Autoencoding a Soft Touch to Learn Grasping from On-Land to Underwater

Ning Guo, Xudong Han, Xiaobo Liu, Shuqiao Zhong, Zhiyuan Zhou, Jian Lin, Jiansheng Dai, Fang Wan, and Chaoyang Song**

Robots play a critical role as the physical agent of human operators in exploring the ocean. However, it remains challenging to grasp objects reliably while fully submerging under a highly pressurized aquatic environment with little visible light, mainly due to the fluidic interference on the tactile mechanics between the finger and object surfaces. This study investigates the transferability of grasping knowledge from on-land to underwater via a vision-based soft robotic finger that learns 6D forces and torques (FT) using a supervised variational autoencoder (SVAE). A high-framerate camera captures the whole-body deformations while a soft robotic finger interacts with physical objects on-land and underwater. Results show that the trained SVAE model learns a series of latent representations of the soft mechanics transferable from land to water, presenting a superior adaptation to the changing environments against commercial FT sensors. Soft, delicate, and reactive grasping enabled by tactile intelligence enhances the gripper's underwater interaction with improved reliability and robustness at a much-reduced cost, paving the path for learning-based intelligent grasping to support fundamental scientific discoveries in environmental and ocean research.

1. Introduction

Collecting delicate deep-sea specimens of geological or biological interests with robotic grippers and tools is central to supporting fundamental research and scientific discoveries in environmental and ocean research.[1,2] The human fingers are dexterous in object manipulation thanks to the finger's musculoskeletal biomechanics and skin's tactile perception even in harsh environments such as underwater.[3,4] Much research has been devoted to skilled object manipulation in daily life scenarios.[5] However, limited research focuses on transferring such capabilities to an underwater environment.[6] The ambient environment significantly challenges visual and tactile feedback integration while performing object grasping for visual identification under fluidic interference on the surface of physical interaction.[7,8] As a challenging task for humans, designing and developing robotic solutions for reactive and reliable grasping becomes even more complicated when the end-effector is fully submerged underwater.[9]

1.1. Design toward Soft Grasping for Ocean Exploration

Object grasping is essential for environmental and ocean research to collect in situ specimens, where a trend toward softness in gripper design shows a growing adoption over the years.[10] Classical research on underwater grasping mainly focused on a direct translation of mechanical grippers made from rigid materials with waterproof design for all components, including the actuators, mechanisms, and sensors, resulting in a bulky design that is usually difficult for system integration.[11] Previous research reports a modular continuum finger for dexterous subsea manipulation with force and slip sensing, where the complex integration of a range of mechanical, electrical, and

N. Guo, X. Han, X. Liu
Department of Mechanical and Energy Engineering
Southern University of Science and Technology
Shenzhen 518055, China

S. Zhong, Z. Zhou, J. Lin
Department of Ocean Science and Engineering
Southern University of Science and Technology
Shenzhen 518055, China

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/aisy.202300382.

DOI: 10.1002/aisy.202300382

J. Dai, F. Wan
Shenzhen Key Laboratory of Intelligent Robotics and Flexible Manufacturing
Southern University of Science and Technology
Shenzhen 518055, China

F. Wan
School of Design
Southern University of Science and Technology
Shenzhen, Guangdong 518055, China
E-mail: wanf@sustech.edu.cn

C. Song
Guangdong Provincial Key Laboratory of Human-Augmentation and Rehabilitation Robotics in Universities
Southern University of Science and Technology
Shenzhen, Guangdong 518055, China
E-mail: songcy@ieee.org

computing subsystems limits the use of this prototype out of a laboratory testing tank.[12] Another submarine gripper was developed as part of the European project TRIDENT,[13] demonstrating dexterity for executing grasping and manipulation activities, but suffers from challenges when interacting with the delicate subsea environment and objects.

Recent development in soft robotics adopts a different approach to leverage material softness for grasping.[14] The advantage of soft grippers for underwater scenarios is a systematic integration of fluidic actuation, motion transmission, and form-closed adaptation enabled by the soft, lightweight, low-cost material and fabrication against an aquatic environment with reduced complexity using simple open-loop control.[2] These soft grippers demonstrated successful, compliant interaction with various objects underwater.[15] A recent review shows an emerging research gap in introducing sensory capabilities to soft grippers underwater for closed-loop grasping feedback.[16]

1.2. The Need for Vision-Based Tactile Grasping Underwater

Inspired by the tactile perception of human fingers, a wide range of robotic research has been devoted to integrating with object grasping in industrial or daily life settings.[17] Current research on tactile perception often leverages material softness for skin-like design.[4] Recent work in 3D tactile tensegrity has expanded the adoption of tactile sensing to the underwater environment, presenting a promising direction through the integration of soft self-powered triboelectric nanogenerators and deep learning-assisted data analytics for underwater exploration.[18] While recent work gave an exhaustive investigation of tactile sensors and their applications in intelligent systems,[19] there is also an emerging field of vision-based tactile sensors in robotics under-represented in this field.[20]

Vision-based tactile perception leverages machine vision to provide multimodal contact information with detailed spatial resolution.[21] The focus is to deploy soft media that deform under external forces and infer tactile information from visual observation.[22] Sato et al.[23] built a linear approximation model to estimate the contact forces by tracking two colored spherical markers arranged at different depths of an elastomer surface. Yamaguchi et al.[24] presented another low-order approximation model to infer the contact forces from observed makers' variations in the camera. Unfortunately, the current literature has not yet explored the adoption of vision-based tactile robotics in the underwater scenario.

1.3. Machine Learning for Latent Intelligence in Tactile Robotics

The performance of machine learning algorithms heavily depends on the choice of data representation.[25] When projecting complex soft robotics deformation into image space,[26] a growing trend of research is devoted to treating the representation of a captured image as the latent variables of an appropriate generative model.[27] The generative models are usually highly interpretable in understanding the causal relations of the observations,[28] making it a potential solution to increase the robustness of vision-based, soft, tactile sensing underwater where the environmental uncertainties are much worse than the daily life or industrial settings, as shown in **Figure 1**.

Variational autoencoder (VAE) recently emerged as a powerful generative model that learns the distribution of latent variables and is widely used for visual representation in robot

