurement sample at Node 11.

Fig. 3. Contact force estimation evaluation. (a) Experiment setup for tactile sensing assessment. Figure (b) shows the force estimation error rate measured at 12 contact nodes plus 1 out-bound node, repeatedly five times with the indention depth up to 3 mm. Figure (d) reports the error distribution (Full-scale (FS) errors) over a quarter of the tactile skin. (c) A data sample of Node 11 demonstrates the good matching between estimated values and actual values, thus validating the reliability of the tactile model in force sensing.



(a) Accuracy of contact location localization.



(b) Error distribution across the skin. (c) Deviation of contact location in x and y coordinates.

Fig. 4. Contact location localization evaluation. (a) Comparison of contact localization accuracy (l_2-norm accumulated in five trials) for the whole range of indentation depth. (b) The spectrum illustrates the overall precision across a quarter of the tactile skin. (c) Deviation in x and y coordinates for the case of 1-3 mm indentation depth. The fact that all deviation bars are within the contact region verifies the reliability of the tactile model.

in the central area is slightly lower than that of the edge regions, though the difference is not substantial. This variation can be attributed to the inaccuracy of the skin simulation model. Despite this limitation, the precision of this sensing ability is still validated via Fig. 3d which shows good matching between estimated force values and actual values at Node 11.

2) Contact location estimation: The ability to localize the contact region will be assessed via the l_2-norm metric $\|\mathbf{e}\|_2$. The equation for calculating $\|\mathbf{e}\|_2$ is shown below:

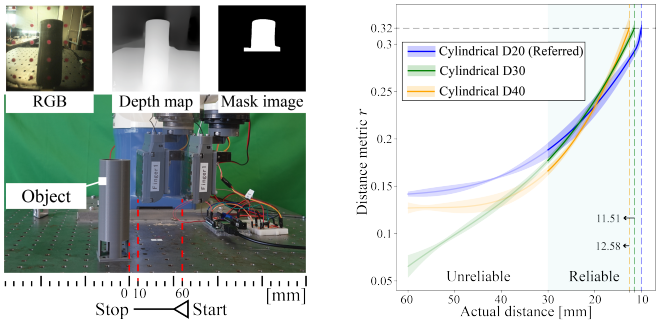
$$\|\mathbf{e}\|_2 = \sqrt{\sum_{i=1}^{n}((\mathcal{P}^i_{x,est} - \mathcal{P}^i_{x,true})^2 + (\mathcal{P}^i_{y,est} - \mathcal{P}^i_{y,true})^2)}, \tag{10}$$

where subscripts $_{est}$ and $_{true}$ are presenting for estimated and true values of the contact location, while n is the number of trials. Equation 10 quantifies the deviation of the estimated values from the actual values and sums the error across all trials. This approach not only assesses the sensor performance but also validates its stability.

Figure 4a presents the total estimation error across the examined points. The plot indicates that sensor precision decreases when contact occurs near the edges, compared to areas closer to the center. This observation is further validated by the error distribution map shown in Fig. 4b. The potential explanation lies in the nature of the perception inference model which is based on marker movement and the fact that the skin is vertically aligned with the camera axis. Specifically, spatial markers movements, when contacts occur at the edge regions, are most likely manifested less significantly than in the central. Furthermore, according to Fig. 4c which denotes the deviation domain in x and y axes, we can claim that the estimated contact location will not fall outside the RoC region, *i.e.,* irregular values or outliers. This statement in combination with the force estimation results completes the verification of the reliability of the tactile sensing modality of the sensorized soft finger based on *Vi2TaP*.

3) Proximity sensing mode: This section evaluates the performance of the *Vi2TaP* finger's proximity sensing modality, particularly in estimating the distance between an object and the soft skin. The results presented here will demonstrate the feasibility of extracting the metric r from the depth map $\mathbf{Z}^{\text{est}}$ as a reliable substitute for the distance measurement.

