

# Robot Learning System Based on Target Localization and Human Demonstration for Medical Examination

Weiyong Si , Member, IEEE, Zhehao Jin , Member, IEEE, Ning Wang , Member, IEEE, and Chenguang Yang , Fellow, IEEE

**Abstract**—Robot-assisted ultrasound scanning has been proven to be an effective solution for medical examination; however, most of the existing work is fully teleoperation or semiautonomous due to dynamic tasks and uncertain environments. In this article, we propose a novel robot learning system for autonomous ultrasound scanning that integrates dynamic target localization with motion generation based on learning from demonstration (LfD). First, a new LfD model, Gaussian process dynamic movement primitives (GPDMP), is introduced to encode human motion skills and generate motion trajectories online. The proposed GPDMP framework supports multiple demonstrations and guarantees convergence toward the target even in the presence of external disturbances, such as subject movement during ultrasound scanning. To further enhance generalization to moving subjects through robust target localization, a vision-based localization method is developed to identify and localize the neck region. Neck positioning across individuals with different body shapes is addressed by an improved AlphaPose network, while interference from other moving subjects in the scene is mitigated using an enhanced DeepSORT-based tracking algorithm. The proposed system is capable of automatically identifying and

localizing the human neck, generating approach trajectories for the ultrasound probe, and achieving real-time interaction. A robot-assisted neck ultrasound scanning system is implemented to experimentally demonstrate the effectiveness of the proposed framework.

**Index Terms**—Dynamic movement primitive (DMP), Gaussian process (GP), learning from demonstration (LfD), robotic ultrasound system, vision-based target localization.

## I. INTRODUCTION

ULTRASOUND scanning has been proven an effective diagnostic solution in medical examinations due to its low-cost, noninvasive, etc., which has been widely employed within clinical contexts [1]. Nowadays, ultrasound scanning is often conducted by experienced sonographers, who require many years of training and practical experience to acquire proficiency in scanning techniques. Therefore, experienced sonographers are not enough as the aging people grow, and the situation is more serious in rural areas. For example, according to the estimation in the U.K., there is a shortage of experienced sonographers, and a similar situation arises in China [2]. On the other hand, ultrasound scanning often requires the sonographers to execute force on the human body constantly, which may cause muscle fatigue due to working for many years. In addition, the ultrasound images obtained by different sonographers often exhibit inconsistencies. In light of these situations and challenges, solutions incorporating robotic assistance for ultrasound scanning have emerged, as exemplified by the authors in [3], [4], [5], [6], [7], and [8]. The robot-assisted ultrasound scanning (RUS) holds the potential to enhance the working conditions of sonographers, mitigating the risk of muscular fatigue and minimizing occupation-related strain.

Regarding the medical robots, Zhang et al. [9] reviewed the relevant technology from teleoperation to autonomous medical robots. Learning from demonstration (LfD) has been investigated to transfer human skills to robots. Among existing LfD methods, two predominant paradigms have emerged: dynamic system-based method [10], [11] and statistical-based methods [12]. Dong et al. [13] studied robot skill learning, combining the LfD and a broad learning system for contact-rich tasks. The LfD method can be enhanced by the adaptive neural network control for robot skill learning and reproducing [14].

Received 5 April 2025; revised 27 July 2025 and 15 November 2025; accepted 27 December 2025. Recommended by Technical Editor A. K. Hoshir and Senior Editor S. Katsura. (Corresponding authors: Ning Wang; Chenguang Yang.)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Faculty Research Ethics Committee and University of the West of England Research Ethics Committee under Application No. UWE REC and FET-2122-59.

Weiyong Si was with Bristol Robotics Laboratory, BS16 1QY Bristol, U.K. He is now with the School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, U.K.

Zhehao Jin was with Bristol Robotics Laboratory, BS16 1QY Bristol, U.K. He is now with Nanyang Technological University, Singapore 639798.

Ning Wang was with Bristol Robotics Laboratory, BS16 1QY Bristol, U.K. She is now with the School of Computing and Digital Technologies, Sheffield Hallam University, S1 2NU Sheffield, U.K. (e-mail: wang.ning@ieee.org).

Chenguang Yang was with Bristol Robotics Laboratory, BS16 1QY Bristol, U.K. He is now with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong, HKSAR (e-mail: cyang@ieee.org).

In addition, a range of statistical methods from machine learning have been introduced into LfD, such as Gaussian process-movement primitive (GPMP) [15] for encoding primitive skill. LfD methods offer an effective solution for enabling robots to acquire human manipulation skills, contrasting with large data-driven approaches, such as reinforcement learning (RL). In the context of medical examination tasks, training a robot in an RL environment is challenging, even though RL has achieved success in various daily and industrial tasks, such as dressing and assembly [16]. While some studies have explored combining RL and LfD, developing fully autonomous medical robots for practical applications remains difficult. This is largely due to the dynamic character of medical examination tasks and conditions, such as variability among patients and human motion.

RUS has attracted considerable research attention, with a growing body of work focusing on robotic ultrasound examination systems [17], [18], [19]. Remote examination via teleoperation involves ultrasound experts in the control loop. Various teleoperation interfaces and multimodal feedback have been developed to improve the telepresence of human operators. For example, a virtual reality (VR)-based interface is presented for ultrasound scanning [5]. Haptic interfaces, such as the Touch X and Omega devices, are also investigated to provide force feedback to ultrasound experts, enabling them to perceive the interaction between the ultrasound probe and the human body [20]. However, fully autonomous solutions are still under development.

This robot learning system aims to automate the ultrasound scanning procedure for arteries in the neck, including target recognition, localization, and autonomous motion planning based on human demonstrations. The RUS localization process is achieved by teleoperation ultrasound robots based on the shared control method [9]. In recent years, RUS has applied machine vision methods for feature extraction to achieve localization in specific scanning regions [21]. For example, some scholars have used RGB images to segment the human from the background and locate the desired 2-D scan target on the human body. Depth images have been used to resolve the 3-D normal pose of the probe for RUS [17], [22], [23]. Convolutional neural networks (CNN) are popular for their ability to automatically learn features in images and obtain feature structures at different scales by changing the network structure to achieve better region localization. It has been widely used in tasks, such as target detection. For example, both the two-stage detection region CNN [24], [25], [26] and the one-stage detection led by YOLO [27], [28], [29], [30] have achieved good accuracy in object recognition. This work investigates a dynamic object recognition neural network for RUS of the neck artery. The main contributions of this work can be summarized as follows.

- 1) A novel GPDMP model for encoding medical examination skills during the approaching stage. By integrating Gaussian process (GP) with dynamic movement primitive (DMP), the GPDMP model effectively captures and generalizes human motion skills from multiple demonstrations. The GPDMP model could encode human skills from multiple demonstrations, which could synthesize

different experts' skills, enhancing the robustness and quality of the training data.

- 2) An improved AlphaPose network combined with an enhanced DeepSORT tracking algorithm to achieve robust target recognition and localization even under environment disturbance. Compared with the conventional method, the improved algorithm could deal with environmental noise effectively, making it suitable for real-world applications where robustness is critical.
- 3) A robot learning system for autonomous ultrasound scanning, including the target recognition and localization, and autonomous motion generation modules, was developed. The proposed autonomous scanning method could adapt to dynamic targets, such as moving humans, and generate trajectories based on human demonstrations.

The rest of this article is organized as follows. The target localization method is presented in Section II. In Section III, the learning from human demonstration method, GPDMP, is presented. The design of the experiment and the results and performance analysis are presented in Section IV. Finally, Section V concludes this article and provides future work.