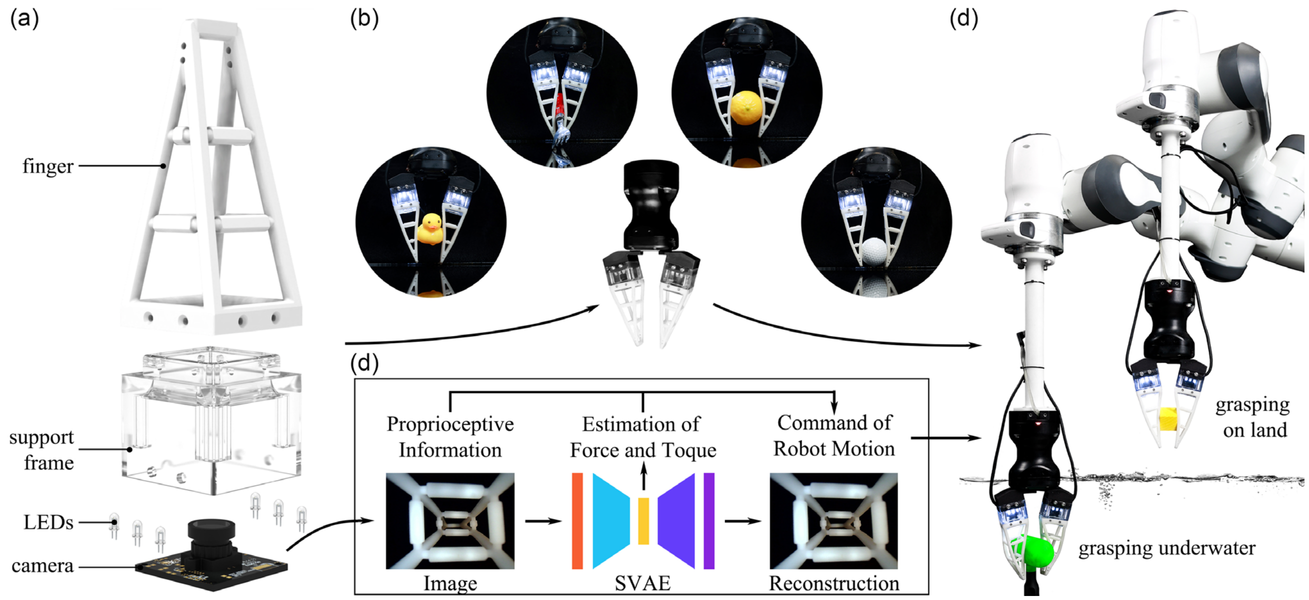

Figure 1. Overview of the soft visual-tactile learning across land and water using SVAE. a) Design of the sensorized soft finger where the camera board is sealed with a silicon layer. b) The integrated amphibious gripper is transformed by replacing the fingertip of a Robotiq Hand-E gripper with the sensorized soft finger with omnidirectional grasping adaptations. The Hand-E gripper has an ingress protection (IP) rating of IP67, which is suitable for our underwater experiment in a tank without extra waterproofing. c) The scheme of visual–tactile learning takes an image of the deformed metamaterial as input, reconstructs the image, and simultaneously predicts the force and torque. d) The amphibious gripper is mounted on a Franka Emika Panda robot arm to execute force control tasks on land and underwater.

learning.[28–30] Since the original publication, many variants and extensions of VAE have been proposed. Semisupervised variational autoencoder (SVAE) was proposed to address the problem of unlabeled data training.[31] Higgins et al.[32] introduced weight to balance the reconstruction error and regularization of latent variable distribution, enabling learning of a disentangled latent representation. The recent adoption of a SVAE model for identifying critical underlying factors for prediction demonstrates the promising potential for application in robotic grasping.[33]

This study investigates the transferability of grasping knowledge from on-land to underwater via a vision-based soft robotic finger that learns 6D forces and torques (FT) using SVAE. Using real-time images collected from an in-finger camera that captures the soft finger's whole-body deformations while interacting with physical objects on-land and underwater, we established a learning-based approach to introduce tactile intelligence for soft, delicate, and reactive grasping underwater, making it a promising solution to support scientific discoveries in interdisciplinary research.

2. Results

2.1. In-Finger Vision for a Soft Tactile Finger

Here we present the in-finger vision design for tactile sensing compatible with both on-land and underwater scenarios, as shown in Figure 1a. The finger is based on a soft metamaterial with a shrinking cross-sectional geometry toward the tip, capable of omnidirectional adaptation on the finger surface to unknown object geometries, enabling a passive form closure for robotic grasping.[34] A monocular RGB camera (120 frames per second) is mounted inside a support frame under the finger to obtain high-framerate images of the finger's adaptive deformations at a resolution of 640 × 480 pixels. The support frame is 3D printed with the optically transparent material (Somos WaterShed XC 11122). All electronics inside are waterproofed by dipping the camera board, except the lens, into transparent silicon. We added six light-emitting diodes (LEDs) to the camera board for improved lighting conditions, resulting in an integrated design of a water-resistant, soft robotic finger with machine vision from the inside.

Figure 1b shows the integration of the proposed finger with a Robotiq's Hand-E gripper, which has an ingress protection rating of IP67 for testing in lab tanks. The proposed soft finger exhibits spatial adaptive deformations, conforming to the object's geometry during physical contact and exhibiting both regular and twisted adaptions for enhanced robustness for grasping, as shown in Figure 1d. For more intensive use in the field, one can directly mount the soft finger to the tip of existing grippers on an underwater vehicle. We demonstrated the effeteness of using the soft finger by grasping some Yale–CMU–Berkeley objects of various shapes and softness underwater or floating on the water surface.[35] See Movie S1 in the Supplementary Materials for further details.

The advantage of the proposed design is a complete separation of the sensory electronics from the soft interaction medium by design, resolving the issues of an enclosed chamber that may suffer from severe surface deformation when used underwater.[20] Such design enables us to collect real-time image streams of the physical interaction between the soft finger and external object using the in-finger vision, as shown in Figure 1c, which can be further implemented with generative models, such as the SVAE, to provide the tactile perception of grasping interactions, both on-land and underwater.

2.2. Generative Tactile Learning via Supervised Variational Autoencoder

Here presents a generative learning architecture for tactile perception in both on-land and underwater scenarios with latent explanations using a SVAE in **Figure 2**.

The generative model is illustrated in Figure 2a, which includes an encoder $q_\phi(\boldsymbol{Z}|\boldsymbol{X})$ to process real-time images from the in-finger vision, then processed through a latent space operation to estimate latent distribution of the interactive physics $\boldsymbol{Z} \sim \boldsymbol{N}(\boldsymbol{Z}_\mu|\boldsymbol{X}_\theta)$, assuming a normal distribution. Here, we added a force and torque prediction based on $\boldsymbol{Z}_\mu$ to produce the 6D tactile estimation as an auxiliary output. Finally, through a generative decoder $p_\theta(\boldsymbol{Z}|\boldsymbol{X})$, our model reproduces images of the tactile interactions based on the learned SVAE model. Note that θ and ϕ are the parameters of the encoder and decoder neural networks, which must be optimized during training. Note that the SVAE model's loss function for training is the combination of image reconstruction loss, force/torque prediction loss, and latent representation regularization loss. Following the detailed formulation of the SVAE model in the Experimental Section, for tuning with small datasets, we introduced two hyperparameters, α and β, to modify the objective function into the following, where the parameter $\alpha \geq 0$ is used to adjust the relative importance during optimization between the image reconstruction and force/torque prediction tasks.

$$\begin{aligned}\tilde{L}(\theta, \phi; \boldsymbol{X}, \boldsymbol{Y}) = & -\frac{\alpha}{1+\alpha}\|\boldsymbol{X} - \hat{\boldsymbol{X}}\| - \frac{1}{1+\alpha}\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\| \\ & + \beta D_{\mathrm{KL}}[\boldsymbol{N}(\boldsymbol{Z}_\mu, \boldsymbol{Z}_\sigma)||\boldsymbol{N}(\boldsymbol{0}, I)]\end{aligned} \tag{1}$$

Figure 2b shows the predicted 6D FT via SVAE against the ground truth. The R^2 scores are higher than 0.98 for 6D force and torque predictions, indicating the SVAE model's excellent performance in tactile sensing on the test dataset. We also plot the distributions of prediction errors in each 6D force/torque dimension over different ranges in Figure 2c. For applied forces ranging between $[0, 2)$, $[2, 4)$, $[4, 6)$, $[6, 8)$, and $[8, 10)$ N, the standard deviations of the prediction are 0.07, 0.06, 0.09, 0.12, and 0.24 N, respectively. For applied torques ranging between $[0, 120)$, $[120, 240)$, $[240, 360)$, $[360, 480)$, and $[480, 600)$ N· mm, the standard deviations of the prediction are 4.6, 3.9, 6.0, 9.1, and 20.3 N mm, respectively. These results follow a Gaussian distribution with a near-zero mean and an increasing standard deviation as the range becomes more considerable. The force-sensing errors are comparable in the x and y axes and more prominent in the z axis, while the torque-sensing errors are the least in the z axis. This characteristic is primarily due to the metamaterial's structural design, which is less sensitive to the force along the z axis.

