Robust Landmark-based Stent Tracking in X-ray Fluoroscopy

Luojie Huang[1]*, Yikang Liu[2], Li Chen[3]*, Eric Z. Chen[2], Xiao Chen[2], and Shanhui Sun[2] **

[1] Johns Hopkins University, Baltimore, MD, USA
[2] United Imaging Intelligence, Cambridge, MA, USA shanhui.sun@uii-ai.com
[3] University of Washington, Seattle, WA, USA

Abstract. In clinical procedures of angioplasty (i.e., open clogged coronary arteries), devices such as balloons and stents need to be placed and expanded in arteries under the guidance of X-ray fluoroscopy. Due to the limitation of X-ray dose, the resulting images are often noisy. To check the correct placement of these devices, typically multiple motion-compensated frames are averaged to enhance the view. Therefore, device tracking is a necessary procedure for this purpose. Even though angioplasty devices are designed to have radiopaque markers for the ease of tracking, current methods struggle to deliver satisfactory results due to the small marker size and complex scenes in angioplasty. In this paper, we propose an end-to-end deep learning framework for single stent tracking, which consists of three hierarchical modules: a U-Net for landmark detection, a ResNet for stent proposal and feature extraction, and a graph convolutional neural network for stent tracking that temporally aggregates both spatial information and appearance features. The experiments show that our method performs significantly better in detection compared with the state-of-the-art point-based tracking models. In addition, its fast inference speed satisfies clinical requirements.

Keywords: Stent enhancement, Landmark tracking, Graph neural network

1 Introduction

Coronary artery disease (CAD) is one of the primary causes of death in most developed countries [19]. The current state-of-the-art treatment option for blocked coronary arteries is the percutaneous coronary intervention (PCI) (Fig. 1). During this minimally invasive procedure, a catheter with a tiny balloon (the tracked dark object in Fig. 1c) at the tip is put into a blood vessel and guided to the blocked coronary artery. Once the catheter arrives at the right place, the balloon is inflated to push the artery open, restoring room for blood flow. In most cases, a stent, which is a tiny tube of wire mesh (Fig. 1), is also placed in the

* Work was done during an internship at United Imaging Intelligence America.
** Corresponding author.

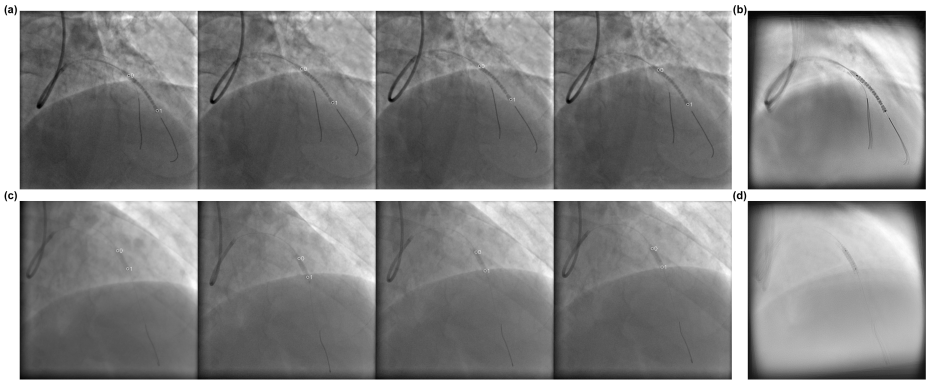

Fig. 1. Examples of stent tracking with the proposed method (a,c) and stent enhancement based on the tracking results (b,d). a and c show four frames from a video.

blocked artery after the procedure to support the artery walls and prevent them from re-narrowing. Intraoperative X-ray fluoroscopy is commonly used to check the location of stent/balloon before expansion. However, stent visibility is often limited (Fig. 1a and c) under X-ray because the minimal level radiation dose out of safety concerns. Furthermore, stents keep moving rapidly with heartbeat and breathing in the complicated environment of patients' anatomy.

Compared to other physical approaches, such as invasive imaging or increasing radiation dose, a more cost-effective solution is to enhance the stent appearance through image processing (a.k.a., digital stent enhancement), as shown in Fig. 1b and d. A common method is to track the stent motion, separate the stent layer from the background layer, and average the stent layers from multiple frames after motion compensation. Stent tracking is achieved by tracking two radiopaque balloon markers that locate at two ends of the stent (Fig. 1).

Stent tracking for enhancement remains quite challenging due to multiple reasons. First, the balloon markers are very small compared to the whole field of view (FOV) while the movements are large. Second, the scenes in PCI procedures are very complex: organs and other devices can form a noisy background and 3D organs and devices can be projected into different 2D images from different angles. Third, stent enhancement requires high localization accuracy and low false positives. Fourth, fast tracking speed is needed to meet the clinical requirements (e.g., 15fps for a 512x512 video). Fifth, data annotations are limited just like other medical imaging applications.