## II. TARGET RECOGNITION AND LOCALIZATION

In this work, we studied a region localization algorithm for robot ultrasound scanning tasks, as shown in Fig. 1. In this section, the target recognition and localization methods are presented, including the coarse-grained human skeleton segmentation, fine-grained neck segmentation, and spatial localization algorithms. For target detection and regional localization, we adopted the coarse-grained and fine-grained concepts from Viola Jones Detectors [31]. A large framework in the background window was utilized to extract the entire human body region, followed by the application of the weakly supervised learning method from deformable part-based model [32]. The initial YOLOv3-tiny and Alpha pose were considered as human feature point detectors, with the subsequent neck detector and associated operations serving as a complementary filter to address the human neck problem. This approach can also address the challenge of limited labeled training data. By training the YOLOv3-tiny and Alpha pose networks, all possible configurations of the subsequent neck detectors as part filters could be automatically learned as latent variables.

### A. Coarse-Grained Human Skeleton Segmentation and Tracking

1) *Targeted Human Tracking Based on YOLOv3-Tiny and DeepSORT*: It is difficult to directly segment and track the neck region using human skeletal information in complex environments, such as those involving multiple people. To improve system adaptability across different scenes and individuals, we trained a YOLOv3-tiny network to detect a single human body class; the network structure is illustrated in Fig. 2. Compared with YOLOv3, YOLOv3-tiny features a lightweight architecture and a substantially reduced number of trainable parameters. The dataset used in this study includes 1200 images selected from the

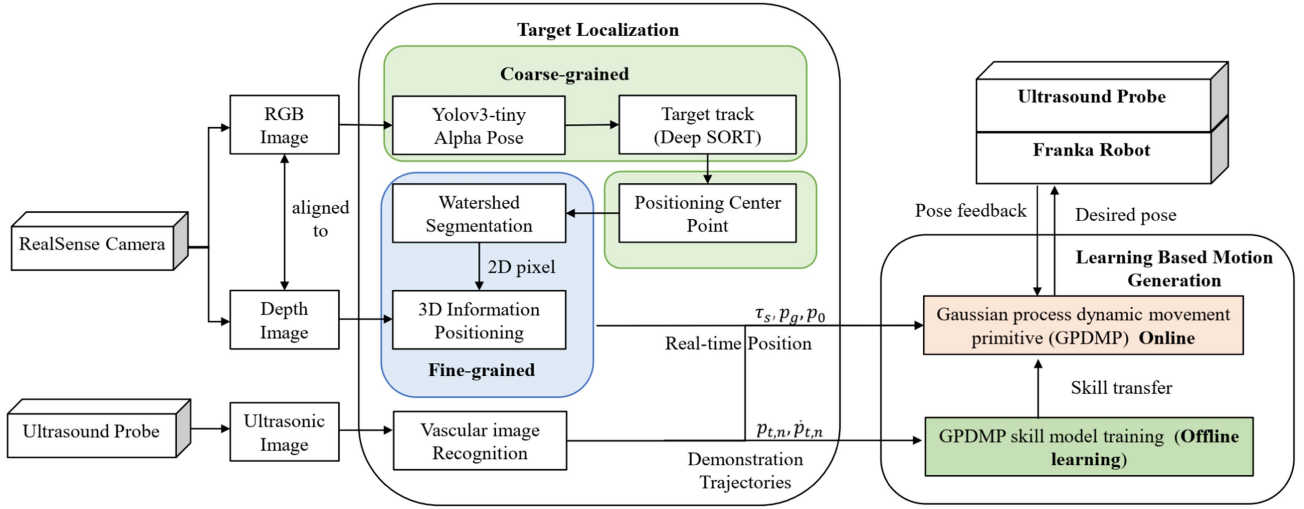


Fig. 1. Overview of the proposed RUS system.

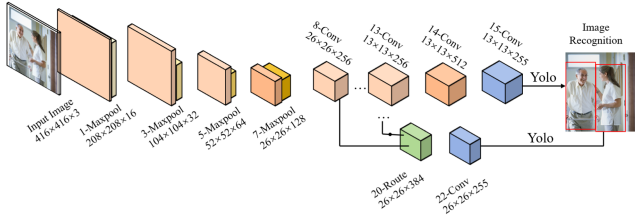


Fig. 2. Structure of YOLOv3-tiny network.

person category of the MS COCO 2017 object detection dataset, along with additional human body images corresponding to real ultrasound scanning scenarios, manually annotated using LabelMe.

Incorporating a DeepSORT-based people tracking algorithm improves system tracking performance and mitigates misjudgments in scenes involving multiple individuals. Moreover, matching newly detected targets over three consecutive frames using DeepSORT effectively reduces the false positive rate of the original YOLOv3-tiny detector. The DeepSORT-based primary people tracking algorithm incorporates a major identity-locking module into the traditional DeepSORT framework, enabling reliable estimation of candidate bounding boxes corresponding to the primary person in the scene. As shown in Fig. 3, the primary identity module adopts a standard nonmaximum suppression (NMS) strategy. All detected individuals in the scene are first considered, and their corresponding bounding boxes are ranked according to their distance from the image center, from nearest to farthest. The detections are then processed sequentially: if the overlap area between a candidate bounding box and the previously selected bounding box exceeds a predefined threshold, then the candidate is suppressed. This procedure prevents interference caused by overlapping detections from multiple individuals and ensures stable selection of the primary target. Finally, among the remaining candidate bounding boxes, the one with the highest intersection over union (IoU) value

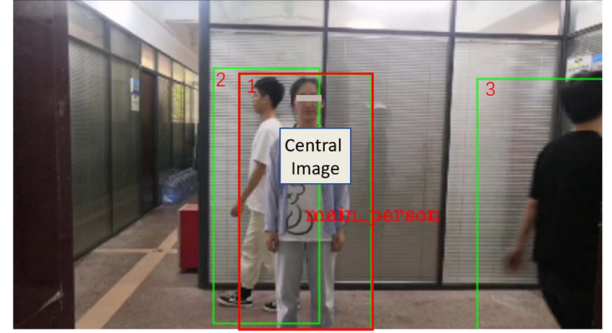


Fig. 3. Primary identity module illustration. The detection bounding box labeled as 1 has the largest IoU with the central reference image and is therefore selected as the main target.

with the central reference region is selected as the primary target.

**2) Region of Interest (ROI) Positioning Based on Improved AlphaPose:** The ROI is defined as the neck region for ultrasound scanning and is localized using skeletal keypoints predicted by the improved AlphaPose. Using the target detection algorithm, the human body region of the primary person is first extracted. This region is then fed into the AlphaPose network [33] for human skeletal keypoint extraction. The architecture of the improved AlphaPose network is shown in Fig. 4. To localize the neck artery for ultrasound scanning, a regional center localization module is incorporated after the Pose NMS stage in the improved AlphaPose framework.

The architecture of the spatial transformer network (STN) [34] is illustrated in Fig. 5. The STN applies a learnable spatial transformation to the input feature maps, operating on the entire feature representation. Its key advantage lies in its ability to compensate for geometric variations—including translation, rotation, and scaling—by mapping the input features into a canonical form suitable for subsequent network stages, thereby facilitating accurate candidate region identification.

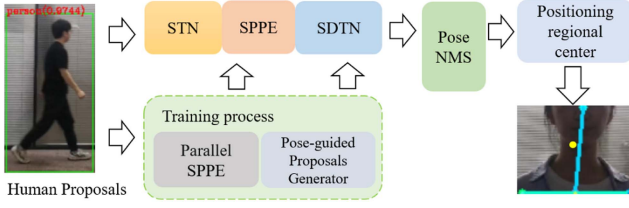


Fig. 4. Overview of the improved AlphaPose network. The detailed architectures of the STN and SDTN modules are presented in Figs. 5 and 6, respectively.

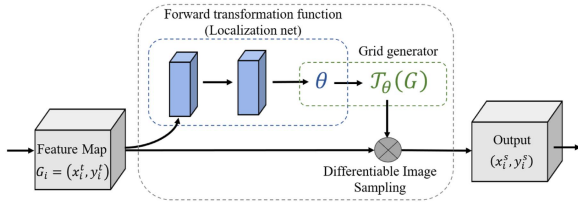


Fig. 5. Architecture of the STN, consisting of a localization network, a grid generator, and a differentiable image sampling module.