We also conducted a comparative study to evaluate the proposed SVAE model against two baseline models, including a ConvNet model for force and torque prediction only and a VAE model for image reconstruction only, with results

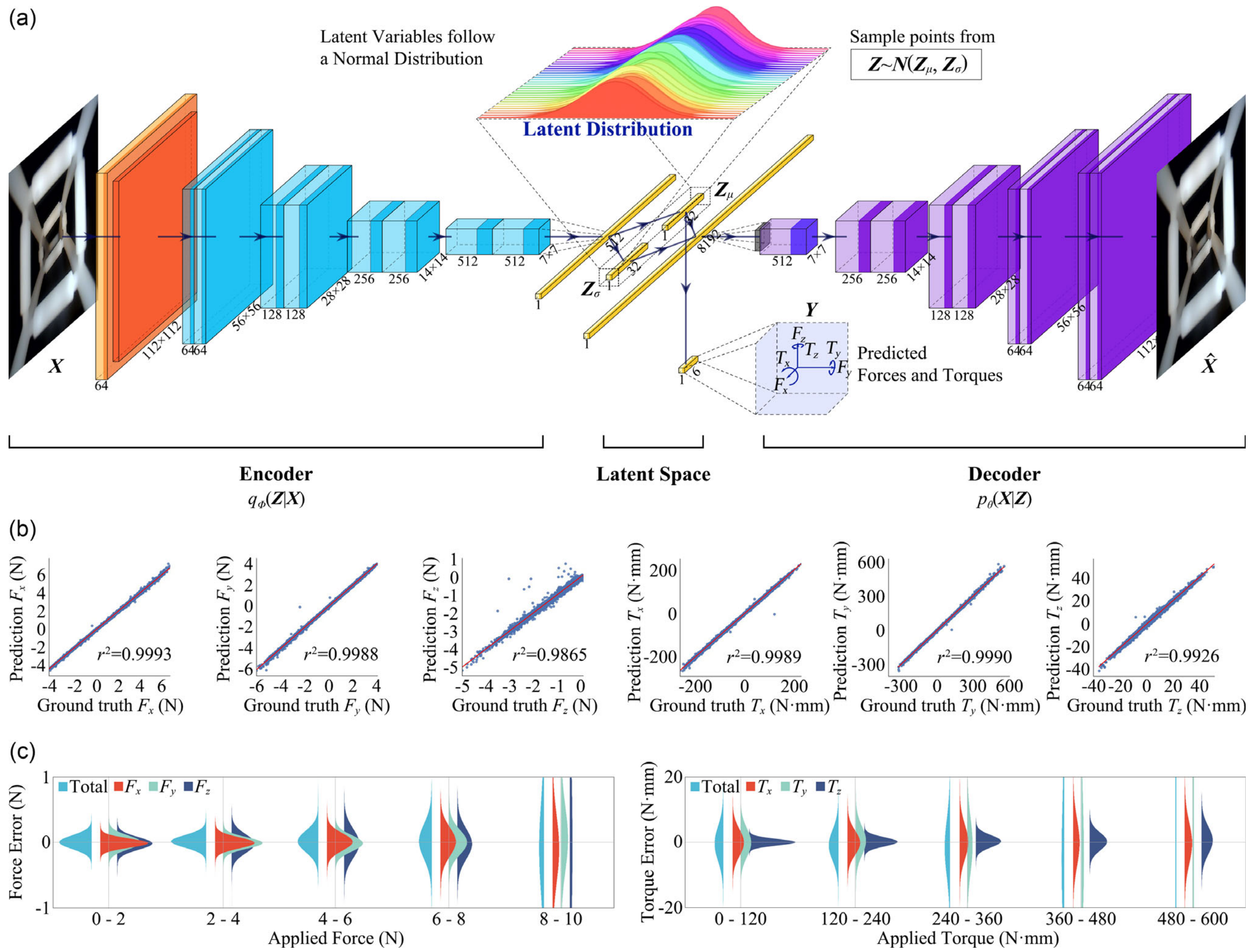

Figure 2. Latent deformation learning model for the soft metamaterial. a) The architecture of the SVAE model, where a VAE is combined with a supervised regression task for force and torque prediction. b) Predicted force/torque versus the ground truth in each of the six dimensions on the test dataset. c) Distributions of prediction errors in each 6D force/torque dimension over different ranges.

summarized in **Table 1**. These two models share the same network architecture as the SVAE but are trained separately. However, the ConvNet model is a deep regression network with a convolutional layer that only takes force/torque prediction loss.

Table 1. Comparative analysis of the proposed SVAE's performances.

Models	Settings	Image reconstruction error (MSE)	Force/torque prediction accuracy (avg.)
ConvNet	Vanilla	–	96.04%
SVAE	$\alpha = 0.001$	5.19×10^{-2}	99.53%
	$\alpha = 0.01$	5.17×10^{-2}	99.52%
	$\alpha = 0.1$	2.02×10^{-2}	99.53%
	$\alpha = 1$	9.47×10^{-3}	99.45%
	$\alpha = 10$	6.68×10^{-3}	96.22%
	$\alpha = 100$	5.65×10^{-3}	61.46%
VAE	Vanilla	5.36×10^{-3}	–

VAE model is a VAE that only takes image reconstruction and latent representation regularization loss. We used mean square error between the original and reconstructed images to evaluate the representation learning task and the coefficient of determination R^2 to assess the overall force/torque prediction task. See Methods S1 in the Supplementary Materials for further details on training data collection.

The SVAE has shown comparable performance over the vanilla VAE in the representation learning task while α is approaching infinity. Meanwhile, SVAE outperforms the deep regression model ConvNet in the force/torque prediction task when $\alpha \leq 1$ and the training is focused more on the prediction task, achieving over 99.45% on the validation set. Since SVAE is a multitask learning framework, the hyperparameter α is vital in balancing the reconstruction and prediction tasks. Here, $\alpha = 1$ is chosen for all validation tests and real-time experiments. The results show that the cotrained representation learning enhances the force/torque prediction task.

Considering the insufficient soft finger underwater deformation images and corresponding ground truth 6D force/torque data pairs, we trained the underwater SVAE model from one pretrained on-land SVAE model and fine-tuned it using limited underwater data. To test the transferability of the 6D force/torque prediction performance of the fine-tuned underwater SVAE model, we conducted the same collision experiments using soft fingers in both on-land and underwater scenarios for real-time 6D force/torque prediction. See Movie S2 in the Supplementary Materials for a video demonstration of real-time 6D force/torque prediction in the collision experiments. See Methods S2 in the Supplementary Materials for further discussion on the transferability of the model's predictive performance.

2.3. Land2Water Generalization of Tactile Representation

We also investigated the generalization of tactile representations learnt via SVAE in a Land2Water skill transfer problem for tactile