Current stent tracking methods lie under the tracking-by-detection category and assume only one stent presents in the FOV. They first detect all possible radiopaque balloon marker from each frame, and then identify the target stent track based on motion smoothness [2,3] or consistency score [14,23]. However, these methods are prone to large detection and tracking errors caused by strong false alarms. Deep learning techniques dramatically improve detection and tracking accuracy. However, it is difficult to apply these techniques to the

stent tracking problem because of the the small object size and complex PCI scenes and overfitting issue on small dataset. Therefore, we tackle the above issues by incorporating some basic prior knowledge into our framework design. For example, the stent has distinctive hierarchical features: two dark markers that can be detected by low-level features and complicated patterns between the marker pairs, such as wire, mesh and balloon tubes, to be recognized from high-level semantic analysis. Additionally, the association of marker pairs in different frames requires long temporal dimension reasoning to tolerate inaccurate detections in certain frames with limited image quality. Moreover, most deep learning frameworks for keypoint tracking problems (such as human pose tracking) train detection and tracking modules separately. However, it is generally harder to detect small markers from a single X-ray image than keypoints from natural images due to the limited object features and complex background.

Therefore, we propose an end-to-end trainable deep learning framework that consists of three hierarchical modules: a landmark detection module, which is a U-Net [21] trained with small patches to detect potential balloon markers with local features; a stent proposal and feature extraction module, which is a ResNet [11] trained with larger stent patches to extract high-level features located between detected marker pairs; and a stent tracking module, which is a graph convolutional neural network (GCN) to associate marker pairs across frames using the combination of extracted features and spatial information. Our ablation study demonstrated that end-to-end learning of the whole framework can greatly benefit the performance of final stent tracking. For example, detector can be boosted by incorporating the feedback from trackers by learning to suppress some false positives that cause bad outcome in trackers.

In summary, the major contributions of this paper are as follows: 1) We propose the first deep learning method, to our best knowledge, to address the single stent tracking problem in PCI X-ray fluoroscopy. 2) To handle the challenge of tracking small stents in complex video background, we propose an end-to-end hierarchical deep learning architecture that exploits both local landmarks and general stent features by CNN backbones and achieves spatiotemporal associations using a GCN model. 3) We test the proposed method and several other state-of-the-art (SOTA) models on both public and private datasets, with hundreds of X-ray fluoroscopy videos, data of a scale that has not been reported before. The proposed method shows superior landmark detection, as well as frame-wise stent tracking performance.

2 Related Work

2.1 Digital Stent Enhancement

The digital stent enhancement (DSE) algorithm typically follows a bottom-up design and can be generally divided in to 4 main steps: landmark detection, landmark tracking, frame registration, and enhanced stent display.

First, the region of interest needs to be located from each frame. Due to the limited visibility of stent and large appearance variation between folded and

expanded stents, it is challenging to extract the stent directly. Instead, potential landmarks such as the radiopaque balloon marker pairs at two ends of the stent or the guidewire in between is more commonly used to indicate the location of the stent. Throughout the X-ray image sequence, the stent location is constantly changing with cardiac and breathing motions. Based on the landmark candidates, the most promising track needs to be identified to associate the target stent across frames. Next, frames can be registered based on motion inferred by the landmark trajectory. The motion compensation is often performed by aligning tracked landmarks together using rigid registration [15,7] or elastic registration [2]. To enhance stent visualization, the stent layer is extracted while the background is suppressed [8].

2.2 Balloon Marker Detection

Two markers on the balloon used to deliver the stent are considered the most prominent feature of the stent structure due to the consistent ball shape and radio-opacity from high absorption coefficient. Various strategies are previously studied to achieve efficient balloon marker detection.

Conventional image processing methods are applied, including match filtering or blob detection, to extract candidate markers from the X-ray image. Bismuth et al.[3] proposed a method involving a priori knowledge and dark top hat preprocessing to detect potential markers from local minimum selection. Blob detectors [18,13,23] locate markers by differentiating regions with unique characteristics from neighborhood, such as brightness or color. However, the extremely small size of balloon markers and common noises from the background, such as guidewire tips, Sharp bone edges and other marker-like structure, make those methods prone to a high false positive rate.

Learning-based methods are also proposed to incorporate more extended context information for better markers localization. Lu et al.[14] used probabilistic boosting trees combining joint local context of each marker pair as classifiers to detect markers. Chabi et al.[6] detected potential markers based on adaptive threshold and refined detections by excluding non-mask area using various machine learning classifiers, including k-nearest neighbor, naive Bayesian classifier, support vector machine and linear discriminant analysis. Vernikouskaya et al.[21] employed U-Net, a popular encoder-decoder like CNN designed specifically for medical images, to segment markers and catheter shafts during pulmonary vein isolation as binary masks. The maker segmentation performances from the above methods are still limited by the super imbalance between foreground and background areas. Moreover, all the candidate refinements only focus on considering more local context information at single frame, while the temporal correlation has never been exploited to enhance the classifiers.

In our work, balloon marker detection is considered as a heatmap regression task, which has shown superior performances in other landmark detection applications, such as face recognition [25], human pose estimation [20] and landmark detection in various medical images [10,1]. To obtain potential markers, we use

U-Net as the heatmap regression model, which represents each landmark as a 2D Gaussian distribution for more accurate localization.