Localization network performs operations, such as rotation, translation, and scaling of 2-D images, as shown in (1), resulting in a mapping  $\theta$  of the input  $(x, y)$  to the output  $(x', y')$ . Grid generator performs a spatial data transformation, transforming a normal image or feature map into a more familiar look for network training and recognition, as shown in (2)

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} e \\ f \end{pmatrix} \quad (1)$$

where  $\begin{pmatrix} x \\ y \end{pmatrix}$  represents the original coordinates in the input image, and  $\begin{pmatrix} x' \\ y' \end{pmatrix}$  are the transformed coordinates in the output image.

The matrix  $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$  is the linear transformation submatrix, where  $a$  and  $d$  scale the coordinates along the  $x$  and  $y$  axes, respectively and  $b$  and  $c$  control the shearing (skewing) effects. The vector  $\begin{pmatrix} e \\ f \end{pmatrix}$  is the translation component, where  $e$  represents the horizontal translation (shift along the  $x$ -axis) and  $f$  represents the vertical translation (shift along the  $y$ -axis).

This equation describes how each pixel  $(x, y)$  in the input image is mapped to  $(x', y')$  in the output image. The STN learns parameters  $a-f$  automatically to apply transformations (rotation, scaling, and translation) that optimize downstream processing

$$\begin{pmatrix} x_i^s \\ y_i^s \end{pmatrix} = T_\theta(G_i) = A_\theta \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix} \quad (2)$$

where  $x_i^t$  and  $y_i^t$  are the target coordinates (a regular grid in the output image), and  $x_i^s$  and  $y_i^s$  are the corresponding source coordinates in the input image.  $\begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix}$  is the extended affine transformation matrix, where  $a$  and  $e$  are scaling factors,  $b$  and  $d$  are shearing factors, and  $c$  and  $f$  are translation components. The

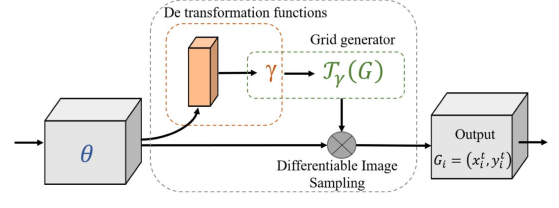


Fig. 6. Architecture of the inverse SDTN, which maps the estimated pose from the normalized coordinate space back to the original image coordinates via a grid generator and differentiable image sampling.

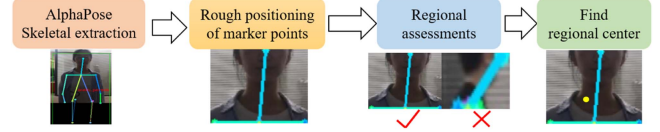


Fig. 7. Workflow of regional positioning. The neck region is coarsely localized from skeletal landmarks, evaluated for posture validity, and refined to obtain the center of neck.

1 in  $\begin{pmatrix} x_i^t \\ y_i^t \\ 1 \end{pmatrix}$  enables the matrix to represent affine transformations that include translation.

The architecture of the inverse spatial detransformer network (SDTN) is illustrated in Fig. 6. The SDTN maps the estimated human pose from the normalized coordinate space back to the original image coordinates, thereby generating the pose proposal. As the inverse transformation of the STN, the relationship between SDTN and STN is defined by the following equation:

$$\gamma = \theta^{-1}. \quad (3)$$

Between the STN and SDTN, a parallel single-person pose estimator (Parallel SPPE), implemented as a four-layer hour-glass network, is employed to evaluate whether the centroid of the estimated human pose is aligned with the center of the human detection bounding box. If a misalignment is detected, then a larger error is produced and backpropagated, which in turn drives the optimization of the STN. This feedback mechanism encourages the STN to extract more accurate human detection regions.

The positioning regional center network explicitly models the geometric relationship between key anatomical landmarks and the ROI. Based on human skeletal segmentation, major anatomical landmarks are first extracted to achieve a coarse localization of the ultrasound scanning area. Subsequently, a medical human body scale model [35] is introduced to compute the average distance ratios between the selected landmark points and the ROI, thereby normalizing intersubject variations. Using these scale-consistent relationships, the central position of the ultrasound diagnostic region is accurately determined. The detailed algorithmic procedure is illustrated in Fig. 7. Posture conformity is determined according to the following requirements.

- 1) The human body is approximately centered in the image, and both shoulder landmark points and the nose landmark point are visible.

- 2) The subject is facing forward, such that the nose landmark lies horizontally between the left and right shoulder landmarks and is vertically positioned above the shoulder landmarks.
- 3) The neck region exhibits an approximately symmetric configuration, where the horizontal position of the nose landmark lies within a 50% threshold of the midpoint between the shoulder landmarks.

### B. Fine-Grained Watershed-Based Neck Segmentation

The proposed fine-grained image segmentation and spatial localization framework comprises three main components: a watershed-based neck segmentation module, an image post-processing module, and a depth-data-based spatial localization module. In this study, the neck region is selected as the primary experimental target, and the algorithm is accordingly tailored for this application. The watershed algorithm is employed to segment the neck region, while anatomical feature points are incorporated during postprocessing to improve the accuracy of neck boundary extraction.

Specifically, the input image is first preprocessed by grayscale conversion, binarization, and median filtering to suppress noise. Image gradients are then computed using the Sobel operator, followed by a region-growing process to obtain initial segmented regions. A watershed transform is subsequently applied to the labeled image to produce distinct region segments. The segmentation results are filtered to retain only the region containing the centroid of interest, after which a broad search is performed from the centroid along the horizontal and vertical directions to accurately determine the neck boundaries.

Moreover, due to variations in lighting conditions and other environmental factors, the point cloud acquired by the RGB-D sensor contains noise, which can lead to inaccuracies in depth estimation. Such depth errors directly affect Z-axis estimation for motion planning. Therefore, postprocessing of the depth data is required. In this work, a spatial positioning module for depth data is proposed by refining the data normalization strategy. Specifically, the depth value is extracted from a small rectangular region located near the detected neck edge and offset inward by five pixels. This inward offset is introduced to better approximate the depth of the underlying neck artery, which lies beneath the skin surface, and the resulting value is taken as the representative depth of the neck edge.

Furthermore, the depth data extracted at the neck edge may contain outliers. Outlier detection is therefore performed using a box plot-based method, in which the interquartile range (IQR) is computed from the overall statistical characteristics of the dataset. According to standard box plot criteria, data points lying beyond 1.5 times the IQR above the upper quartile or below the lower quartile are identified as outliers, as defined in the following equation:

$$\begin{cases} \text{Outliers.value} < (Q1 - 1.5 * IQR) \\ \text{Outliers.value} > (Q3 + 1.5 * IQR) \end{cases} \quad (4)$$

where Q1 represents the first quartile, Q3 denotes the third quartile, and IQR is defined as Q3 - Q1. Data points falling

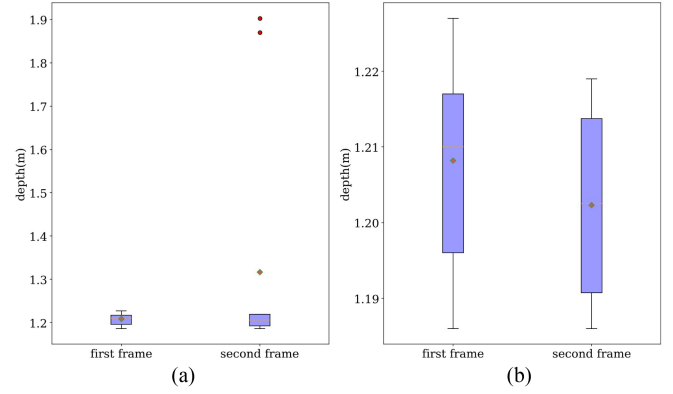


Fig. 8. Comparison of depth data distributions before and after outlier removal using the box plot method: (a) original data and (b) filtered data after removing outliers.

beyond  $(Q3 + 1.5 * IQR)$  or  $(Q1 - 1.5 * IQR)$  are considered as outliers, and “\*” is considered as standard elementwise multiplication.