The experiment setup is illustrated in Fig. 5a. Similarly, the *Vi2TaP* finger will be driven toward an object by the Denso robot arm with constant speed. Initially, the finger pad will be positioned 60 mm away from the object. This gap is expected to go down to 10 mm at the end of the travel. During this motion, the metric r will be computed and matched to the actual distance derived from the robot arm's movement. This procedure will be repeated five times for each trial to assess the repeatability of the proximity mode. According to Eq. 9, r is calculated w.r.t the reference object. In this test, we allocated a cylinder with a diameter of 20 mm ($D20$) as the reference. The evaluation was conducted with two other cylinders with diameters of 30 mm and 40 mm, so-called $D30$ and $D40$,



(a) Experiment setup. (b) Distance r estimation results.

Fig. 5. Proximity sensing evaluation. (a) Presents the experiment setup for distance measurement. (b) Report the effectiveness of distance estimation via metric r for two different-sized objects (D30 and D40) w.r.t the reference one (D20). Generally, the result reveals that the proposed algorithm executes accurate estimation within the reliable region (from 35 mm downward).

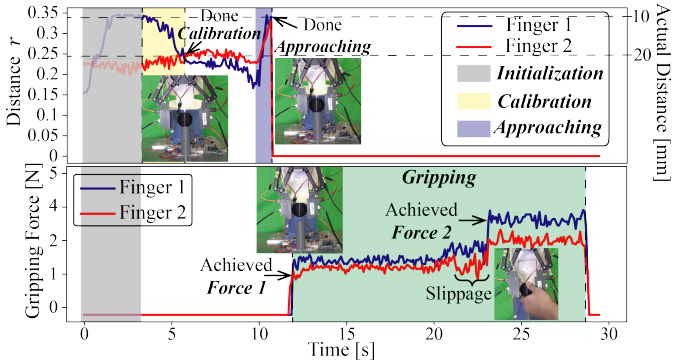


Fig. 6. Gripping demonstration includes several stages: ***Calibration*** - utilizing proximity feedback to relocalize the gripper jaws so that the object is in the middle, corresponding to the approximate distance of 20 mm. ***Approaching*** - driving two fingertips toward the object until the distance is within 10 mm and switch to Tactile mode. ***Gripping*** - gripping the object with contact forces *Force1* and *Force2* for lifting and preventing slippage, respectively.

respectively. The sampling rate of the system in this mode is approximately 22 Hz.

Figure 5b reports the variation of r for three tested objects that the arm relies on to stop the motion. Notably, via the metric r, the *Vi2TaP* finger accurately estimates the end-distance value of 10 mm, with only a small error (11.51 mm and 12.58 mm for $D30$ and $D40$, respectively). This highlights the utility of the metric r in determining the optimal point for transitioning between vision and tactile perception domains, which is crucial for our device's performance. From another perspective, as the finger moves, the estimated distance r is only reliable within the range from 30 onward. Outside this range, significant deviations of r are observed for both cases, $D30$ and $D40$, compared to that of the reference object. Nevertheless, given that the interaction space for most gripper devices is typically small and fixed, this detection range remains practical and effective. This finding also suggests the proper point where the initial area $\mathcal{A}_0$ is obtained.

B. Gripping demonstration

This section elaborates on how a multi-modal visual-tactile sensing scheme could benefit the grasping performance of a robotic gripper, particularly in object localization and in-hand stabilization. In this paper, we draw inspiration from the challenges inherent in parallel grippers. These grippers typically have each finger connected to a synchronized system driven by a common actuator, enabling concurrent motion. Although uniform and coordinated movement of all fingers allows for a secure and stable grasp, it requires the gripper to accurately position itself so that the object aligns with the central axis. We envisage that the seeing-through view of the *Vi2TaP*-based sensor (*i.e.,* $\theta = 90°$) can give a hint on the relative position of the target w.r.t each finger. This experiment use Robotiq Gripper 2F-140 with the jaw stroke after *Vi2TaP*-based fingertips installment is 90 mm. A ball-shaped object with a diameter of 50 mm was the target. Also, this ball was flattened creating irregular portrails to the camera to prove the generalization of the proximity sensing capability.