2.3 Graph Based Object Tracking

Given the set of marker candidates across X-ray image sequence, either a priori motion information [3] or a heuristic temporal coherence analysis [14], which calculates consistency score between frames base on predefined criteria, is used to identify the most prominent landmark trajectory. Wang et al.[23] proposed a offline tracking algorithm as graph optimization problem, by constructing a trellis graph with all the potential marker pairs and then employed the Viterbi algorithm to search the optimal path across frame from the graph. Similar graph models are also applied to other general object tracking tasks [22,5,16] as min-cost flow optimization problem. However, these static graph models will fail when the information contained by nodes or edges is not representative enough or outdated. Brasó et al.[4] demonstrated superior results in multiple object tracking by constructing a dynamical graph of object detections and updating node and edge features using GCN.

In this work, we first interpret the whole video into a graph, where the nodes are associated with encoded appearance features of potential stent from marker pair detections and edges are temporal coherency across frames. A graph neural network is trained as a node classification model to update both node and edge features via message passing. The stent tracking is achieve by learning both context and temporal information.

To our knowledge, the proposed CNN-GCN based DSE algorithm is the first deep learning model to achieve robust balloon marker tracking and 2D stent visual enhancement by incorporating both extended context and temporal information.

3 Approach

In this work, we propose an effective end-to-end trainable framework for landmark based stent/balloon tracking (Fig. 2) with a hierarchical design: a U-Net based landmark detection module that generates a heatmap to localize marker candidates with local features, a ResNet based stent proposal and feature extraction module to extract global stent features in a larger context, and a GCN based stent tracking module to identify the real stent by temporally reasoning with stent features and marker locations.

3.1 Landmark Detection

In landmark based stent/balloon detection, each candidate object is represented by a detected landmark pair: $\mathcal{O}_i = (\mathcal{D}^L_{i1}, \mathcal{D}^L_{i2})$. The first step is to detect landmarks from each frame using a U-Net [21]. In contrast to conventional object detection, the major challenge of landmark detection is the highly unbalanced

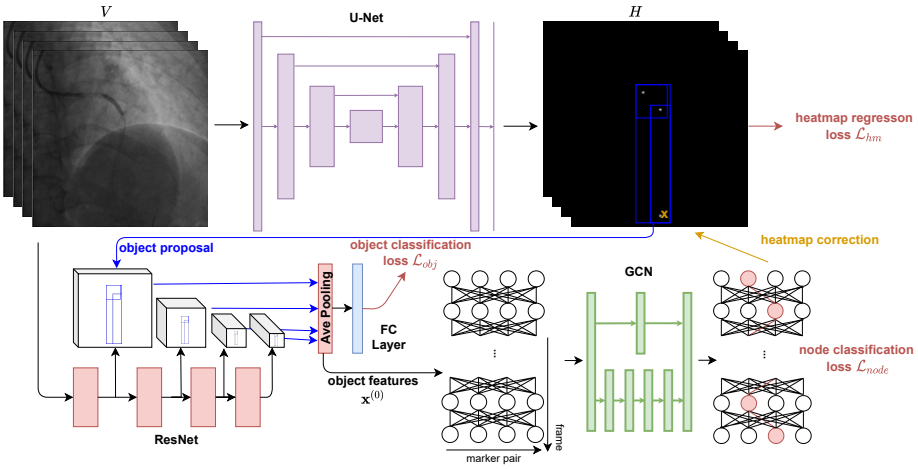

Fig. 2. The proposed end-to-end deep learning framework for stent tracking.

foreground/background ratio, as landmarks are commonly tiny dots of few pixels compared to the frame size. Therefore, we treat landmark detection as a heatmap regression problem and pretrain the detector with smaller positive landmark patches, thus to increase fore-to-background ratio. The input video $\mathbf{V} \in \mathbb{R}^{T\times H\times W\times C}$ is fed into the landmark detector (U-Net) that generates heatmaps $\mathbf{H} \in \mathbb{R}^{T\times H\times W}$, where detected landmarks are represented as 2D Gaussian distributed points. From a predicted heatmap, peak points are extracted as landmark detections, represented as 2D coordinates and a confidence score: $\mathcal{D}_i^L = (x_i^L, y_i^L, s_i^L)$. During training, an false negative regularization(**Sec 3.4**) is implemented to further enforce the detector to focus on landmarks.

With an ideal landmark detection, the target stents can be directly located by landmarks and tracked over time with simple temporal association. However, due to the lack of extended context information for perfect landmark localization, the landmark detector is inevitably limited by a high false positive rate which further hinders stent tracking. Hence, we proposed a delicate pipeline to simultaneously refine object detection and tracking.

3.2 Stent Proposal & Feature Extraction

Given a set of landmark detections $\mathcal{D}_t^L$ at frame (t), candidate objects can be formed by all possible combination of landmark pairs $\mathcal{O}_t = \{(\mathcal{D}_{ti}^L, \mathcal{D}_{tj}^L) \mid \mathcal{D}_{ti}^L, \mathcal{D}_{tj}^L \in \mathcal{D}_t^L\}$, where $\mathcal{D}_{ti}^L, \mathcal{D}_{tj}^L$ denote the ith and jth Landmark Detections in $\mathcal{D}_t^L$ at frame (t). As the landmark pair is always located at two ends of the corresponding object, We can assign a confidence score to the candidate object using the average of landmark confidence scores:

$$S_i^{\mathcal{O}} = \frac{1}{2}(S_{i1}^L + S_{i2}^L). \tag{1}$$

We can generate a rectangular bounding box for the object based on the landmark locations, of which the center is the middle point of the landmark pair and side lengths are the distance between the landmarks along the corresponding axis. A ResNet is used to extract appearance features of candidate objects. The outputs of ResNet at multiple levels within the corresponding bounding boxes were averaged and stored into a D-dimension feature vector $\mathbf{x}^{(0)} \in \mathbb{R}^D$ for each candidate object, which are used later in GCN for temporal reasoning (Fig. 2). In addition, to facilitate feature learning with a deep supervision, we feed the feature vector into a fully-connected layer and use a weighted cross-entropy loss ($\mathcal{L}_{obj}$ in Eq. 9) between the its outputs and labels indicating if the proposed bounding box contains the object of interests.

3.3 Stent Tracking

With the object candidates at every frame, we first construct an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ across all frames, where vertices $\mathcal{V}$ represent candidate objects proposed by detected landmark pairs and edges $\mathcal{E}$ are full connections of candidate objects between adjacent frames. Every object at frame t is connected with all the candidate objects at frame $(t-1)$ and frame $(t+1)$.

The attributes of vertices are the appearance feature vectors $\mathbf{x}^{(0)}$ extracted from the feature extractor. The edge weights in the initial graph are calculated as a weighted combination of object confidence scores and the spatial similarity by comparing sizes, rotations and locations of objects:

$$\mathbf{w}_{i,j} = \frac{S_i^{\mathcal{O}} + S_j^{\mathcal{O}}}{2}(\alpha_1 IoU(\mathcal{O}_i, \mathcal{O}_j) + \alpha_2 AL(\mathcal{O}_i, \mathcal{O}_j)), \tag{2}$$

where α_1, α_2 are weighting factors, $IoU(\cdot)$ is the IoU between object bounding boxes and $AL(\cdot)$ measures the objects similarity by comparing angles and lengths of the landmark pair vector, defined as:

$$AL(\mathcal{O}_i, \mathcal{O}_j)) = \max(0, \frac{|\vec{v}_i \cdot \vec{v}_j|}{|\vec{v}_i||\vec{v}_j|} - \frac{||\vec{v}_i| - |\vec{v}_j||}{\sqrt{|\vec{v}_i||\vec{v}_j|}}). \tag{3}$$

Here, $\vec{v}_i$,$\vec{v}_j$ are the 2D vectors between landmark pairs of $\mathcal{O}_i$ and $\mathcal{O}_j$.

The initialized graph is a powerful representation of the whole video for object tracking, as the appearance features are embedded into vertices and spatiotemporal relations are embedded in the edge weights. To track objects over time, we perform node classification on the graph using a GCN, which identifies the tracked objects at different frames as positive nodes of a corresponding object class while false object detections and untracked objects are classified as negative nodes.

The tracking model is a GCN with a full connected bypass (Fig. 2). The GCN branch consists of a weighted graph convolution layer (wGCL) [12] and two edge convolution layers (ECLs) [24]. Weighted graph convolution layer with self-loop is defined as:

$$\mathbf{x}_i = \mathbf{\Theta} \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{\mathbf{w}_{j,i}}{\hat{d}_i} \mathbf{x}_j^{(0)}, \tag{4}$$

with a normalization term $\hat{d}_i = 1 + \sum_{j \in \mathcal{N}(i)} \mathbf{w}_{j,i}$, where $\mathbf{w}_{j,i}$ denotes the edge weight between node j and node i.

Within an edge convolution layer, the edge features are first updated by a FC layer with the features of corresponding vertices pairs connected with each edge:

$$\mathbf{e}_{i,j} = h_{\mathbf{\Theta}}(\mathbf{x}_i, \mathbf{x}_j), \tag{5}$$

where $h_{\mathbf{\Theta}}$ is a nonlinear function with learnable parameters $\mathbf{\Theta}$. Then, ECL updates node features by the summation of updated edge features associated with all the edges emanating from each vertex:

$$\mathbf{x}_i^{EC} = \sum_{j \in \mathcal{N}(i)} \mathbf{e}_{i,j}. \tag{6}$$

The GCN branch effectively updates features of candidate objects by most similar objects from adjacent frames. Moreover, a sequence of convolution layers enables information propagation from even further frames. However, node features solely updated from the GCN are susceptible to noisy neighborhood. For example, if the target object is missed by the upstream detection at a certain frame, such errors would propagate to nearby frames and thus worsen general tracking performance. Therefore, we add a simple parallel FC bypass to the GCN branch. In the FC bypass, all the node features are updated independently without influence from connected nodes:

$$\mathbf{x}_i^{FC} = h_{\mathbf{\Theta}}(\mathbf{x}_i^{(0)}). \tag{7}$$

In the last layer, node features from the GCN branch $\mathbf{x}^{EC}$ are enhance by the FC bypass outputs $\mathbf{x}^{FC}$ for robust object tracking.

Heatmap correction GCN results are then used to correct the heatmaps generated in landmark detection. Specifically, we multiply heatmap values in the a $w \times w$ window centered around a detected landmark with the maximum probability of the graph nodes containing the marker. In this way, the landmark detector can ignore the false positives that can be easily rejected by the GCN model and increase detection sensitivity.