Using depth data extracted from two randomly selected frames, a box plot was constructed [see Fig. 8(a)], revealing that outliers significantly distorted the mean depth estimation. After applying the box plot-based outlier removal method, the resulting depth data exhibited a more representative and stable distribution, as shown in Fig. 8(b). The mean value computed from the filtered dataset was then adopted as the depth estimate corresponding to the neck edge point.

## III. MOTION PLANNING

### A. GP-Based DMP for Primitive Motion Planning

DMPs have been widely employed in LfD to encode human motor skills [36]. A recent comprehensive review provides an overview of DMP formulations and discusses the latest developments in this area [11]. The fundamental formulation of DMPs is presented in the following equation:

$$\begin{aligned} \tau_s \dot{v} &= \alpha_z (\beta_z (p_g - p) - v) + F(s) \\ \tau_s \dot{p} &= v \end{aligned} \quad (5)$$

where  $p_g$  indicates the target position,  $p$  indicates the current position,  $\tau_s$  is the temporal scaling parameter, and  $v$  is the velocity.  $\alpha_z$  and  $\beta_z$  need to be designed separately, but these parameters generally satisfy  $\alpha_z = 4\beta_z$ .  $F(s)$  is a nonlinear forcing term that affects the trajectory.  $F(s)$  can be defined as a Gaussian function

$$F(s) = \frac{\sum_{j=1}^N \psi_j(s) w_j}{\sum_{j=1}^N \psi_j(s)} s(p_g - p_0) \quad (6)$$

$$\psi_j(s) = \exp(-\eta_j (s - c_j)^2) \quad (7)$$

where  $p_0$  denotes the initial position of the expected trajectory. The function  $F(s)$  is the result of locally weighted regression using  $N$  basis functions  $\psi_j$ . The variable  $s$  guarantees that the effect of  $F(s)$  decreases as  $s$  decreases. The parameters  $c_j$  and

$\eta_j$  represent the center and width of  $\psi_j$ , respectively.  $w_j$  denotes the weight assigned to the corresponding basis function  $\psi_j$ , and  $j$  denotes the index of the corresponding basis function  $\psi_j$ .

A canonical system can coordinate different degrees of freedom, which can be described as

$$\tau_s \dot{s} = -\alpha_s s, \quad s \in [0, 1]; \quad s(0) = 1 \quad (8)$$

where  $\tau_s$  is the scaling parameter, and the positive coefficient  $\alpha_s$  are denoted, where  $s(0) = 1$  represents the initial value of  $s$ , which exponentially converges to zero.

We studied learning from multiple demonstration methods to improve the learning and generalization capability. GP was used to model the multiple demonstration data. GP regression is used to obtain the trajectory features of multidemonstrations to train the forcing term  $F(s)$ . The Gaussian distribution can be described by the following formula:

$$P_Z(z) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right) \quad (9)$$

where  $Z$  denotes a Gaussian random variable and  $z$  represents its realization. The normal distribution of  $Z$  is commonly denoted as  $Z \sim \mathcal{N}(\mu, \sigma^2)$ .

Given a training set  $D_i = \{t_k, q_{k,i}\}$ , where  $t_k$  is the time of the  $k$ th sampling step, and  $q_{k,i}$  represents the dimension of the  $i$ th step at time  $t_k$ , we can assume that the training set is generated by the following model:

$$q_{k,i} = h_i(t_k) + \epsilon_{k,i} \quad (10)$$

where  $\epsilon_{k,i}$  represents the observation noise at the  $i$ th position in time  $t_k$ , following a Gaussian distribution  $\mathcal{N}(0, \sigma_{no,i}^2)$ . Furthermore, we assume that the function  $h_i(t^*)$  is fitted by a  $\mathcal{GP}$ , which leads to the following equation:

$$h_i(t^*) \sim \mathcal{GP}(m_i(t^*), k_i(t, t')) \quad (11)$$

where  $m_i(t^*) : R \rightarrow R$  and  $k_i(t, t') : R \times R \rightarrow R_{++}$  are, respectively, the mean function and covariance function (kernel function) of the GP. Equation (11) implies that the output vector  $\mathbf{q}_i = [q_{1,i}, \dots]^T \in R^N$ , where  $N$  is the size of the training set, follows the following multivariate Gaussian distribution:

$$p(\mathbf{q}_i|\mathbf{t}) = \mathcal{N}(\mathbf{m}_i(\mathbf{t}), K_i(\mathbf{t}, \mathbf{t}) + \sigma_{no,i}^2 I) \quad (12)$$

where  $\mathbf{m}_i(\mathbf{t}) \in R^N$  is a vector where the  $k$ th component is given by  $\mathbf{m}_{i,k} = m_i(t_k)$ .  $K_i(t, t') \in R^{N \times N}$  is the covariance matrix with the form  $K_i(t, t')_{ab} = k_i(t_a, t'_b)$ .  $\mathbf{t} = [t_1, t_2, \dots, t_N]^T$ . As in most GP applications, the mean function  $m_i(t)$  is set to  $m_i(\cdot) = 0$ . The hyperparameters  $\theta_i$  of the GP are hidden in the kernel function  $k_i(\cdot, \cdot)$ , the form of which is determined by the chosen kernel function. In this article, since most of the approximated target functions are smooth and continuous differentiable, the Gaussian kernel function  $k_i(t, t') = \sigma_{ker,i}^2 \exp\left\{-\frac{(t-t')^2}{l_i^2}\right\}$  is chosen. Therefore, the hyperparameters are defined as  $\theta_i = \{\sigma_{ker,i}, l_i\}$ . The optimal hyperparameters  $\theta_i^*$  can be obtained by maximizing the likelihood function  $\log p(q_i|t)$  on the training set. Once the optimal hyperparameters are obtained, we can use the multivariate Gaussian distribution to predict the values  $h_i(t^*) \in R^M$  of the input vector  $t^* \in R^M$ ,

expressed as follows:

$$p(h_i(t^*)|t^*, D_i) = \mathcal{N}(\mu_i(t^*, D_i), \Sigma_i(t^*, D_i)) \quad (13)$$

where  $M$  is the number of inputs. The mean function  $\mu_i(t^*, D_i) \in R^M$  and the covariance function  $\Sigma_i(t^*, D_i) \in R^{M \times M}$  of the posterior distribution can be analyzed and calculated as follows [37]:

$$\begin{cases} \mu_i(t^*, D_i) = K_i(t^*, \mathbf{t}) (K_i(\mathbf{t}, \mathbf{t}) + \sigma_{no,i}^2 I)^{-1} \mathbf{q}_i \\ \Sigma_i(t^*, D_i) = K_i(t^*, t^*) - \\ K_i(t^*, \mathbf{t}) (K_i(\mathbf{t}, \mathbf{t}) + \sigma_{no,i}^2 I)^{-1} K_i(\mathbf{t}, t^*). \end{cases} \quad (14)$$

The nonlinear forcing term in DMP can guarantee convergence by multiplying the phase variable  $s$ , which will converge to 0. Inspired by our previous work [15], the GP-MP could guarantee the reproducing trajectory via a given point. By investigating the property of the GP-MP, we guarantee convergence by adding a desired point to the demonstration dataset. For example, we add  $(t_{\text{end}}, 0)$  to the demonstration dataset. When the running time is close to infinite, the forcing term will be 0. The constrained points form the dataset  $\bar{D}_i = \{\bar{t}, \bar{q}_i\}$ . Combining the constrained set  $\bar{D}_i$  and the original demonstration data set  $D_i$ , we get the new set  $\tilde{D}_i = \{D_i, \bar{D}_i\}$ . The GP-MP could guarantee the forcing term converges to 0.