The demonstration is summarized in Figure 6, showcasing key snapshots and sensor feedback from both tactile (contact force λ_z) and proximity sensing modalities (distance metric r) in response to various actions. The experiment begins with the object positioned randomly between the two fingers, though off-center; specifically, it is closer to Finger 1, meaning $r_1 < r_2$. Then, the proximity modality is used to relocalize the gripper position so that the object is exactly middle, *i.e,* $r_1 \approx r_2$ as seen in Fig. 6, so-called ***Calibration*** phase. Next, the fingers begin the ***Approaching*** phase, moving toward the object until they are within a 10 mm distance to the object. At this point, the tactile sensing mode is activated in both fingers for the ***Gripping*** phase. The gripping action continues until both fingers endure contact forces of 1 N (referred to as $Force$ 1), allowing the gripper to securely lift the object. In the final test, the object is manually pulled downward to simulate slippage. To respond to this condition, the system monitors changes in the contact location as an indicator to detect potential slippage and apply additional grip force (referred to as $Force$ 2) to secure the object. The whole demonstration can be reviewed in the Supplementary Video.

As observed in the Supplementary Video and Fig. 6, the gripper successfully executed the task thanks to the aid of both tactile and proximity sensing modalities. However, several issues remain for consideration. During the calibration phase, the r values for both fingers converge at 0.238, while the actual distances from the object to Finger 1 and Finger 2 are 21.27 mm and 18.73 mm, respectively. This outcome highlights two key points: First, the calibration result is within an acceptable range, with only a small deviation from the expected 20 mm; Second, each finger's performance is highly sensitive to background conditions, leading to unavoidable inconsistencies in r estimation, particularly in gripper applications. For tasks involving tactile sensing, a discrepancy is observed in the estimated force between the two fingers, which becomes more pronounced as the grip force increases. This deviation can likely be attributed to the limited generalizability of the tactile perception model, which relies on an affine transformation procedure mentioned in Section III-A2. To address this issue, more extensive randomization could be applied to improve the model's robustness.

V. Conclusion

This article presents a unified paradigm for multi-modal visual-tactile soft sensors, featuring an innovative mechanism that actively shifts back and forth between vision and tactile perceptive fields. *Vi2TaP* mechanism offers several advantages including simple design architecture, reasonable cost, easy-to-fabrication and most importantly, complete separation of the visual and tactile perception field. Also, the first showcase of *Vi2TaP*, a soft sensorized gripper, has promised a broad range of other applications in the soft robotic field.

Beyond the contributions discussed above, several technical challenges remain. First, although the current mode-switching mechanism is mechanically simple, its switching rate requires improvement. The system takes approximately 0.5-0.6 seconds to fully transition between modes, with the primary bottleneck being the serial communication between the central computer (responsible for processing sensory data) and the Arduino microcontroller (which controls the motor). To ensure a seamless transition between two sensing modes, adopting a more powerful microcontroller and a faster communication protocol, such as I2C, should be considered. The second challenge concerns the precision of the mechanical characterization of the composite soft layer, where the polarizer plays a crucial role. In this work, its mechanical contribution is approximated using a series of virtual springs distributed across the skin. While this approach shows promise, further refinement is needed. Specifically, the determination of the film stiffness coefficient z_i (see Eq. 4) currently relies on force data obtained from only four locations near the skin's edges. Improving stiffness distribution accuracy requires incorporating additional reference contact points broadly distributed across the entire surface. Moreover, a more robust optimization algorithm is necessary to determine the optimal set of z_i. Finally, the presented demonstration is limited to a pick-and-place task, suitable for clustered environments where the object position is unknown in advance. To fully showcase the advantages of the multimodal visuotactile sensory system in soft grippers, future studies should explore more complex scenarios, including object recognition, path planning, and dexterous manipulation.

With significant potential for further development, *Vi2TaP* is expected to initialize a new wave of soft robotic devices at various scales with multi-modal perceptual modes for challenging sensing and control tasks.

