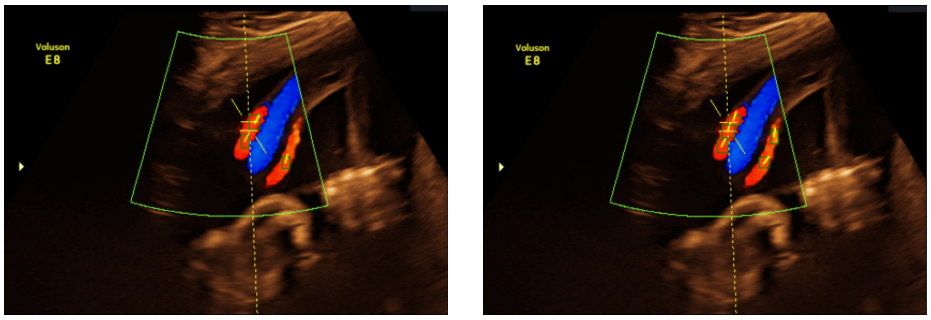
Appendix 1 Evaluating Model's Performance



(A) ground truth (B) prediction

Fig. I: Manually-annotated (A) versus model-predicted (B) bounding boxes and angles. The correct anatomical sites are sparsely annotated due to laborious nature of the task, which means not every correct anatomical sites are annotated in our ground truth dataset.

Since the correct anatomical sites are under-annotated in our dataset (see Fig. I), we are unable to determine the false-positive in our predictions and, subsequently, the precision of our model. Hence, we chose to evaluate our model performance with a sensitivity-based approach. For every ground-truth box (see Fig. IA), we identify the corresponding nearest predicted box (see Fig. IB) by determining box pair that yields the shortest Euclidean distance:

$$f(\mathbf{t_i^*}) = \mathbf{t_{\tilde{j}}}, \quad \tilde{j} = \underset{j}{\operatorname{argmin}} \|\mathbf{t_i^*} - \mathbf{t_j}\|_2$$

where $\mathbf{t_i^*}$ represents centroid of the i-th ground-truth box, $\mathbf{t_j}$ centroid of the j-th predicted box, $\tilde{j}$ the index of the nearest predicted box, and f the mapping function. Then, for a given threshold n, we say the ground truth box is successfully detected if its nearest predicted box is less n pixels away:

$$T \coloneqq \{\, \mathbf{t_i^*} \,\}$$
$$\widetilde{T} \coloneqq \begin{cases} \{\, \mathbf{t_i^*} \mid \|\mathbf{t_i^*} - \mathbf{t_{\tilde{j}}}\|_2 < n \,\}, & \text{if } \{\, \mathbf{t_j} \,\} \neq \varnothing \\ \varnothing, & \text{otherwise} \end{cases}$$

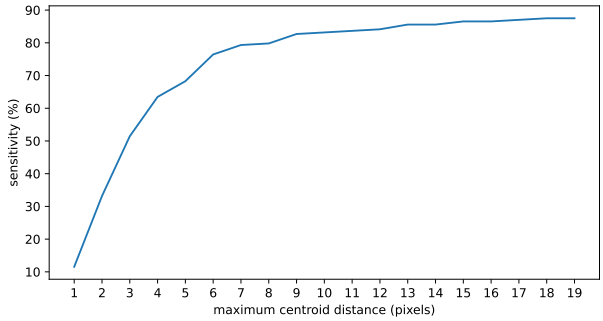
Finally, the sensitivity of our model is derived by calculating the percentage of successfully detected ground truth box:

$$\textit{sensitivity} = \frac{|\widetilde{T}|}{|T|} \times 100\%$$

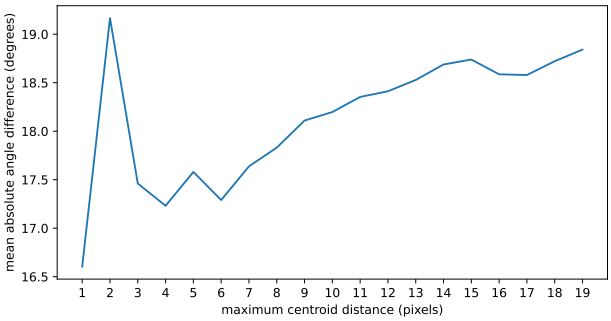
while the inaccuracy in angle prediction is estimated using the mean L_1 distance between the ground truth angles $\{ a_i^* \}$ and the predicted angles $\{ a_j \}$:

$$\Delta A = \{ \| a_i^* - a_{\tilde{j}} \| \}$$
$$\bar{L}_1 = \frac{\sum \Delta A}{|\Delta A|}$$

As expected, both the sensitivity and the angle inaccuracy increased with a larger threshold n (see Fig. II). Since a higher sensitivity and lower angle inaccuracy is desired, we reported the values with n fixed to 10 (see Sec. 4.2 of the main article) to avoid over-glorifying our model in either of the two performance metrics.



(A) *sensitivity* against n (the higher the better)



(B) $\bar{L}_1$ against n (the lower the better)

Fig. II: Evaluation of model's performance