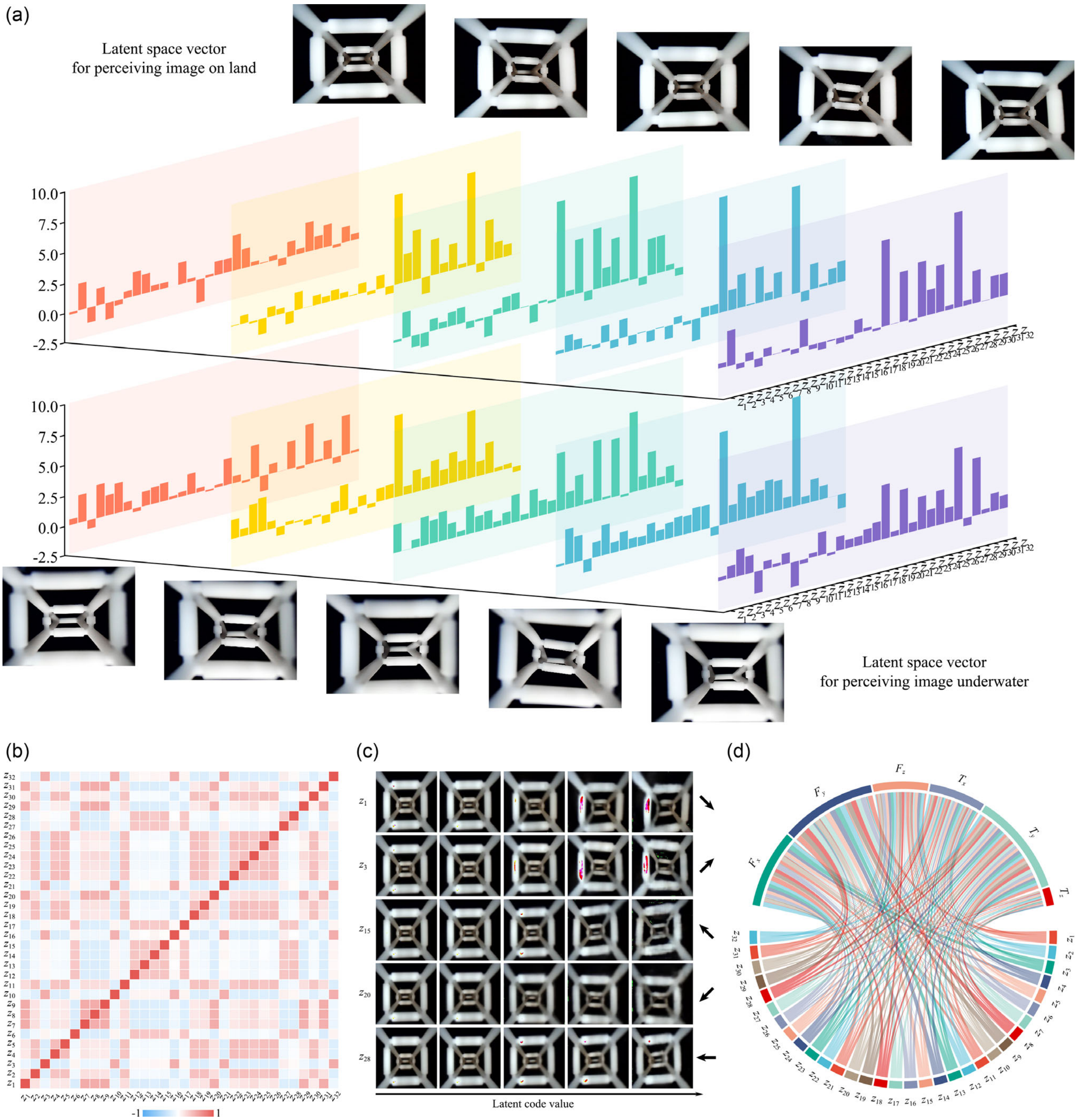

Figure 3. Representation learning of deformations of the proposed soft metamaterial. a) The complex deformations of the soft metamaterial on land and underwater are represented in the latent space. b) Correlation map of learnt 32 latent variables. c) Reconstructed images of varying selected latent variables. d) The relative correspondences between latent variables and force/torque.

sensing in **Figure 3**. While implementing the SVAE model, we chose a 32-dimension definition with a balanced trade-off between reconstruction error and dimensional complexity in explanatory power. See Methods S3 in the Supplementary Materials for further discussion.

Figure 3a shows the comparison of the 32D latent space vectors for tactile perception between on-land (top) and underwater (bottom) scenarios when the soft finger experiences the same deformation delivered by the robotic arm. Five random instances of the in-finger vision are chosen for each scenario and plotted with their corresponding latent variable distributions. We identified a similar distribution between the upper and lower plots for these five random instances. This suggests the transferability of the latent variables' explanatory power in tactile perception between the on-land and underwater scenarios. This is because the learnt latent representation could be close to the intrinsic dimension of the soft finger deformation, minimizing the information loss during tactile image encoding. The segment of the soft finger interacting with objects is made from 3D-printed metamaterial without any electronic parts, whose mechanical properties are not affected by water, indicating the generalization of tactile representation in Land2Water transfer, which is reported for the first time in the vision-based tactile-sensing literature.

The correlation map plotted in Figure 3b suggests that these 32 latent variables learnt from our SVAE model are generally unrelated, which is a preferred property for representation learning.[25] However, for variable clusters such as $\{Z_7, Z_8, Z_9\}$ and $\{Z_{12}, \ldots, Z_{15}\}$, regional correlation is observable at a relatively small scale. We also demonstrate the latent interpolation for the metamaterial's deformation projected in the image plane on selected dimensions of $\{Z_1, Z_3, Z_{15}, Z_{20}, Z_{28}\}$ in Figure 3c, which gives an intuitive sense of what are the latent variables in physical space. For example, we found that Z_1 and Z_3 are related to pushing right-downward and right-upward when their values go from negative to positive, while Z_{15} and Z_{20} are related to moving left-upward and left-downward. Furthermore, Z_{28} has a prominent horizontal movement. These latent variables are strongly related to representing the complex deformations of the soft metamaterial in terms of image reconstruction but are not disentangled. As shown in Figure 3d, the correlation between the 6D force/torque and the 32D latent variables is complex and diversified. For example, the latent variable Z_{28} strongly correlates with F_y, which agrees with the reconstructed horizontal movement along the corresponding axis of the camera coordinate.

2.4. Land2Water Grasping Knowledge Transfer

This section presents two experiment results that implement the Land2Water grasping knowledge obtained through tactile sensing, including one for object grasping against location uncertainties and another for tactile sensorimotor grasping adaptability from on-land to underwater scenarios.

2.4.1. Object Grasping Against Location Uncertainties

This experiment demonstrates the equal necessity of tactile perception when grasping underwater in **Figure 4**, which is generally acknowledged to increase the robustness in on-land conditions. Figure 4a,c shows the open-loop grasping without force feedback, where the gripper reaches the target grasping point, closes the fingers to a given gripping width, and lifts the object. In the case of closed-loop grasping with force feedback, as illustrated in Figure 4b,d, the gripper adjusts the gripping width according to the force estimation from SVAE until a grasping confirmation signal is triggered and then lifts the object. The prediction output from state-of-the-art learning-based grasp planning models usually contains grasping point position and gripper width.[36,37] To compare the performance of different grasping policies during grasping execution, we manually selected the gripper's grasping positions and gripping widths for each test object shown in Figure 4e and then added a small noise with standard deviation $\sigma = 5$ mm to simulate noises in gripper width prediction from grasping planning model. Figure 4f summarizes the ten grasping trials for each object using both methods and reports success rates. The average success rate for on-land grasping of the five test objects is 44% without contact feedback. After adding tactile feedback, the success rate is significantly enhanced to 100%. After adding tactile feedback, our results show a similar enhancement for underwater grasping, boosting the average success rate from 30% in open-loop grasping to 90% in closed-loop grasping. Figure 4g, h shows the histograms of the forces applied in 100 successful on-land and underwater grasps of the egg each, using different grasping policies. Compared with the open-loop grasping policy in both on-land and underwater scenarios, on the condition of simulated noisy gripper width, the variance of grasping forces applied using the closed-loop grasping policy is significantly reduced. See Movie S3 in the Supplementary Materials for a video demonstration. See Method S4 in the Supplementary Materials for a detailed force–time profile during each experiment.

2.4.2. Tactile Sensorimotor Grasping Adaptability

This experiment demonstrates sensorimotor grasping using tactile perception enabled by the proposed SVAE model, which is transferable from on-land to underwater scenarios. **Figure 5**a demonstrates the overall experiment process. Once contact begins, the in-finger vision captures the whole-body deformations of the soft finger and feeds the SVAE model with real-time image streams of the physical interaction at 120 Hz. The 6D FT are predicted for both on-land and underwater scenarios and then compared against a predefined threshold for reactive grasping. During this process, the width between the two soft fingers is actively adjusted to accommodate the disturbances in object status, that is, fluidic disturbances for grasping underwater and sudden collision for on-land grasping. We execute the reactive grasping by sending reference position commands to a position controller in the robot system using a motion generator calculated by the measured gripper position and force error detected on the fingers.

