Supplementary Material

1 Hyperparameters

Neural network architecture We adapted the default architecture in [2] for our U-Net marker detection module. The difference was that we only used three downsampling operations to prevent overfitting. We used ResNet-18 [1] for the object feature extraction module. For the GCN tracking module, we used the PyTorch Geometric library [1]. The GCN branch contains a *GCNConv* layer (1024 input channels and 256 output channels), a *EdgeConv* layer (512 input channels and 128 output channels), and a *EdgeConv* layer (256 input channels and 64 output channels), with a *ReLU* layer after each layer. The FC branch has one fully-connected layer with 1024 input channels and 256 output channels. Then outputs of GCN branch and FC branch were combined and passed to a fully-connected layer with 320 input channels and 1 output channel.

Loss function In Eq. 8, $\lambda_1 = 1, \lambda_2 = 2$. In Eq. 9, $\alpha = 1, \beta = 2$.

2 Results

We used a fixed threshold cutoff (0.6) for identifying positive predictions (i.e. markers with probability greater than 0.6 were selected) to calculate the evaluation metrics shown in Table 1, 2, and S3. Here, we also demonstrate the performance of CenterTrack and our method (Table S1 and Table S2) when two markers with the highest probabilities in each frame were identified as markers detected, since it is the most straightforward way to identify a single stent based on the outputs of neural networks. Fig. S1 shows examples of tracking results from CenterTrack and the proposed method with this top-2 selection criterion, and the stent enhancement results based on the tracking results. It can observed that false positives dramatically affect enhancement results (Fig. S1b) and our method has a high precision score (Table S1 and Table S2), which demonstrates its robustness in clinical applications. Lower two rows of Fig. S1 show results of CenterTrack and the proposed method with representative MAEs (0.382 and 0.511) respectively. It can be observed that even though the MAE of our method is worse than that of CenterTrack, the enhancement results do not show much difference.

[1] https://pytorch-geometric.readthedocs.io/en/latest/

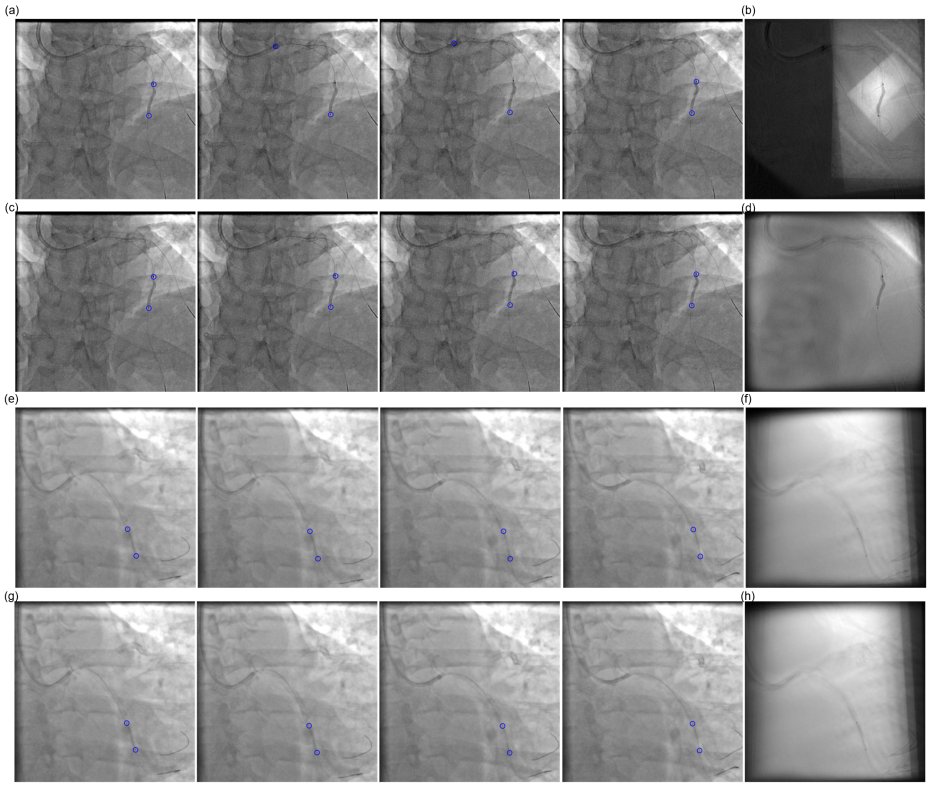

Fig. S1. Example tracking results (4 frames) from CenterTrack (a,e) and the proposed method (c,g), and the corresponding stent enhancement results (b,f) and (d,h). 7 frames were used for enhancement in all cases.

Table S1. Evaluations on **In-house Dataset** (top 2). CR means coordinate regression model, and CT means CenterNet. ↑ indicates that higher is better, ↓ indicates that lower is better.

Model		Detection				Localization	
Type	Backbone	Precision↑	Recall↑	F1↑	Accuracy↑	MAE↓	RMSE↓
CT	MobileNetV2	0.752	0.803	0.777	0.635	0.443	0.827
	DLA34	0.813	0.805	0.809	0.679	**0.391**	**0.742**
Ours		**0.979**	**0.882**	**0.928**	**0.866**	0.502	0.891

Table S2. Evaluations on **TAVI Dataset** (top 2).

Model		Detection				Localization	
Type	Backbone	Precision↑	Recall↑	F1↑	Accuracy↑	MAE↓	RMSE↓
CT	MobileNetV2	0.919	0.905	0.907	0.831	**5.172**	**6.054**
	DLA34	0.927	0.896	0.911	0.837	5.415	6.150
Ours		**0.986**	**0.915**	**0.949**	**0.903**	5.966	6.705

Table S3. Evaluations on **TAVI Dataset** with cross validation. CR means coordinate regression model, and CT means CenterNet. All the reported values are mean values from 5-fold cross validation. ↑ indicates that higher is better, ↓ indicates that lower is better.

Model		Detection				Localization	
Type	Backbone	Precision↑	Recall↑	F1↑	Accuracy↑	MAE↓	RMSE↓
CR	MobileNetV2	0.851	0.741	0.7892	0.656	13.44	14.56
	ResNetV2	0.861	0.799	0.829	0.709	12.145	13.168
CT	MobileNetV2	0.751	**0.934**	0.832	0.713	**5.276**	**6.248**
	DLA34	0.819	0.927	0.869	0.768	5.362	6.490
Ours		**0.913**	0.902	**0.901**	**0.820**	5.802	6.524

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
2. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)