Taking the 1-D position DMP as an example, the nonlinear forcing term is learned via GP regression in the time domain. The forcing term is then evaluated along the phase trajectory through the canonical system. Accordingly, the forcing term can be expressed as

$$F(s(t)) = h_i(t) \sim \mathcal{GP}(m_i(t), k_i(t, t')). \quad (15)$$

## B. Control System Design

The dynamics of the general serial manipulator robot in joint space can be modeled as

$$M(q)\ddot{q} + C(q, \dot{q})\dot{q} + G(q) + \tau_f(\dot{q}) + \tau_{ic} + d_u = \tau_c + \tau_{\text{ext}} \quad (16)$$

where  $M(q)$  is the inertia matrix,  $C(q, \dot{q})$  is the Coriolis term, and  $G(q)$  represents gravitational force.  $\tau_{ic}$  is the internal coupling and static friction,  $\tau_f(\dot{q})$  is velocity-dependent friction torque,  $d_u$  accounts for unmodeled dynamics and uncertainties, and  $\tau_c$  is the control torque.  $q$ ,  $\dot{q}$ , and  $\ddot{q}$  represent the joint position, velocity, and acceleration, respectively.

*Property 1:* Matrix  $M(q)$  is symmetric and positive definite.

*Property 2:* Matrix  $\dot{M}(q) - 2C(q, \dot{q})$  is a skew-symmetric matrix.

The relationship between the joint torque and the wrench can be described as

$$\tau_{\text{ext}} = J^T(q) f_e \quad (17)$$

where  $J(q)$  is the Jacobian matrix and  $f_e$  is the interaction force with the environment.

In addition, the controller designed in the Cartesian space will also benefit the generalization across different platforms with various configurations. Therefore, we derive the controller in the Cartesian space. Based on the kinematic equation of the

manipulator, we describe the position, velocity, and acceleration of the end-effector as

$$x(t) = f(q) \quad (18)$$

$$\dot{x}(t) = J(q)\dot{q} \quad (19)$$

$$\ddot{x} = \dot{J}(q)\dot{q} + J(q)\ddot{q} \quad (20)$$

where  $\dot{x}$  and  $\ddot{x}$  are the velocity and acceleration of the robot end-effector in the Cartesian space, respectively.  $\dot{J}(q)$  is the derivative of the Jacobian matrix.

To track the desired pose and contact force, there are a number of human robot controllers [38], [39], [40]. The unified motion and force controller [40] is adopted. In the experiment, the robot moves safely toward the human body and establishes initial contact. To ensure compliant interaction, we employ a Cartesian admittance control strategy, which adjusts both the desired position and velocity in response to external forces.

Let  $x \in \mathbb{R}^6$  denote the end-effector pose, and  $f_e \in \mathbb{R}^6$  the external force/torque measured by the F/T sensor. The admittance control law is defined as

$$M_a\ddot{x}_d + D_a\dot{x}_d + K_a(x_d - x) = f_e \quad (21)$$

where the following holds.

- 1)  $M_a \in \mathbb{R}^{6 \times 6}$ : virtual inertia matrix.
- 2)  $D_a \in \mathbb{R}^{6 \times 6}$ : damping matrix.
- 3)  $K_a \in \mathbb{R}^{6 \times 6}$ : stiffness matrix.

This model is numerically integrated to generate

- 1) the desired Cartesian position  $x_d(t)$ , and
- 2) the desired Cartesian velocity  $v_d(t) = \dot{x}_d(t)$ .

These two quantities form the control signal for the robot's tracking controller. The desired joint velocity and acceleration are obtained from the Cartesian velocity and acceleration using differential kinematics

$$\dot{q}_d = J^{-1}(q) \dot{x}_d \quad (22)$$

$$\ddot{q}_d = J^{-1}(q) \left( \ddot{x}_d - \dot{J}(q, \dot{q}) \dot{q}_d \right). \quad (23)$$

We assume a nonsingular square Jacobian matrix  $J(q)$ . The reference motion in the Cartesian space is mapped to joint space using inverse kinematics and then transformed into desired joint torques  $\tau_d$  using

$$\tau_d = M(q)\ddot{q}_d + C(q, \dot{q})\dot{q}_d + G(q) - J^T(q)f_e \quad (24)$$

where  $\dot{q}_d$  and  $\ddot{q}_d$  are obtained from  $v_d$  and  $\ddot{x}_d$  via differential kinematics. To update the desired Cartesian position and velocity over time, the following numerical integration is used:

$$\begin{aligned} \dot{x}_d(t) &= \dot{x}_d(t - \Delta t) + \Delta t \cdot \ddot{x}_d(t) \\ x_d(t) &= x_d(t - \Delta t) + \Delta t \cdot \dot{x}_d(t). \end{aligned} \quad (25)$$

To maintain safety and stability, the admittance parameters  $D_a$  and  $K_a$  are designed to ensure compliant but stable behavior. In the Cartesian admittance controller, the damping matrix  $D_a \in \mathbb{R}^{6 \times 6}$  and stiffness matrix  $K_a \in \mathbb{R}^{6 \times 6}$  are defined to regulate both translational and rotational motions of the end-effector.

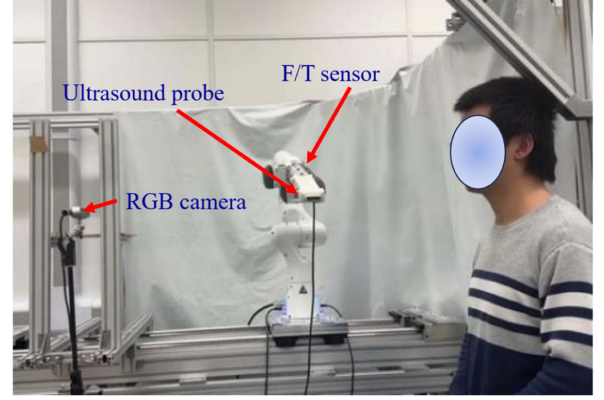


Fig. 9. RUS for neck.

TABLE I  
PARAMETERS OF THE ADMITTANCE CONTROLLER

| End-effector | Stiffness matrix |                                                                          |         | Damping matrix                                                         |  |  |
|--------------|------------------|--------------------------------------------------------------------------|---------|------------------------------------------------------------------------|--|--|
| Position     | $K_p =$          | $\begin{bmatrix} 200 & 0 & 0 \\ 0 & 200 & 0 \\ 0 & 0 & 50 \end{bmatrix}$ | $D_p =$ | $\begin{bmatrix} 30 & 0 & 0 \\ 0 & 30 & 0 \\ 0 & 0 & 15 \end{bmatrix}$ |  |  |
| Orientation  | $K_O =$          | $\begin{bmatrix} 50 & 0 & 0 \\ 0 & 50 & 0 \\ 0 & 0 & 50 \end{bmatrix}$   | $D_O =$ | $\begin{bmatrix} 15 & 0 & 0 \\ 0 & 15 & 0 \\ 0 & 0 & 15 \end{bmatrix}$ |  |  |

Specifically, the admittance matrices are constructed in a block-diagonal form as

$$D_a = \begin{bmatrix} D_p & \mathbf{0} \\ \mathbf{0} & D_O \end{bmatrix}, \quad K_a = \begin{bmatrix} K_p & \mathbf{0} \\ \mathbf{0} & K_O \end{bmatrix} \quad (26)$$

where  $D_p, K_p \in \mathbb{R}^{3 \times 3}$  correspond to the position damping and stiffness matrices, and  $D_O, K_O \in \mathbb{R}^{3 \times 3}$  correspond to the orientation damping and stiffness matrices, respectively. This decoupled design enables independent compliance regulation along the translational and rotational degrees of freedom.

## IV. EXPERIMENT

### A. Experimental Platforms

An RUS system was developed for experiments, which is shown in Fig. 9. A 7-DoF Franka Emika Panda robotic arm is equipped with an ultrasound probe, RGB-D camera, and F/T sensor to perform the scanning task. A Touch X was used as the teleoperation device. A RealSense camera (D435i) was used to record RGB images, which could also provide visual feedback during human teleoperation. Control computer running Ubuntu 18.04, which was connected to the Touch X device, Franka Emika Panda, and the camera. ROS was used to integrate different components. The whole experimental procedure was conducted in accordance with the Faculty Research Ethics Committee, and the protocol was approved by the UWE Research Ethics Committee (UWE REC REF No: FET-2122-59).