Figure 5b illustrates the experiment process that tests the gripper system's responsiveness of tactile-reactive grasping, a desirable capability for both on-land and underwater grasping of objects with known properties. After making contact with a

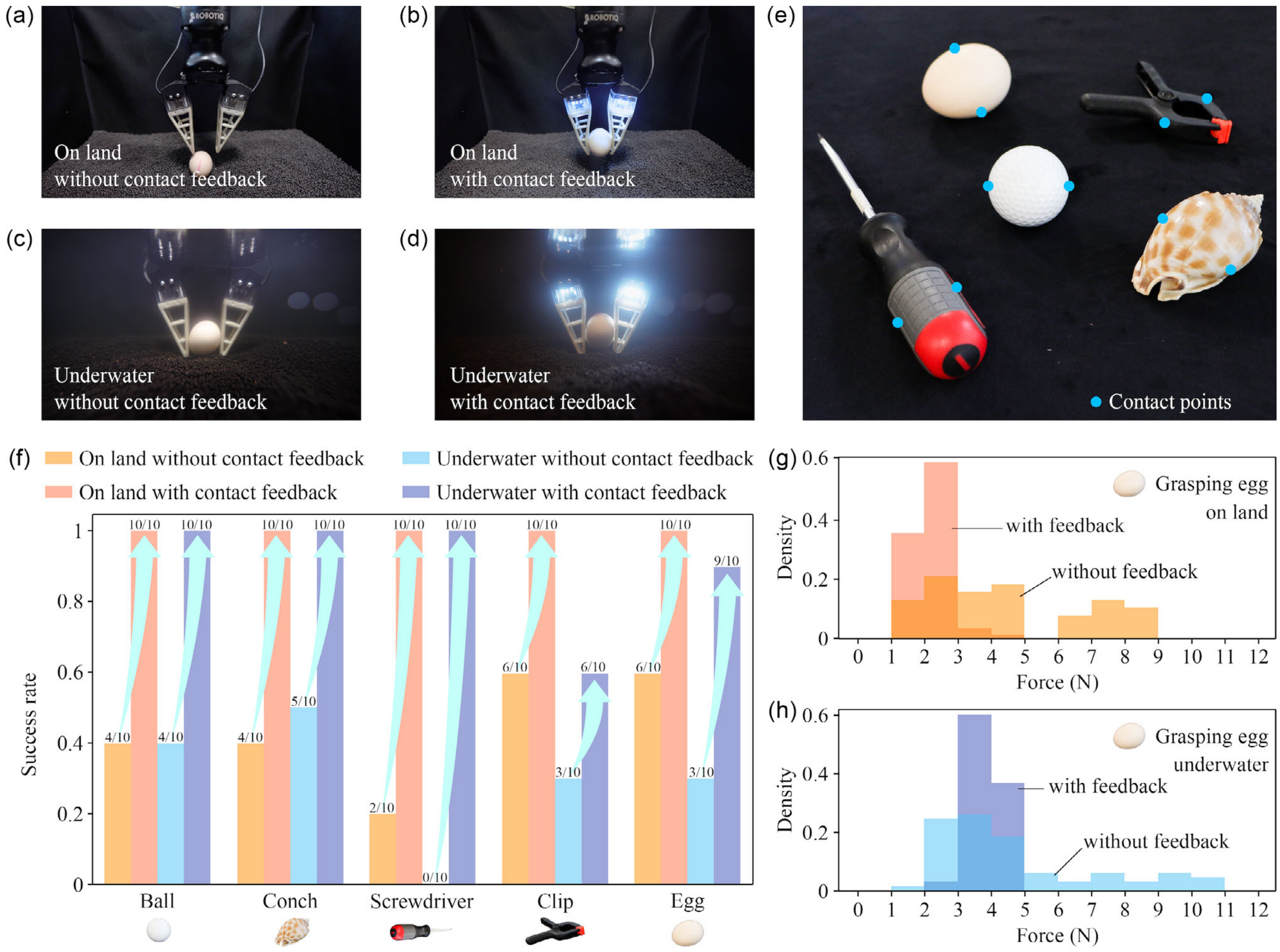

Figure 4. Tactile grasping results with or without SVAE in on-land and underwater scenarios. a) Open-loop object grasping on land with a predefined grasping position. b) Closed-loop grasping on land with contact force feedback. c) Open-loop object grasping underwater with a predefined grasping position. d) Closed-loop object grasping underwater with contact force feedback. e) Test objects with the predefined grasping points marked. f) The grasp result summary. g) Histogram of the forces applied in the successful on land grasps of the egg. h) Histogram of the forces applied in the successful underwater grasps of the egg.

slightly rotated tube, fixed, of oval cross section, we send force commands to the gripper to maintain a contact force at 0.4, 1.6, and 3 N sequentially. Shown in Figure 5c,d are the recorded force (in blue) against the commanded force (in red) when the experiment was conducted on-land and underwater. In both scenarios, the force controller successfully transited and stabilized at the commanded contact force within seconds. See Movie S4 in the Supplementary Materials for a video demonstration.

Figure 5e illustrates another experiment that tests the gripper's capability to maintain a specified contact force while reacting to disturbances, a preferred but more challenging skill for both on-land and underwater grasping of objects with unknown yet delicate properties. In this experiment, the oval-shaped tube is commanded to rotate clockwise in 45° and 60° first and then counterclockwise in 90° to simulate the changing interaction between the gripper and target object. During the process, the gripper needs to maintain a 0.4 N force for the on-land experiment in Figure 5f and a 1.6 N force for the underwater experiment in Figure 5g. When the target object changes its pose during rotation, the gripper reacts to the shape variation based on the estimated force from SVAE. See Movie S5 in the Supplementary Materials for a video demonstration. We also tested the gripper's reactive grasping under rotational disturbances by turning a cylinder along the z-axis. In this case, the SVAE model successfully predicted a torque while the fingers started twisting and commanded the gripper to rotate while maintaining a zero torque τ_z in reactive motion. See Movie S6 in the Supplementary Materials for a video demonstration.

3. Discussion

It has been a challenge to introduce robotic intelligence into underwater grasping by adding the sense of touch,[16] which supports delicate and autonomous interactions with the unstructured aquatic environment for scientific activities in environmental, biological, and ocean research. Classical solutions usually take a mechanical approach with various sealing technologies