3.4 Training

The landmark detector was trained as a heatmap regression model. Since landmark detection results are used for object proposal, feature extraction, and graph construction, missed landmark would cause irreversible corruption to tracking as the missed object cannot be recovered, while false positives can be filtered out during object proposal or node classification. We used a modified cost term $\mathcal{L}_{hm}$ to ensure fewer false negatives, defined as:

$$\mathcal{L}_{hm} = \frac{\lambda_1}{N} \sum_{i=0}^{N} (y_i - \hat{y}_i)^2 + \frac{\lambda_2}{N} \sum_{i=0}^{N} (ReLU(y_i - \hat{y}_i))^2, \tag{8}$$

where λ_1, λ_2 are weighting factors, y_i, $\hat{y}_i$ are pixel intensities from ground truth and predicted heatmap (corrected by GCN outputs), respectively.

The feature extractor and GCN tracking model are trained as classification problems. We use weighted cross entropy as the cost function to handle the unbalanced labels (most object candidates are negative), which is defined as $-\sum_{i=1}^{C} w_i p_i \log(\hat{p}_i)$, where p, $\hat{p}$ denote the ground truth and predicted object/node class, respectively, and w_i is the predefined weight for class i.

Taken together, the total loss for end-to-end training is

$$\mathcal{L} = \mathcal{L}_{hm} + \alpha\mathcal{L}_{obj} + \beta\mathcal{L}_{node}, \tag{9}$$

where $\mathcal{L}_{obj}$ and $\mathcal{L}_{node}$ are weighted cross entropy losses for object classification and node classification respectively.

4 Experiments

4.1 Datasets

Our in-house stent dataset consists of 4,480 videos (128,029 frames of 512 $\times$ 512 8-bit frames) acquired during PCI procedures. The data acquisition was approved by Institutional Review Boards. For in-house videos, the landmarks are radiopaque balloon marker pairs located at two ends of the stent (Fig. 1). There are 114,352 marker pairs in the dataset, which were manually annotated by trained experts. The dataset was split into training, validation, and testing set with a 8:1:1 ratio, which resulted in 3584 videos (103892 frames), 448 videos (12990 frames), and 448 videos (11147 frames) respectively.

In addition, to verify generalization of our method, we included a public dataset in our experiment. The transcatheter aortic valve implantation (TAVI) dataset is a public intraoperative aortography X-ray imaging dataset including 35 videos of 1000×1000 pixels 8-bit frames. The original dataset consisted of 11 keypoint annotations including 4 anatomical landmarks, 4 catheter landmarks and 3 additional background landmarks. TAVI is different from PCI procedure but it contains landmark pairs: Catheter Proximal (CP) and Catheter Tip (CT) whose constellations are similar to the stents. We excluded irrelevant landmarks resulting the final TAVI dataset contains 2,652 frames from 26 videos. The TAVI dataset was randomly divided into a training set with 2,027 images from 18 videos and a test set with 625 images from 8 videos. We ran K-fold (k=5) cross-validation for all models on both private dataset and the TAVI dataset, and the detailed results are reported in the Supplementary Materials.

4.2 Comparative Models

To demonstrate the efficacy of our algorithm, we compared it with several SOTA models on both datasets. First, we selected two coordinate regression models, ResNet V2 [11] and MobileNet V2 [17]. Such regression models detect landmarks in each frame by predicting the landmark center coordinates, which have shown

superior performance regressing TAVI catheter landmarks in [9]. Moreover, we include a center based multi-object tracking (MOT) model, CenterTrack [27], with two most powerful backbones: DLA-34 [26] and MobileNet V2. CenterTrack detects objects as heatmap regression of centers and simultaneously tracks objects over time by predicting the translations of center points between adjacent frames. CenterTrack has demonstrated extraordinary performance on various MOT benchmark datasets.

4.3 Evaluation Metrics

As our final goal is to enhance the stents by aligning landmark points across frames, detection success rate and localization accuracy are the most important factors to ensure high-quality enhancement. To compare the landmark prediction performance, we use the following detection and localization metrics for evaluation. For detection performance, we used *Precision*, *Recall*, F_1 and *Accuracy*. Landmark locations extracted from heatmaps were paired with the closest ground truth(GT) greedily. A stent prediction was matched if distances of its both landmarks to paired GT were smaller than 5pxs(in-house) or 15pxs(TAVI).

$$Precision = \frac{TP}{TP + FP}; \quad Recall = \frac{TP}{TP + FN}$$
$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FN + FP}; \quad Accuracy = \frac{TP + TN}{TP + FN + TN + FP}$$

On the successfully detected landmarks, we also evaluated landmark localization accuracy using pixel-wise MAE and RMSE:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - \hat{p}_i|; \quad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(p_i - \hat{p}_i)^2},$$

where p_i, $\hat{p}_i$ denote predicted and ground truth landmark coordinates.