Table I outlines the parameters of the admittance controller. In the case of the orientation controller, the quaternion representation is utilized.

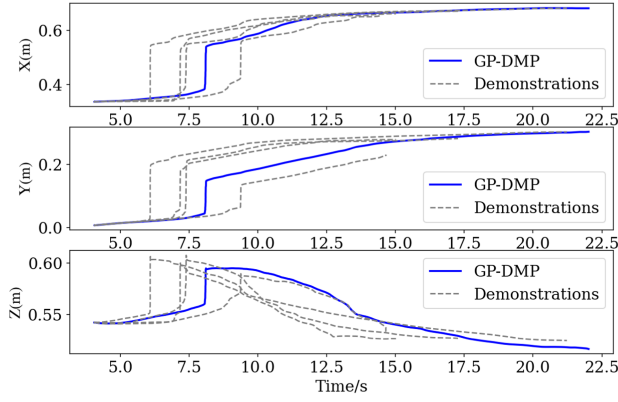


Fig. 10. 2-D demonstration and reproducing trajectories generated by GP-DMP.

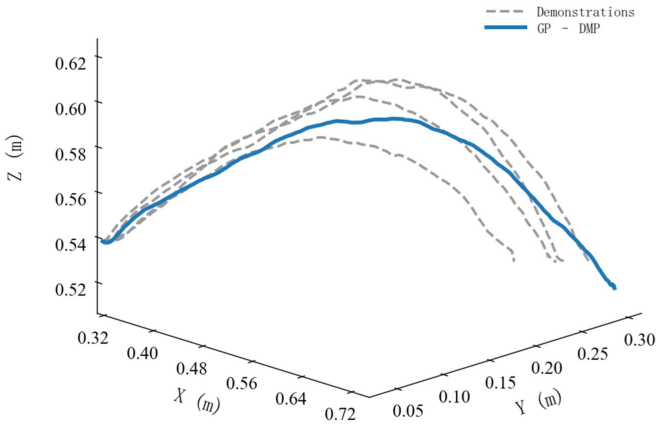


Fig. 11. 3-D demonstration and reproducing trajectories generated by GP-DMP.

## B. Experimental Results

During the experiment, teleoperation-based human demonstrations were used to collect a demonstration dataset for learning the motion skill from the initial position to the desired neck pose. Five demonstrations were performed and their demonstration trajectories are shown in Figs. 10 and 11. Fig. 10 illustrates the GP fitting of the demonstration trajectories. By comparing the reproduced trajectories with the demonstrations, it can be observed that the proposed GPDMP method accurately captures and reproduces the demonstrated motion patterns. Moreover, even when the target position deviates significantly from the demonstrations, GPDMP is able to generate smooth trajectories that converge to the desired goal, demonstrating generalization capability to novel target positions. The 3-D trajectories are shown in Fig. 11.

The GPDMP could generate the desired trajectory for the ultrasound probe based on target localization results online. The experimental results are, respectively, shown in Figs. 12 and 13. Fig. 12 illustrates the RGB video stream of the ultrasound probe approaching the human neck, along with the corresponding ultrasound images. The yellow contour overlaid on the RGB frames indicates the target position for neck artery scanning.

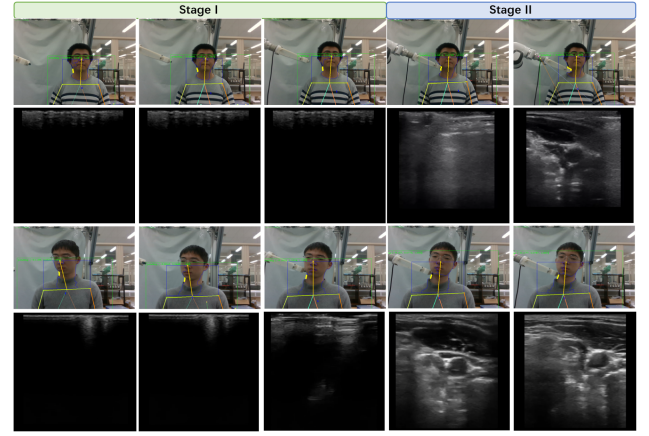


Fig. 12. Robotic ultrasound scanning approach process on the neck. In Stage I, the robotic arm approaches the neck region based on visual tracking. In Stage II, the robot reaches the target position and acquires corresponding ultrasound images. Experimental results on two subjects are shown.

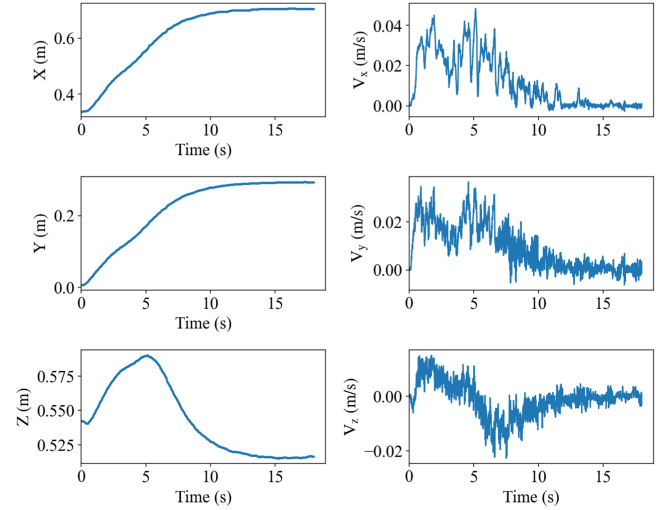


Fig. 13. Position and velocity trajectories of the ultrasound probe.

As shown, the robotic arm gradually approaches the neck and accurately reaches the target location. This neck tracking method can successfully work on different people without the need of retraining or setting new parameters. We have demonstrated the generation capability on two subjects as shown in Fig. 12. In fact, we have evaluated the method on a total of six subjects with different body shapes. Fig. 13 illustrates the position and velocity profiles of the ultrasound probe during the experiment. As shown, the probe exhibits smooth and gradual positional changes along all directions, indicating flexible and stable motion behavior. Jerk, defined as the time derivative of acceleration, is commonly used to quantify trajectory smoothness, with lower jerk values corresponding to smoother motion trajectories

$$J = \int_{t_0}^{t_f} \left\| \frac{d^3 \mathbf{x}(t)}{dt^3} \right\|^2 dt$$



Fig. 14. RUS under human motion. Sequential frames illustrate the human body movement process, and the aggregated trajectory of the subject is shown on the right. The red dashed line represents the human motion trajectory.

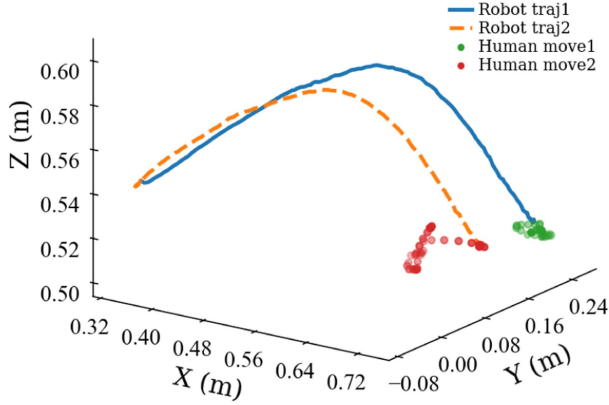


Fig. 15. Online motion generation for two moving subjects. Solid curves indicate the reproduced robot trajectories, while scattered points represent the corresponding human motion trajectories.

where  $\mathbf{x}(t)$  is the position at time  $t$

$$J \approx \sum_{k=2}^{N-1} \left\| \frac{\mathbf{a}_{k+1} - 2\mathbf{a}_k + \mathbf{a}_{k-1}}{\Delta t^2} \right\|^2$$

where  $\mathbf{a}_k$  is the acceleration at time step  $k$ , and  $\Delta t$  is time interval. Lower  $J$  means smoother motion. We compared the GPDMP and the DMP trajectory; the  $J_{\text{DMP}}$  is 24.53 and the  $J_{\text{GPDMP}}$  is 19.32. The trajectory generated by GPDMP is smoother than the DMP's trajectory.