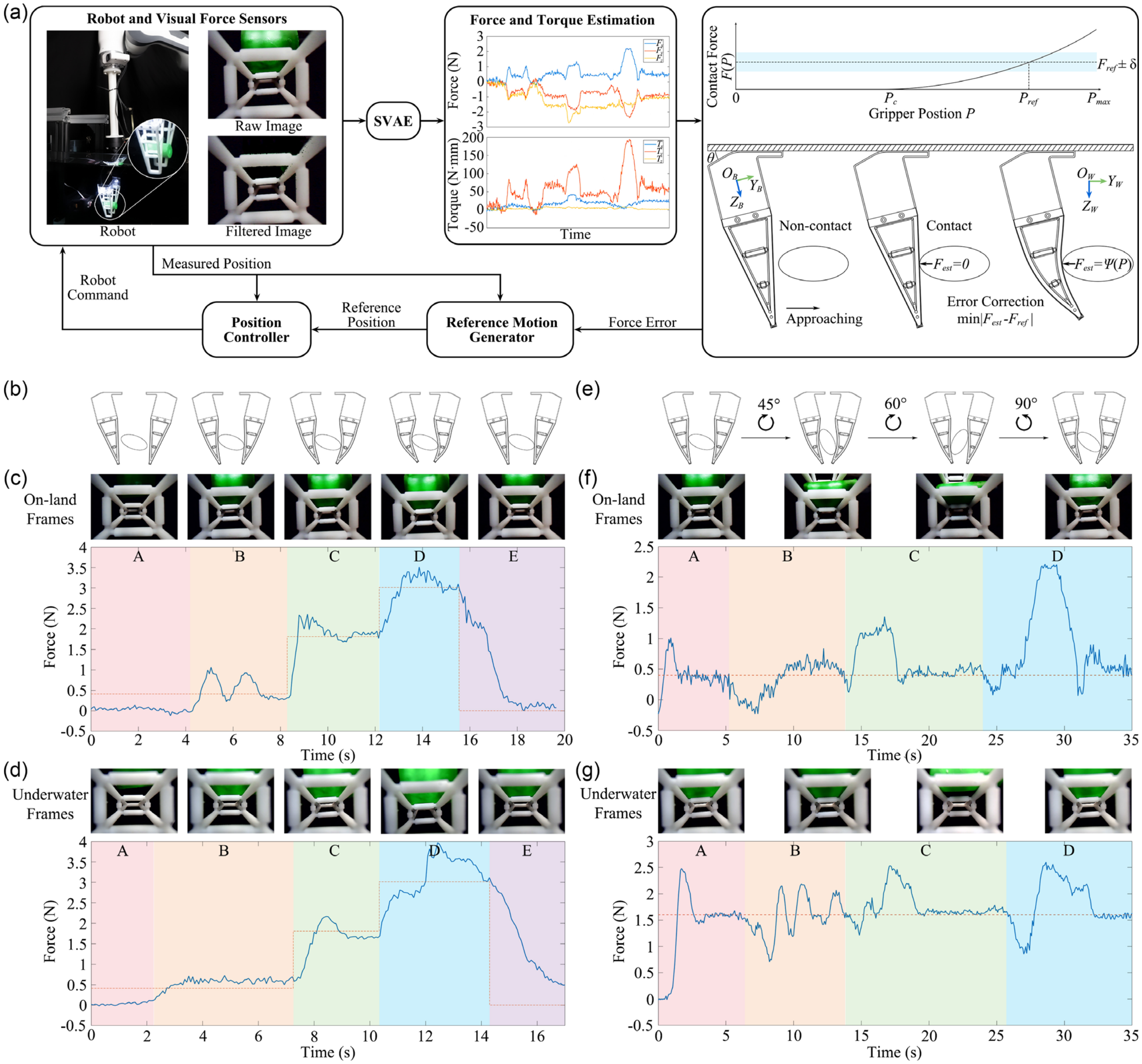

Figure 5. Tactile perception of soft finger for real-time robotic grasping control. a) Experiment setup of the force control tasks with soft tactile sensing. The goal of the force control tasks is to maintain the contact force at the required values by controlling the position of the soft finger. b–d) Desired gripping force following experiments. The gripping force is commanded to a series of expected values, and the corresponding gripper adaptation stages are illustrated in different colors. e–g) In-hand object shape adaptation experiments. The grasped object's shape changes constantly, and the gripper sensitively adjusts its position to maintain the constant gripping force.

to deal with fluidic pressure and corrosive contamination, suffering trade-offs in engineering flexibility and intelligent perception. This work proposes a vision-based approach to achieve high-performing tactile sensing underwater by combining the emerging advancement in soft robotics and machine learning. The simplicity of the design enables a minimum set of mechanical components, avoiding dynamic seals for enhanced robustness underwater. The soft finger's passive adaptation and in-finger vision enable a seamless integration of the proposed SVAE to learn tactile sensing through visual sensing underwater. The latent representations learnt from the SVAE algorithm enable a generative solution to infer the 6D FT during physical interactions underwater with explanatory reasoning. As a result, we successfully transferred the tactile intelligence of the proposed gripper system from on-land to underwater. We achieved tactile force prediction accuracy above 98% along each axis on the testset, using the same hardware with minimal algorithmic parameter adjustment. Real-time grasping experiment results in a lab tank demonstrate the effectiveness of the soft tactile finger for reliable and delicate grasping in both environments.

Model explainability and generalization are primarily concerned in machine learning research. Considering the transferability of tactile intelligence from on-land to underwater, we leveraged the VAE model's powerful representation learning capability to express the soft finger's deformation patterns in latent space. Results show that the extracted latent features of the same finger deformation in different environments exhibit a similar distribution. From the statistical inference perspective, learning this low-dimensional deformation pattern is closely related to the dimension reduction problem[38] where the learnt latent representation corresponds to an approximated sufficient statistics of original data.[39] In contrast with conventional dimension reduction techniques such as principal component analysis, the convolutional neural network usually performs better in finding the low-dimensional representation from image data.[40]

Performance degradation of tactile force prediction from on-land to underwater is unavoidable due to the significant change of visual input to the SVAE model. Due to unpredictable fluid dynamics, object grasping underwater is generally more challenging than on-land, which is the same case with or without tactile feedback, as demonstrated in our results. However, adding tactile feedback to the gripper system effectively enhanced the reliability of underwater grasping. The finger's network design cuts the fluids while closing, generally causing fewer disturbances for underwater grasping, a common problem usually suffered by fingers with a rigid structure.[9]

Tactile perception is generally desired to achieve effective grasping behaviors in underwater environments but is underexplored in research and practice compared with the on-land scenarios.[16,41] Operational tasks for underwater robotics are usually associated with a lack of vision, leading to ambiguous recognition of objects,[42] in which tactile perception plays an important role. Our results in grasping success rates demonstrate the benefit of tactile perception when visual perception is underperforming. Besides, reactive control architecture based on the perception–action cycle can be integrated with our tactile soft finger to achieve more intelligent manipulation underwater.

Our presented work has several limitations, which need future research for optimization in structural design and learning algorithms. For example, visual input tends to be corrupted by background noise in an underwater environment, which could be alleviated mechanically by adding a layer of silicone skin on the finger surface.[43] We could also enhance the tactile perception using XMem[44] to track the soft finger's deformation from the in-finger vision or use inpainting algorithms[45] to use the in-finger vision for visual perception. The proposed underwater grasping system is yet to be tested on remotely operated vehicles in shallow and deep water for further engineering enhancement.

4. Experimental Section

Formulating the Supervised Variational Autoencoder: Accurately deriving the relationship between deformation and force of soft structure can significantly improve the efficacy of visual–tactile sensing.[46] However, the geometry-dependent deformation of the soft structure is complex to represent. Even though we can discretize the structure with standard node elements using the finite-element method, measuring the displacements of corresponding nodes from a monocular camera can be another problem.

The standard solution involves a two-step method by first building a force–displacement mapping of soft structure and then solving the partial observable vision problem using a monocular camera.[47] Here, we leveraged the interpretability of latent variables in the original VAE model and constrained these learnt factors to image-based features of our soft finger deformation using in-finger vision, where the restored force could be measured during training and acted as a supervised signal to guide the learning of latent space.