4.4 Implementation Details

All deep learning models were implemented with PyTorch and run on NVIDIA V100. For the proposed method, the marker detection module was pre-trained on 128 $\times$ 128 image patches, and then the whole model was trained on 10-frame video clips. We used Adam optimizer with a learning rate of 1e-5. For the coordinates regression models, we follow the multi-task learning schemes provided by Danilov et al. [9], using Binary Cross Entropy for the classification branch and Log-Cosh loss for the regression branch, optimized with Adam with a learning rate of 1e-5. Similarly for CenterTrack, we trained the models based on the configuration described in the original publication. The major modification we made was to remove the branch of object bounding box size regression since we do not need the landmark size estimation for our task. We used the focal loss

Table 1. Evaluations on **In-house Dataset**. 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↓
CR	MobileNetV2	0.620	0.557	0.587	0.415	1.172	1.283
	ResNetV2	0.618	0.604	0.611	0.440	1.064	1.125
CT	MobileNetV2	0.485	0.932	0.638	0.469	0.455	0.837
	DLA34	0.591	**0.936**	0.725	0.568	**0.398**	**0.748**
Ours		**0.907**	0.908	**0.908**	**0.831**	0.597	0.963

Table 2. Evaluations on **TAVI Dataset**.

Model		Detection				Localization	
Type	Backbone	Precision↑	Recall↑	F1↑	Accuracy↑	MAE↓	RMSE↓
CR	MobileNetV2	0.839	0.735	0.784	0.644	12.904	14.129
	ResNetV2	0.857	0.846	0.851	0.741	11.571	12.490
CT	MobileNetV2	0.785	**0.961**	0.864	0.761	**5.100**	**6.159**
	DLA34	0.868	0.930	0.898	0.815	5.418	6.357
Ours		**0.918**	0.957	**0.938**	**0.882**	5.975	6.831

for heatmap regression and L1 loss for offset regression, optimized with Adam with learning rate 1.25e-4. Please see the supplementary material for more details on hyperparameters.

5 Results and Discussion

5.1 Main Results

Table 1 and Table 2 list the results of proposed model and baseline models on the in-house dataset and the public TAVI dataset [4]. The results are consistent on both datasets. In terms of detection, our framework significantly outperforms the prior state of the art on both datasets. Firstly, tracking models generally excels pure detection models as the additional temporal information is helpful to enhance landmark detection. Another major limitation of the coordinate regression models is that the number of detections is always fixed. Therefore, some targeted landmarks can be easily overwhelmed by strong background noises in coordinate regressor, resulting in a common trend of lower *recall*. On the contrary, heatmap regression models have the flexibility to predict more possible landmarks as long as the desired features are identified from the image. This would help achieve higher recall but also resulting in a large number of false positive landmark detections, indicated as the worse CenterTrack *precision* values compared to coordinate regressors. To solve this issue, in our framework design,

[4] Example of stent tracking and enhancement comparisons are included in the Supplementary Materials.

we introduced both additional spatial information and temporal information to refine the noisy preliminary detections.

As tracking models, the proposed model shows a remarkable detection margin over CenterTrack. The results demonstrate the effectiveness of our two major innovation in the framework: stent proposal and GCN-based tracking. Instead of tracking multiple landmarks as individual points as in CenterTrack, our model enhanced the isolated detections by introducing stent proposal stage. The feature of possible stents patches between candidate landmark pairs enables the model to enhance landmarks learning and context relationship between landmark pairs. Moreover, as a pure local tracking model, landmarks association of CenterTrack is limited to adjacent frames, where the landmark association is simply learnt as spatial displacements. The information propagation of our multi-layer GCN model enables stent feature nodes in graph, with the combination of both spatial and appearance features, to interact with other nodes in longer time range. We will further prove the effectiveness of our designs in the following ablation studies.

To compare two datasets, our in-house dataset includes more videos with more complicated background compared to the TAVI dataset, which makes the stent tracking a more challenge task for all models, with more likely false positive detections. From the results, we could observe the notable declines in in-house dataset detection metrics, especially in *precision*, for each model compared to TAVI's. However, the *precision* of our framework only dropped from 0.918 to 0.907, which indicates that our framework is more robust to suppress false positives and maintain high accuracy detection in more complex PCI scenes. This robustness advantage is a good reflection of our hierarchical landmark and stent feature extraction efficacy at the stent proposal stage. Even though more complicated background would cause confusion to individual landmark detection, our model can still successfully target the desired landmark pairs by identifying the prominent stent feature in between.

As for localization evaluation, the MAE and MSE values we got from TAVI is about 10 times larger than our own dataset. This is because, firstly, the original frame size is significantly larger than our data; secondly, the landmarks in TAVI, CP and CD, are also about 10 times larger than our landmarks. For example, the CD landmarks are more of a dim blob rather than a single opaque point as in our dataset. To compare the results among models, heatmap regression models perform generally better than coordinate regression models, as the numerical regression still remain an quite arbitrary task for CNN model and localization accuracy would be diminished during the single downsampling path in regressors. CenterTrack achieves the lowest MAE and MSE errors on both datasets. The high accurate localization of CenterTrack is realized by both sophisticated model architecture and the additional two-channel output branch specifically designed for localization offsets regression. The cost for the localization accuracy improvement is computational complexity and time (CenterTrack-DLA34: 10.1 FPS). As for our model, the localization solely depends on the heatmap regression from the faster U-Net (21.7 FPS). We have not tried to add any localization refinement design for the framework efficiency. On our dataset, the mean MAE

Table 3. Ablation study on **In-house Dataset**.