Experiments on generalization to moving subjects were conducted to evaluate the adaptive performance of the learning system under dynamic scenarios. During the visual servoing phase, when the subject moves and the neck position changes dynamically, a comparative study was carried out between a traditional nonvisual servoing approach and the proposed visual servoing method for real-time neck tracking. The subject movement process and its corresponding trajectory during visual servoing are illustrated in Fig. 14, while the robot motion trajectories together with the subject motion trajectories are shown in Fig. 15.

The ultrasound scanning system without visual servoing is only capable of reaching the initial neck position and fails to track the target in real time when the subject moves, resulting in an inability to maintain the desired scanning pose. In contrast, by integrating target localization with LfD, the proposed method enables adaptive motion trajectory planning under dynamic conditions, which is crucial for practical medical examination scenarios.

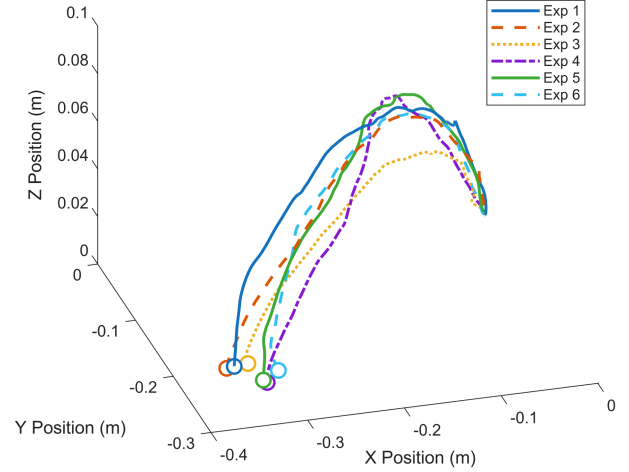


Fig. 16. 3-D trajectories generated by the GPDMP framework across six subjects.

## V. DISCUSSION

The GPDMP method uses GP modeling to encode the forcing term directly from multiple demonstrations. GP-based movement primitives can effectively capture smooth, nonlinear variations among demonstrations without requiring explicit clustering or merging. In the proposed GPDMP model, the following holds.

- 1) The original DMP structure ensures convergence and preserves the benefits of the original formulation.
- 2) The GP representation of  $F(s(t))$  captures the variability and features of multiple demonstrations.
- 3) The method does not depend on perfectly smooth or perfectly aligned demonstrations, as the GP naturally models the distribution of trajectories.
- 4) Multiple demonstrations allow the system to cover a larger portion of the workspace, supporting reliable generalization to new subject pose.

An experimental comparison was conducted between single-demonstration DMPs with scaling and the proposed GPDMP model. Six subjects participated in the experiment, and ten trials were performed for each subject. In this evaluation, the subject was instructed to stand randomly within a 30 cm  $\times$  30 cm square region, and we tested whether the robotic ultrasound probe could successfully perform the initial target registration and acquire a valid ultrasound image. A trial was considered successful if the ultrasound image of the neck artery was correctly detected. As illustrated in Fig. 12, the artery feature in the ultrasound image appears as a circular, indicated by the black circle at the center. The evaluation criteria for ultrasound neck artery scanning were discussed in detail in our previous work [8].

The results show that the GPDMP achieves a 95% task success rate, whereas the single-demonstration DMP with scaling achieves only 75% under varying human postures. The 3-D trajectories generated by GPDMP for six different targets in Fig. 16 and the Cartesian velocities in Fig. 17. The significantly higher success rate of GPDMP demonstrates its stronger generalization capability when adapting to variations in human

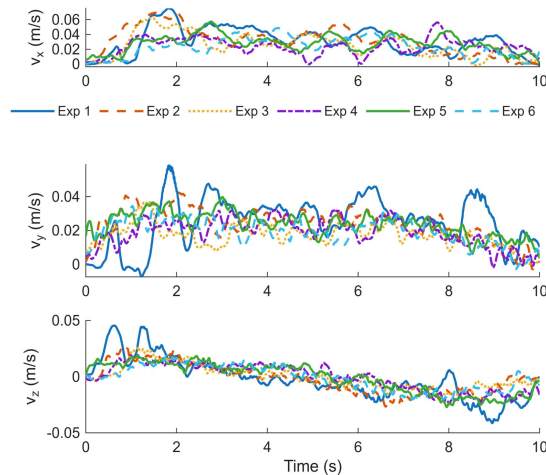


Fig. 17. Cartesian velocity profiles generated by the GPDMP framework for six subjects.

position. While motions may appear scalable across different states, our experiments showed that generalization using a single demonstration led to inconsistent contact behavior and reduced task success rates. GPDMP produces trajectories that retain the key features of human demonstrations while adapting to new start/target positions within the demonstrated workspace, and the success rate is higher than the original DMP with only one demonstration model. Therefore, multiple demonstrations are essential to capture task variability and ensure robust and reliable performance.

## VI. CONCLUSION

In this article, we proposed a robotic learning system with target localization and human demonstration to achieve autonomous robotic sonography. This learning system could recognize and localize human targets robustly and generate trajectories for approaching targets online. The medical examination task of neck artery scan has been taken as an example to validate the performance of the RUS system. In terms of target localization, a vision-based ultrasound scanning localization method, which can automatically identify and locate the human neck, has been proposed. For motion planning, a modeling method named GPDMP was proposed to encode human skills. With the skills learned from GPDMP, we can plan the trajectory of the ultrasound probe when it approaches the human neck without human intervention, even under external disturbance, e.g., subject changing to another clothes, or scanning subjects with different body shapes. This study offers new insights into vision-based interactive ultrasound scanning systems in dynamic real-world situations. In future research, we will investigate the foundation models, e.g., vision-language models and large language models, to advance autonomous robot-assisted scanning for diverse regions of the human body.

## ACKNOWLEDGMENT

The authors would like to thank Jinyan Tian for providing assistance with the machine vision.