As shown in Figure 2a, suppose the collected, labeled data pairs $(\boldsymbol{X}, \boldsymbol{Y})$ are independent and identically distributed, where $\boldsymbol{X}$ and $\boldsymbol{Y}$ are images and vectors of force/torque, respectively. The aim is to find an optimal representation $\boldsymbol{Z}$ of $\boldsymbol{X}$ containing sufficient information about $\boldsymbol{Y}$. To tackle both representation learning and force/torque prediction tasks, we extended the optimization framework of the original VAE[28] to an additional supervised task and maximized the log-likelihood function of marginal probability $\log p_\theta(\boldsymbol{X}, \boldsymbol{Y})$.

$$\log p_\theta(\boldsymbol{X}, \boldsymbol{Y}) = L(\theta, \phi; \boldsymbol{X}, \boldsymbol{Y}) + D_{\mathrm{KL}}[q_\phi(\boldsymbol{Z}|\boldsymbol{X}) || p_\theta(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{Y})] \tag{2}$$

where $L(\theta, \phi; \boldsymbol{X}, \boldsymbol{Y})$ is the evidence lower bound (ELBO) for SVAE, which can be extended as

$$\begin{aligned} \log p_\theta(\boldsymbol{X}, \boldsymbol{Y}) &\geq L(\theta, \phi; \boldsymbol{X}, \boldsymbol{Y}) \\ &= E_{\mathbf{Z} \sim q_\phi(\boldsymbol{Z}|\boldsymbol{X})}[\log p_\theta(\boldsymbol{X}|\boldsymbol{Z})] + E_{\boldsymbol{Z} \sim q_\phi(\boldsymbol{Z}|\boldsymbol{X})} \\ &\quad [\log p_\theta(\boldsymbol{Y}|\boldsymbol{Z})] - D_{\mathrm{KL}}[q_\phi(\boldsymbol{Z}|\boldsymbol{X}) || p_\theta(\boldsymbol{Z})] \end{aligned} \tag{3}$$

In Equation (3), for continuous data of image, force, and latent variables $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$, the prior distribution of the latent variables $p_\theta(\boldsymbol{Z})$, distribution of probabilistic encoder $q_\phi(\boldsymbol{Z}|\boldsymbol{X})$, and decoder $p_\theta(\boldsymbol{X}|\boldsymbol{Z})$, $p_\theta(\boldsymbol{Y}|\boldsymbol{Z})$ were assumed to follow a normal distribution.

$$\begin{aligned} p_\theta(\boldsymbol{Z}) &\sim \boldsymbol{N}(\mathbf{0}, I) \\ q_\phi(\boldsymbol{Z}|\boldsymbol{X}) &\sim \boldsymbol{N}(\boldsymbol{Z}_\mu(\boldsymbol{X}, \phi), \boldsymbol{Z}_\sigma(\boldsymbol{X}, \phi)) \\ p_\theta(\boldsymbol{X}|\boldsymbol{Z}) &\sim \boldsymbol{N}(\mathbf{X}_\mu(\boldsymbol{Z}, \theta), \boldsymbol{I}) \\ p_\theta(\boldsymbol{Y}|\boldsymbol{Z}) &\sim \boldsymbol{N}(\boldsymbol{Y}_\mu(\mathbf{Z}, \theta), \boldsymbol{I}) \end{aligned} \tag{4}$$

Maximization of the new ELBO in Equation (3) was equivalent to maximizing the following optimization object, where the outputs from two decoders were denoted as $\hat{\boldsymbol{X}}$ and $\hat{\boldsymbol{Y}}$, respectively.

$$\tilde{L}(\theta, \phi; \boldsymbol{X}, \boldsymbol{Y}) = -\|\boldsymbol{X} - \hat{\boldsymbol{X}}\| - \|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\| + D_{\mathrm{KL}}[\mathbf{N}(\boldsymbol{Z}_\mu, \boldsymbol{Z}_\sigma) || \boldsymbol{N}(\mathbf{0}, \boldsymbol{I})] \tag{5}$$

Therefore, we built a hierarchical, convolutional, multiscale model for the encoder and decoder to model the long-range correlations in image data. We used four residual serial blocks to extract and reconstruct image features in different scales.[48] The first two terms in Equation (5) measured reconstruction errors and force/torque prediction errors, respectively. The third term encouraged the approximated posterior $q_\phi(\boldsymbol{Z}|\boldsymbol{X})$ to match the prior $p_\theta(\boldsymbol{Z})$, which controlled the capacity of latent information bottleneck. Although the derived optimization objective function Equation (5) implicitly balanced the three sources of loss, its optimization could be complex in practice. To resolve this issue, we proposed the formulation of Equation (1) in Section 2.2 by introducing hyperparameters α and β to Equation (5).

Introducing parameter $\beta \geq 0$ ahead of the third term of Equation (1) was inspired by the work of Higgins et al.[32] so that the optimal β could be estimated heuristically in unsupervised scenarios. We tested several choices of β in a candidate set, ranging from 10^{-4} to 10^2, and fixed $\beta = 0.1$ in our experiment.

All networks were trained on a computer with NVIDIA GTX 1080Ti GPU, a batch size 64, and Adam optimizer.[49] Considering the relatively small dataset size, the initial learning rate was set to 5×10^{-5} and decreased with the training epoch.

Algorithm 1. Reference Motion Generator.

```
Input:  Raw image $I_{raw}$; Measured position $P_m$;
        Reference force $F_{ref}$
Output: Reference position $P_{ref}$
Initialize force tolerance $\delta$ and control gain $K$
while True do
    Filtered image $I_f \leftarrow ColorThreshold(I_{raw})$
    Estimated Force $F_{est} \leftarrow VisualForceNet(I_f)$
    if $|F_{ref} - F_{est}| > \delta$ then
        $\Delta_P = \frac{1}{K}(F_{ref} - F_{est})$
        $P_{ref} = P_m + \Delta_P$
    else if $|F_{ref} - F_{est}| \leq \delta$ then
        $P_{ref} = P_m$
    end if
end while
```

Tactile Grasping from On-Land to Underwater: We conducted object grasping experiments in a lab tank with and without contact force feedback for both on-land and underwater conditions to demonstrate the benefit of tactile learning in reliable object grasping against environmental uncertainties. We tested the grasping success rate using objects of different shapes, sizes, and materials from on-land to underwater. With the adoption of learnt tactile perception, two more tasks were tested to demonstrate intelligent grasping behaviors in both on-land and underwater conditions.

As is shown in Figure 5a, to achieve an intelligent closed-loop grasping behavior, it is an essential requirement for the grasping system to maintain a specified contact force while reacting to the varying environment. The industrial gripper could achieve reliable position commands at a high bandwidth due to the built-in low-level position controller. It is our goal to design a high-level position control policy $u = \pi(P_m, F_{est}, F_{ref})$ with measured gripper position P_m and estimated contact force F_{est} that achieves the desired contact force F_{ref}.

$$\pi(P_m, F_{est}, F_{ref}) = \arg\min_u |F_{est} - F_{ref}| \tag{6}$$

Thanks to the proposed tactile force proprioceptive soft finger, which acts simultaneously as an end-effector and a sensor, a heuristic control policy $\pi = P_{ref}$ was presented to generate the reference motion command for the inner low-level position control loop, as shown in **Algorithm 1**. The frequency of tactile perception feedback was determined by the computational time cost of the proposed SVAE model and the frame rate of the USB camera. We used a 1060Ti 6G GPU laptop in all grasping experiments, and the average inferring time was 5 ms. As a result, the force controller frequency was bound by the camera frame rate at 120 Hz. Note that to estimate the contact force parallel to the gripping direction, modification of SVAE output was necessary. See Methods S5 and Methods S6 in the Supplementary Materials for a detailed derivation of controller design.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

This work was partly supported by the Ministry of Science and Technology of China (2022YFB4701200), the National Natural Science Foundation of China (62206119), the Science, Technology, and Innovation Commission of Shenzhen Municipality (ZDSYS20220527171403009 and JCYJ20220818100417038), Guangdong Provincial Key Laboratory of Human-Augmentation and Rehabilitation Robotics in Universities, and the SUSTech-MIT Joint Centers for Mechanical Engineering Research and Education.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Github at https://github.com/bionicdl-sustech/AmphibiousSoftFinger, reference number 4291526.