Model	Detection				Localization	
	Precision↑	Recall↑	F1↑	Accuracy↑	MAE↓	RMSE↓
Detection only	0.551	**0.918**	0.688	0.525	**0.596**	0.998
Detection refinement	**0.927**	0.532	0.677	0.511	0.598	**0.946**
Separate learning	0.893	0.872	0.883	0.790	**0.596**	0.955
Ours	0.907	0.908	**0.908**	**0.831**	0.597	0.963

value at 0.597 is already sufficient for our final stent enhancement task. Further localization refinement with the sacrifice of time would not cause any noticeable improvements. However, we want to note that our framework is very flexible that the current U-Net backbone can be replaced by any other delicate heatmap regression models, if a more accurate localization is necessary for an application.

5.2 Ablation Studies

The proposed end-to-end learning framework consists of three major modules: heatmap regression based landmark detection (U-Net), landmark-conditioned stent proposal and feature extraction (ResNet), and stent tracking GCN. To demonstrate the benefits of our framework design, we ablated the main components of stent proposal ResNet and stent tracking GCN, as well as the end-to-end learning regime.

Detection only directly utilizes the U-Net to detect individual landmarks at each frame. This model only needs separate frames as inputs and individual landmark locations as supervision.

Detection refinement improves landmark detections in the heatmap prediction from the U-Net backbone and filters false positives by incorporating additional spatial information of the stent between candidate landmarks using a ResNet patch classifier. This two-step model also only requires single frame inputs and no temporal association is used.

Separate Learning includes all three proposed major modules (landmark detection, stent proposal and stent tracking). Instead of simply filtering out false stent patches by the CNN model, this model first uses the convolution layers from well-trained patch classification ResNet to extract feature vectors from candidate stent patches. Then, the features are reconstructed into stent graph and fed into the GCN model for stent tracking over frames. The final model outputs would be the tracked stents at each frame of the input video. However, this three-step approach is achieved by training each model independently: U-Net for landmark detection, ResNet for stents patch classification and GCN for stent tracking.

The ablation study results are shown in Table 3. Our full end-to-end model performs significantly better than the baselines in the detection task. The standalone U-Net yields a very high false positive rate (44.9%) as it is difficult for this model to learn meaningful features to differentiate small landmarks from dark spot noises in the background.

In the detection refinement results, the stent patch classifier significantly reduced the false positives from U-Net predictions, as *precision* surged to 92.7%. However, simply applying the patch classifier to the U-Net outputs would also filter out true stent patches with weaker patch features, resulting in a large drop in *recall*. The above results indicate a trade-off between *precision* and *recall* while applying spatial information based models.

The results of separate learning demonstrate that incorporating GCN temporal stent tracking improved recall and maintained the high precision from detection refinement, resulting in a boost in overall detection accuracy. False negatives are effectively suppressed by the information propagation mechanism in GCN, which helps to enhance the feature of weak but true stent nodes with nearby strong stent nodes in both space and time.

Compared with the separate three-stage learning model, our proposed end-to-end model achieved further improvements in all detection evaluation metrics and reached a better balance between *precision* and *recall*. Although different components of our framework have their specific tasks along the detection and tracking process, the end-to-end learning brings extra benefits, especially by optimizing the data flow between modules. For example, the back-propagated gradients from GCN can also guide the convolutional layers at stent proposal to extract better patch features that would be fed into GCN.

In regard to localization accuracy, all baseline models and the final model show similar performance, as we used the same U-Net backbone for all experiments. The MAE and RMSE values fluctuate within 0.002 and 0.052, which we believe are only from experimental uncertainty and would not have a sensible influence on the final stent enhancement task. For many multi-task learning models on limited data, there is conventionally a trade-off between excellency on specific metrics and good overall performance. The results suggest that the complicated multi-task learning of our end-to-end model would both maintain high localization accuracy and improve detection.

6 Conclusion

In this work, we proposed a novel end-to-end CNN-GCN framework for stent landmarks detection and tracking. The model includes three major modules: (1) U-Net based heatmap regression for landmark candidate detection, (2) a ResNet for landmark-conditioned stent proposal and feature extraction, and (3) residual-GCN based stent tracking. We compared the proposed model with SOTA coordinate regression models and multi-object tracking models. Our experiments demonstrated that the proposed model remarkably outperformed previous SOTA models in stent detection. We further discussed the flexibility of the proposed framework to accommodate new heatmap regression backbones to overcome the current localization limitations. The ablation experiments showed the benefits of our novel designs in stent proposal ResNet, stent tracking GCN, and end-to-end learning scheme.