## REFERENCES

- [1] Z. Jiang, S. E. Salcudean, and N. Navab, "Robotic ultrasound imaging: State-of-the-art and future perspectives," *Med. Image Anal.*, vol. 89, 2023, Art. no. 102878.
- [2] *Clinical Radiology Census Report 2021*. London, UK: The Royal College of Radiologists, 2022.
- [3] H. Huang et al., "Robust passivity-based dynamical systems for compliant motion adaptation," *IEEE/ASME Trans. Mechatron.*, vol. 27, no. 6, pp. 4819–4828, Dec. 2022.
- [4] Z. Jiang et al., "Autonomous robotic screening of tubular structures based only on real-time ultrasound imaging feedback," *IEEE Trans. Ind. Electron.*, vol. 69, no. 7, pp. 7064–7075, Jul. 2022.
- [5] Y. Fu, W. Lin, X. Yu, J. J. Rodríguez-Andina, and H. Gao, "Robot-assisted teleoperation ultrasound system based on fusion of augmented reality and predictive force," *IEEE Trans. Ind. Electron.*, vol. 70, no. 7, pp. 7449–7456, Jul. 2023.
- [6] Z. Wang et al., "Full-coverage path planning and stable interaction control for automated robotic breast ultrasound scanning," *IEEE Trans. Ind. Electron.*, vol. 70, no. 7, pp. 7051–7061, Jul. 2023.
- [7] J. Zhang et al., "Self-adaptive ultrasound scanning system for imaging human spine," *IEEE Trans. Ind. Electron.*, vol. 69, no. 1, pp. 570–581, Jan. 2022.
- [8] W. Si, N. Wang, R. Harris, and C. Yang, "Deep multimodal imitation learning-based framework for robot-assisted medical examination," *IEEE Trans. Ind. Electron.*, vol. 73, no. 1, pp. 928–936, Jan. 2026.
- [9] D. Zhang, W. Si, W. Fan, Y. Guan, and C. Yang, "From teleoperation to autonomous robot-assisted microsurgery: A survey," *Mach. Intell. Res.*, vol. 19, no. 4, pp. 288–306, 2022.
- [10] Z. Jin, W. Si, A. Liu, W.-a. Zhang, L. Yu, and C. Yang, "Learning a flexible neural energy function with a unique minimum for globally stable and accurate demonstration learning," *IEEE Trans. Robot.*, vol. 39, no. 6, pp. 4520–4538, Dec. 2023.
- [11] M. Saveriano, F. J. Abu-Dakka, A. Kramberger, and L. Peternel, "Dynamic movement primitives in robotics: A tutorial survey," *Int. J. Robot. Res.*, vol. 42, no. 13, pp. 1133–1184, 2023.
- [12] W. Si, N. Wang, and C. Yang, "A review on manipulation skill acquisition through teleoperation-based learning from demonstration," *Cogn. Computation Syst.*, vol. 3, no. 1, pp. 1–16, 2021.
- [13] J. Dong, W. Si, and C. Yang, "A novel human-robot skill transfer method for contact-rich manipulation task," *Robotic Intell. Automat.*, vol. 43, pp. 327–337, 2023.
- [14] L.-H. Kong, W. He, W.-S. Chen, H. Zhang, and Y.-N. Wang, "Dynamic movement primitives based robot skills learning," *Mach. Intell. Res.*, vol. 20, no. 3, pp. 396–407, 2023.
- [15] Z. Jin, A. Liu, W.-a. Zhang, L. Yu, and C. Yang, "Gaussian process movement primitive," *Automatica*, vol. 155, 2023, Art. no. 111120.
- [16] Z. Liu, K. Wang, D. Liu, Q. Wang, and J. Tan, "A motion planning method for visual servoing using deep reinforcement learning in autonomous robotic assembly," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 6, pp. 3513–3524, Dec. 2023.
- [17] X. Ma, Z. Zhang, and H. K. Zhang, "Autonomous scanning target localization for robotic lung ultrasound imaging," in *Proc. 2021 IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 9467–9474.
- [18] X. Ma, W.-Y. Kuo, K. Yang, A. Rahaman, and H. K. Zhang, "A-see: Active-sensing end-effector enabled probe self-normal-positioning for robotic ultrasound imaging applications," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 12475–12482, Oct. 2022.
- [19] X. Ma, M. Zeng, J. C. Hill, B. Hoffmann, Z. Zhang, and H. K. Zhang, "Guiding the last centimeter: Novel anatomy-aware probe servoing for standardized imaging plane navigation in robotic lung ultrasound," *IEEE Trans. Autom. Sci. Eng.*, vol. 22, pp. 6569–6580, 2025.
- [20] Y. Xie et al., "Robot-assisted trans-esophageal ultrasound and the virtual admittance-based master-slave control method thereof," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 5, pp. 2505–2516, Oct. 2023.
- [21] M.-S. Pan, J.-T. Tang, and X.-L. Yang, "Movement invariants-based algorithm for medical image tilt correction," *Int. J. Automat. Comput.*, vol. 7, no. 4, pp. 543–549, 2010.
- [22] R. Tsumura, Y. Koseki, N. Nitta, and K. Yoshinaka, "Towards fully automated robotic platform for remote auscultation," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 19, no. 1, 2023, Art. no. e2461.
- [23] M. Tirindelli et al., "Force-ultrasound fusion: Bringing spine robotic-us to the next 'level'," *IEEE Robot. Automat. Lett.*, vol. 5, no. 4, pp. 5661–5668, Oct. 2020.

- [24] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
- [25] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1440–1448, doi: [10.1109/ICCV.2015.169](https://doi.org/10.1109/ICCV.2015.169).
- [26] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
- [27] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
- [28] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.
- [29] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [30] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [31] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. 2001 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 1–1.
- [32] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *Proc. 2008 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
- [33] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "RMPE: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2334–2343.
- [34] M. Jaderberg et al., "Spatial transformer networks," *Proc. 29th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 2017–2025.
- [35] K. S. Saladin, *Human Anatomy*. Manila, Philippines: Rex Bookstore, Inc., 2005.
- [36] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: Learning attractor models for motor behaviors," *Neural Computation*, vol. 25, no. 2, pp. 328–373, 2013.
- [37] C. E. Rasmussen et al., *Gaussian Processes for Machine Learning*, vol. 1. Berlin, Germany: Springer, 2006.
- [38] Y. Lin, Z. Chen, and B. Yao, "Unified motion/force/impedance control for manipulators in unknown contact environments based on robust model-reaching approach," *IEEE/ASME Trans. Mechatron.*, vol. 26, no. 4, pp. 1905–1913, Aug. 2021.
- [39] S. Sidhik, M. Sridharan, and D. Ruiken, "Hybrid models for control of changing-contact manipulation tasks," in *Proc. 8th Annu. Conf. Adv. Cogn. Syst.*, Menlo Park, CA, USA, 2020.
- [40] X. Xing, E. Burdet, W. Si, C. Yang, and Y. Li, "Impedance learning for human-guided robots in contact with unknown environments," *IEEE Trans. Robot.*, vol. 39, no. 5, pp. 3705–3721, Oct. 2023.



**Weiyong Si** (Member, IEEE) received the M.S. degree in aerospace engineering from the Beijing Institute of Technology, Beijing, China, in 2018, and the Ph.D. degree in robotics from the University of the West of England, Bristol, U.K., in 2023.

He is a Lecturer in robotics and AI with the University of Essex, Colchester, U.K. His research interests include robot learning, robot control, and teleoperation.



**Zhehao Jin** (Member, IEEE) received the B.Eng. degree in automation and the Ph.D. degree in control theory and control engineering from the Zhejiang University of Technology, Hangzhou, China, in 2018 and 2024, respectively.

He is currently a Research Fellow with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. His research interests include robot learning and control.



**Ning Wang** (Member, IEEE) received the M.Phil. and Ph.D. degrees in electronics engineering from The Chinese University of Hong Kong, Hong Kong, in 2007 and 2011, respectively.

She is an Associate Professor in robotics and AI with Sheffield Hallam University, Sheffield, U.K. Her research focuses on robotics and intelligent systems, with a primary focus on human–robot interaction, robot learning, and signal processing.

Dr. Wang has been a key member of EU FP7 Project ROBOT-ERA, EU Regional Development Funded Project ASTUTE2020 and industrial projects with U.K. companies. She was the recipient of several awards, including IET premium award for best paper 2022.



**Chenguang Yang** (Fellow, IEEE) received the B.Eng. degree in measurement and control from Northwestern Polytechnical University, Xi'an, China, in 2005, and the Ph.D. degree in control engineering from the National University of Singapore, Singapore, in 2010.

From 2009 to 2010, he performed postdoctoral studies on motor control in human robotics group with Imperial College London, London, U.K. He is now a full Professor with Department of Computing, The Hong Kong Polytechnic University, Hong Kong, HKSAR. He was leading the Robot Teleoperation Group, Bristol Robotics Laboratory, Bristol, U.K. His research interests include robot control and learning, human–robot interaction, and intelligent system design.

Dr. Yang is a Member of European Academy of Sciences and Arts, and he is also recognized as a Fellow by several prestigious institutions, including the Institute of Electrical and Electronics Engineers (IEEE), Institute of Engineering and Technology, Institution of Mechanical Engineers, Asia-Pacific AI Association, and British Computer Society. He was the recipient of U.K. EPSRC UKRI Innovation Fellowship and individual EU Marie Curie International Incoming Fellowship. He previously served as the President of the Chinese Automation and Computing Society in the U.K. and has organized several conferences as the general chair, including the 25th IEEE International Conference on Industrial Technology and the 27th International Conference on Automation and Computing, Bristol. As the lead author, he was the recipient of the IEEE TRANSACTIONS ON ROBOTICS Best Paper Award (2012) and IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS Outstanding Paper Award (2022).