Keywords

soft robotics, tactile learning, underwater grasping

Received: July 4, 2023
Revised: September 14, 2023
Published online: October 22, 2023

[1] J.-C. Feng, J. Liang, Y. Cai, S. Zhang, J. Xue, Z. Yang, *Sci. Bull.* **2022**, *67*, 1802.
[2] Z. Gong, X. Fang, X. Chen, J. Cheng, Z. Xie, J. Liu, B. Chen, H. Yang, S. Kong, Y. Hao, T. Wang, J. Yu, L. Wen, *Int. J. Robot. Res.* **2021**, *40*, 449.
[3] A. Billard, D. Kragic, *Science* **2019**, *364*, eaat8414.
[4] K. S. Kumar, P.-Y. Chen, H. Ren, *Research* **2019**, *2019*, 3018568.
[5] M. T. Ciocarlie, P. K. Allen, *Int. J. Robot. Res.* **2009**, *28*, 851.
[6] D. Mura, M. Barbarossa, G. Dinuzzi, G. Grioli, A. Caiti, M. G. Catalano, *IEEE Robot. Autom. Mag.* **2018**, *25*, 45.
[7] C. E. Capalbo, D. Tomaino, F. Bruno, D. Rizzo, B. Phillips, S. Licht, *IEEE J. Ocean. Eng.* **2022**, *47*, 975.
[8] K. C. Galloway, K. P. Becker, B. Phillips, J. Kirby, S. Licht, D. Tchernov, R. J. Wood, D. F. Gruber, *Soft Robot.* **2016**, *3*, 23.
[9] H. Stuart, S. Wang, O. Khatib, M. R. Cutkosky, *Int. J. Robot. Res.* **2017**, *36*, 150.
[10] S. Licht, E. Collins, D. Ballat-Durand, M. Lopes-Mendes, in *OCEANS 2016 MTS/IEEE Monterey*, Monterey, CA, USA **2016**, pp. 1–5.
[11] J. Yuh, *Auton. Robots* **2000**, *8*, 7.
[12] D. Lane, J. Davies, G. Robinson, D. O'Brien, J. Sneddon, E. Seaton, A. Elfstrom, *IEEE J. Ocean. Eng.* **1999**, *24*, 96.
[13] J. R. Bemfica, C. Melchiorri, L. Moriello, G. Palli, U. Scarcia, in *IEEE Inter. Conf. on Robotics and Automation (ICRA)*, Hong Kong, China **2014**, pp. 2469–2474.
[14] S. Wang, P. Yan, H. Huang, N. Zhang, B. Li, *Research* **2023**, *6*, 0133.
[15] Z. Wang, W. Cui, *Proc. Inst. Mech. Eng., Part M* **2021**, *235*, 3.
[16] A. Mazzeo, J. Aguzzi, M. Calisti, S. Canese, F. Vecchi, S. Stefanni, M. Controzzi, *Sensors* **2022**, *22*, 648.
[17] R. Bao, J. Tao, J. Zhao, M. Dong, J. Li, C. Pan, *Sci. Bull.* **2023**, *68*, S2095.

[18] P. Xu, J. Zheng, J. Liu, X. Liu, X. Wang, S. Wang, T. Guan, X. Fu, M. Xu, G. Xie, Z. L. Wang, *Research* **2023**, *6*, 0062.
[19] Y. Liu, R. Bao, J. Tao, J. Li, M. Dong, C. Pan, *Sci. Bull.* **2020**, *65*, 70.
[20] R. Li, B. Peng, *Cyborg Bionic Syst.* **2022**, *2022*, 9797562.
[21] K. Shimonomura, *Sensors* **2019**, *19*, 3933.
[22] S. Zhang, Z. Chen, Y. Gao, W. Wan, J. Shan, H. Xue, F. Sun, Y. Yang, B. Fang, *IEEE Sens. J.* **2022**, *22*, 21410.
[23] K. Sato, K. Kamiyama, N. Kawakami, S. Tachi, *IEEE Trans. Haptics* **2010**, *3*, 37.
[24] A. Yamaguchi, C. G. Atkeson, *Int. J. Humanoid Robot.* **2019**, *16*, 1940002.
[25] Y. Bengio, A. Courville, P. Vincent, *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1798.
[26] W. Yuan, S. Dong, E. H. Adelson, *Sensors* **2017**, *17*, 2762.
[27] C. Doersch, A. Gupta, A. A. Efros, in *IEEE Inter. Conf. on Computer Vision (ICCV)*, Santiago, Chile **2015**, pp. 1422–1430.
[28] D. P. Kingma, M. Welling, in *Inter. Conf. on Learning Representations (ICLR)*, Banff, AB, Canada, 14–16 April **2014**.
[29] D. J. Rezende, S. Mohamed, D. Wierstra, in *Inter. Conf. on Machine Learning (ICML)*, Beijing, China **2014**, pp. 1278–1286.
[30] H. Takahashi, T. Iwata, Y. Yamanaka, M. Yamada, S. Yagi, in *AAAI Conf. on Artificial Intelligence (CAI)*, Vol. *33*, Honolulu, Hawaii **2019**, pp. 5066–5073.
[31] D. P. Kingma, S. Mohamed, D. J. Rezende, M. Welling, in *Advances in Neural Information Processing Systems (NIPS)*, Montreal, Quebec, Canada **2014**, pp. 3581–3589.
[32] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, A. Lerchner, in *Inter. Conf. on Learning Representations (ICLR)*, Toulon, France **2017**.
[33] T. Ji, S. T. Vuppala, G. Chowdhary, K. Driggs-Campbell, in *Conf. on Robot Learning*, PMLR, London, UK **2021**, pp. 1443–1455.
[34] F. Wan, X. Liu, N. Guo, X. Han, F. Tian, C. Song, In A. Faust, D. Hsu, G. Neumann (Eds), *Proceedings of the 5th Conf. on Robot Learning, volume 164 of Proceedings of Machine Learning Research*, PMLR, Auckland, New Zealand **2022**, pp. 1269–1278.
[35] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, A. M. Dollar, *IEEE Robot. Autom. Mag.* **2015**, *22*, 36.
[36] D. Morrison, P. Corke, J. Leitner, *Int. J. Robot. Res.* **2020**, *39*, 183.
[37] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, K. Goldberg, *Sci. Robot.* **2019**, *4*, eaau4984.
[38] K. P. Adragni, R. D. Cook, *Philos. Trans. R. Soc. A* **2009**, *367*, 4385.
[39] P. Joyce, P. Marjoram, *Stat. Appl. Genet. Mol. Biol.* **2008**, *7*, 1.
[40] G. E. Hinton, R. R. Salakhutdinov, *Science* **2006**, *313*, 504.
[41] Y. Yan, Z. Hu, Z. Yang, W. Yuan, C. Song, J. Pan, Y. Shen, *Sci. Robot.* **2021**, *6*, eabc8801.
[42] R. A. S. I. Subad, L. B. Cross, K. Park, *Appl. Mech.* **2021**, *2*, 356.
[43] H. Jiang, X. Han, Y. Jing, N. Guo, F. Wan, C. Song, *Front. Robot. AI* **2021**, *8*, 787187.
[44] H. K. Cheng, A. G. Schwing, in *Computer Vision–ECCV 2022: 17th European Conf., Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVIII*, Springer, Cham **2022**, pp. 640–658.
[45] T. Yu, R. Feng, R. Feng, J. Liu, X. Jin, W. Zeng, Z. Chen, arXiv preprint, arXiv:2304.06790, **2023**.
[46] D. Ma, E. Donlon, S. Dong, A. Rodriguez, in *2019 Inter. Conf. on Robotics and Automation (ICRA)*, IEEE, Piscataway, NJ **2019**, pp. 5418–5424.
[47] X. Dong, M. A. Garratt, S. G. Anavatti, H. A. Abbass, *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 16940.
[48] K. He, X. Zhang, S. Ren, J. Sun, in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA **2016**, pp. 770–778.
[49] D. Kingma, J. Ba, in *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA **2014**.