Implementation Details of Our Visual Servoing Baseline

We evaluated UltraDP alongside a rule-based Visual Servo (VS) method modified from [1], [2]. Note that in [1] the first level task of the controller was defined in Cartesian space, rather than ultrasound image space, and thus was unable to perform a visual servoing task. So we made one modification. **Our Modification: We have introduced an Image Space term as the primary hierarchical level, followed by Cartesian Space as the second level, and Joint Space as the third.** We formulated a term in image space

$$\boldsymbol{t}_3 = \boldsymbol{J}_o^T k(u - u_d) \in \Re^{7\times 1}, \tag{1}$$

where $\boldsymbol{J}_o \in \Re^{1\times 7}$ is the overall Jacobian from joint space to image space, and k is the proportional coefficient, u is the pixel position of the artery, which is the output of our regression network. The overall Jacobian is obtained by

$$\begin{aligned}\boldsymbol{J}_o &= \boldsymbol{J}_i \boldsymbol{A} \boldsymbol{J} \\ &= [0, \frac{1}{a}, 0, 0, 0, 0] \begin{bmatrix} {}^{base}_{ee}\boldsymbol{R} & \boldsymbol{0} \\ \boldsymbol{0} & {}^{base}_{ee}\boldsymbol{R} \end{bmatrix} \boldsymbol{J}\end{aligned} \tag{2}$$

where $\boldsymbol{J}_i \in \Re^{1\times 6}$ is the jacobian matrix of image space with respect to Cartesian space in the end effector frame, derived from the equation $\Delta y = a\Delta u$, and $\boldsymbol{A} = diag([{}^{base}_{ee}\boldsymbol{R}, {}^{base}_{ee}\boldsymbol{R}])$ is the adjacent matrix from base frame to end effector frame.

Hence, we have the hierarchical task definition:

$$\boldsymbol{J}_1 = \boldsymbol{J}_o \tag{3}$$

$$\boldsymbol{J}_2 = \boldsymbol{J} \tag{4}$$

$$\boldsymbol{J}_3 = diag([1, 0, 0, 0, 0, 0, 0]) \tag{5}$$

At last, we follow the same line in [1] and obtain our controller:

$$\boldsymbol{\tau} = \boldsymbol{g} + \boldsymbol{\tau}_d + \boldsymbol{\tau}_1 + \boldsymbol{\tau}_2 + \boldsymbol{\tau}_3 - \boldsymbol{\tau}_e, \tag{6}$$

Note that the controller's performance was highly dependent on the accuracy of the regression network, as the primary task operates in image space. If the network failed to recognize the artery, the controller was unable to function correctly. We show some failure cases of our network in Fig. 1.

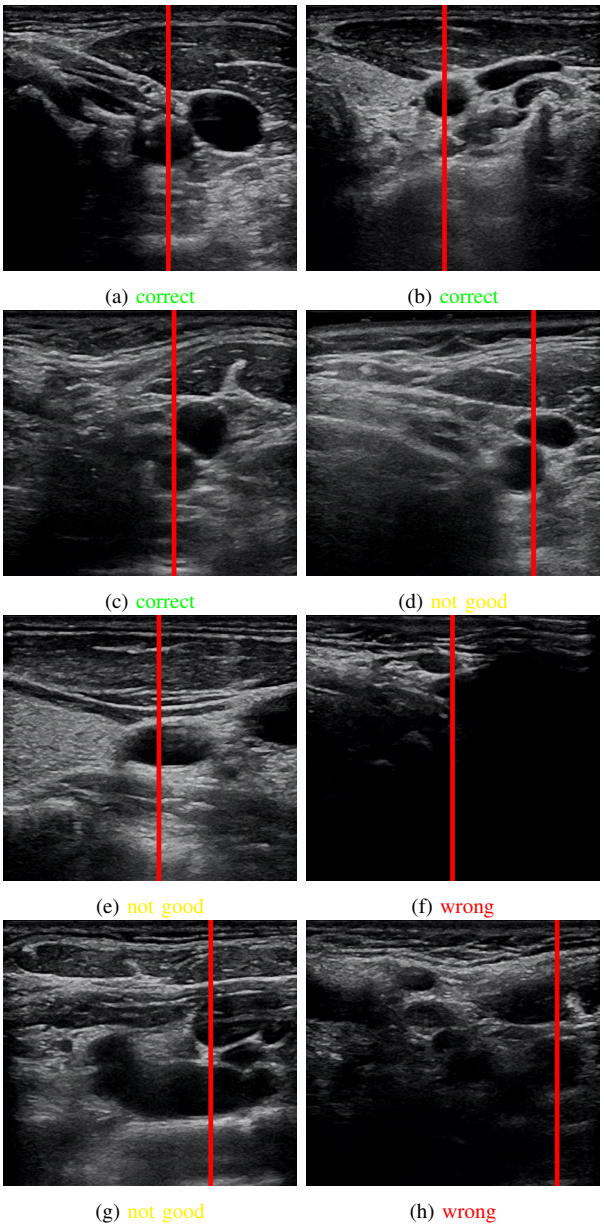

Fig. 1. Illustrations of failure cases in our regression network results: (a)–(c): The network successfully detected the artery. (d)–(e): The network detected the artery, but the results were suboptimal. (f): The probe detached from the neck, causing the artery to disappear, making recognition impossible. (g): Imperfect probe orientation distorted the artery's contour, leading to poor recognition. (h): A black lower region in the image prevented the network from recognizing the artery.

References

[1] X. Yan, S. Luo, Y. Jiang, M. Yu, C. Chen, S. Zhu, G. Huang, S. Song, and X. Li, "A unified interaction control framework for safe robotic ultrasound scanning with human-intention-aware compliance," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 14004–14011, IEEE, 2024.

[2] C. Ott, A. Dietrich, and A. Albu-Schäffer, "Prioritized multi-task compliance control of redundant manipulators," *Automatica*, vol. 53, pp. 416–423, 2015.