References

[1] S. Q. Liu, Y. Ma, and E. H. Adelson, "Gelsight baby fin ray: A compact, compliant, flexible finger with high-resolution tactile sensing," in *2023 IEEE International Conference on Soft Robotics (RoboSoft)*, 2023, pp. 1–8.

[2] R. Gupta, S. Tanwar, S. Tyagi, and N. Kumar, "Tactile-internet-based telesurgery system for healthcare 4.0: An architecture, research challenges, and future directions," *IEEE Network*, vol. 33, no. 6, pp. 22–29, 2019.

[3] W. Mandil, V. Rajendran, K. Nazari, and A. Ghalamzan-Esfahani, "Tactile-sensing technologies: Trends, challenges and outlook in agri-food manipulation," *Sensors*, vol. 23, no. 17, 2023.

[4] 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*, vol. 38, no. 3, pp. 1599–1620, 2022.

[5] Z. Wang, H. Gao, A. Schmitz, S. Somlor, T. P. Tomo, and S. Sugano, ""safe skin" - a low-cost capacitive proximity-force-fusion sensor for safety in robots," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 807–813.

[6] J. Yin, G. M. Campbell, J. Pikul, and M. Yim, "Multimodal proximity and visuotactile sensing with a selectively transmissive soft membrane," in *2022 IEEE 5th International Conference on Soft Robotics (RoboSoft)*, 2022, pp. 802–808.

[7] F. Giovinazzo, F. Grella, M. Sartore, M. Adami, R. Galletti, and G. Cannata, "From cyskin to proxyskin: Design, implementation and testing of a multi-modal robotic skin for human–robot interaction," *Sensors*, vol. 24, no. 4, 2024.

[8] C. Mu, Y. Wang, D. Mei, and S. Wang, "Development of robotic hand tactile sensing system for distributed contact force sensing in robotic dexterous multimodal grasping," *International Journal of Intelligent Robotics and Applications*, vol. 6, no. 4, pp. 760–772, 2022.

[9] J.-C. Sicotte-Brisson, A. Bernier, J. Kwiatkowski, and V. Duchaine, "Capacitive tactile sensor using mutual capacitance sensing method for increased resolution," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 10 788–10 794.

[10] T. Yamamoto, N. Wettels, J. A. Fishel, C.-H. Lin, and G. E. Loeb, "Biotac - biomimetic multi-modal tactile sensor," *Journal of the Robotics Society of Japan*, vol. 30, no. 5, pp. 496–498, 2012.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] W. Fan, H. Li, W. Si, S. Luo, N. Lepora, and D. Zhang, "Vitactip: Design and verification of a novel biomimetic physical vision-tactile fusion sensor," 2024. [Online]. Available: https://arxiv.org/abs/2402.00199

[16] F. R. Hogan, M. Jenkin, S. Rezaei-Shoshtari, Y. Girdhar, D. Meger, and G. Dudek, "Seeing through your skin: Recognizing objects with a novel visuotactile sensor," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 1217–1226.

[17] 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.

[18] *A Dictionary of Physics*. Oxford University Press, 2009. [Online]. Available: https://www.oxfordreference.com/view/10.1093/acref/9780199233991.001.0001/acref-9780199233991

[19] C. Sferrazza and R. D'Andrea, "Design, motivation and evaluation of a full-resolution optical tactile sensor," *Sensors*, vol. 19, no. 4, 2019.

[20] N. H. Nguyen and V. A. Ho, "Mechanics and morphological compensation strategy for trimmed soft whisker sensor," *Soft Robotics*, vol. 9, no. 1, pp. 135–153, 2022.

[21] 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.

[22] 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.

[23] Z. Li, T. Dekel, F. Cole, R. Tucker, N. Snavely, C. Liu, and W. T. Freeman, "Mannequinchallenge: Learning the depths of moving people by watching frozen people," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4229–4241, 2021.

[24] 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.

[25] 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.