References

1. Bier, B., Unberath, M., Zaech, J.N., Fotouhi, J., Armand, M., Osgood, G., Navab, N., Maier, A.: X-ray-transform Invariant Anatomical Landmark Detection for Pelvic Trauma Surgery. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2018. pp. 55–63. Lecture Notes in Computer Science, Springer International Publishing, Cham (2018)
2. Bismuth, V., Vaillant, R.: Elastic registration for stent enhancement in x-ray image sequences. In: 2008 15th IEEE International Conference on Image Processing. pp. 2400–2403 (2008). https://doi.org/10.1109/ICIP.2008.4712276
3. Bismuth, V., Vaillant, R., Funck, F., Guillard, N., Najman, L.: A comprehensive study of stent visualization enhancement in x-ray images by image processing means. Medical Image Analysis **15**(4), 565–576 (2011)
4. Braso, G., Leal-Taixe, L.: Learning a Neural Solver for Multiple Object Tracking. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6246–6256. IEEE, Seattle, WA, USA (Jun 2020)
5. Butt, A.A., Collins, R.T.: Multi-target Tracking by Lagrangian Relaxation to Min-cost Network Flow. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1846–1853. IEEE, Portland, OR (Jun 2013)
6. Chabi, N., Beuing, O., Preim, B., Saalfeld, S.: Automatic stent and catheter marker detection in x-ray fluoroscopy using adaptive thresholding and classification. Current Directions in Biomedical Engineering **6** (2020)
7. Close, R.A., Abbey, C.K., Whiting, J.S.: Improved Localization of Coronary Stents Using Layer Decomposition. Computer Aided Surgery **7**(2), 84–89 (Jan 2002)
8. Close, R.A., Abbey, C.K., Whiting, J.S.: Improved localization of coronary stents using layer decomposition. Computer Aided Surgery **7**(2), 84–89 (2002). https://doi.org/10.3109/10929080209146019
9. Danilov, V.V., Klyshnikov, K.Y., Gerget, O.M., Skirnevsky, I.P., Kutikhin, A.G., Shilov, A.A., Ganyukov, V.I., Ovcharenko, E.A.: Aortography keypoint tracking for transcatheter aortic valve implantation based on multi-task learning. Frontiers in Cardiovascular Medicine **8** (2021)
10. Ghesu, F.C., Georgescu, B., Mansi, T., Neumann, D., Hornegger, J., Comaniciu, D.: An Artificial Agent for Anatomical Landmark Detection in Medical Images. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) Medical Image Computing and Computer-Assisted Intervention - MICCAI 2016. pp. 229–237. Lecture Notes in Computer Science, Springer International Publishing, Cham (2016)
11. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Computer Vision – ECCV 2016. pp. 630–645 (2016)
12. Kipf, T.N., Welling, M.: Semi-Supervised Classification with Graph Convolutional Networks. In: ICLR 2017. pp. 24–26 (2017)
13. Lindeberg, T.: Feature detection with automatic scale selection. International Journal of Computer Vision **30**(2), 79–116 (1998). https://doi.org/10.1023/A:1008045108935
14. Lu, X., Chen, T., Comaniciu, D.: Robust discriminative wire structure modeling with application to stent enhancement in fluoroscopy. In: CVPR 2011. pp. 1121–1127 (2011)
15. Mishell, J.M., Vakharia, K.T., Ports, T.A., Yeghiazarians, Y., Michaels, A.D.: Determination of adequate coronary stent expansion using StentBoost, a novel fluo-

roscopic image processing technique. Catheterization and Cardiovascular Interventions **69**(1), 84–93 (2007)
16. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR 2011. pp. 1201–1208 (Jun 2011)
17. Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 4510–4520 (2018)
18. Schoonenberg, G., Lelong, P., Florent, R., Wink, O., ter Haar Romeny, B.: The effect of automated marker detection on in vivo volumetric stent reconstruction. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008. pp. 87–94. Lecture Notes in Computer Science, Springer (2008)
19. Shao, C., Wang, J., Tian, J., Tang, Y.d.: Coronary artery disease: From mechanism to clinical practice. In: Wang, M. (ed.) Coronary Artery Disease: Therapeutics and Drug Discovery, pp. 1–36. Advances in Experimental Medicine and Biology, Springer (2020)
20. Toshev, A., Szegedy, C.: DeepPose: Human Pose Estimation via Deep Neural Networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1653–1660. Columbus, OH, USA (Jun 2014)
21. Vernikouskaya, I., Bertsche, D., Dahme, T., Rasche, V.: Cryo-balloon catheter localization in X-Ray fluoroscopy using U-net. International Journal of Computer Assisted Radiology and Surgery **16**(8), 1255–1262 (Aug 2021)
22. Wang, B., Wang, G., Chan, K.L., Wang, L.: Tracklet Association with Online Target-Specific Metric Learning. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1234–1241. IEEE, Columbus, OH, USA (Jun 2014)
23. Wang, Y., Chen, T., Wang, P., Rohkohl, C., Comaniciu, D.: Automatic localization of balloon markers and guidewire in rotational fluoroscopy with application to 3d stent reconstruction. In: Computer Vision – ECCV 2012. pp. 428–441 (2012)
24. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic Graph CNN for Learning on Point Clouds. ACM Transactions on Graphics **38**(5), 146:1–146:12 (Oct 2019)
25. Wu, Y., Ji, Q.: Facial Landmark Detection: A Literature Survey. International Journal of Computer Vision **127**(2), 115–142 (Feb 2019)
26. Yu, F., Wang, D., Shelhamer, E., Darrell, T.: Deep layer aggregation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2403–2412 (2018)
27. Zhou, X., Koltun, V., Krähenbühl, P.: Tracking objects as points. ECCV (2020)