2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Multiple Surgical Instruments Tracking-By-Prediction With Graph Hierarchy

Rui Guo, Xi Liu, Ziheng Wang and Anthony Jarc

Abstract— **Current research strive has tremendously changed the horizon of computer vision tasks in multiple agents tracking. Nevertheless, in the research of robotic assisted surgery, reliable surgical instrument tracking imposes challenge due to the high complexity in state modeling for the hierarchical structure of the instrument versus de-coupling the spatial-temporal correlations naturally embedded in the task. In this paper, we present a new tracking paradigm integrating the trajectory prediction to reduce the data association error that is propagated from the false detection. As a key component in the system, a proposed predictor disentangles the hierarchical modeling and agent kinematic learning by introducing inductive attention mechanism in spatial-temporal graph network. Experiments on real anatomical datasets show that our tracking-by-prediction scheme improves overall localization accuracy over the frames by up to 81%, in comparison to the generic pipelines of tracking, even with transductive graph representation learning, with a large margin of gain in terms of precise localization.**

I. INTRODUCTION

Robust visual tracking for multiple agents empowers many autonomy applications. Despite the importance, attempts in building reliable trackers are suffering due to the high complexity in de-coupling spatial-temporal correlations using the neural network. For some other applications, such as robotic assisted surgery, tracking the pose of instrument desires high fidelity model, level up the difficulty and retain the challenge to more researches. Compared to general Multiple Object Tracking (MOT) scenario, surgical instruments used during surgery usually interact with each other. The dynamic modeling of instrument interaction for surgery-specific actions such as dissect, retract and suture directly affects the performance of tracking.

Origins from modeling the social behavior of crowd pedestrian, researchers introduced the social-LSTM as a kernel that describes the dynamics of pedestrians' interaction [1], where the latent motions represented with the hidden states of LSTMs are shared by the mechanism of "social-pooling". Advanced in the social pooling design, individual pedestrian is not treated as isolated entity, but grouped together at the pooling based on defined "neighborhood" relations. Soft attention is also utilized to establish the relative influence among the pedestrians. The attention model calculates a weight matrix that assigned unequal importance to the neighboring pedestrians. It increases the flexibility of the model to understand the crowd behavior based on the spatial interactions. However, social-LSTM models each pedestrian equally using a LSTM. It is not applicable to the complex entity with obvious hierarchical structure. If we would like to model the pose of the pedestrian in a crowd, the plain social-LSTM is less representative to distinguish the detailed structure in the scenario.

Rapidly developed in the theory of graph neural network inspires the advanced modeling using graph representation for the un-structured data [7], [23], [**?**]. To address the limitations mentioned above, we build a novel Spatial-Temporal Graph hierarchy, where the spatial and temporal interactions among surgical instruments are encoded, respectively. The spatial interaction at one time-step are captured by the graph attention scheme, which models over all the surgical instruments shown up in the clinic operation. After assigning the different importance on keypoints, an extra LSTM is used to capture the temporal correlations of interactions. By aggregating all the spatial-and-temporal interaction among all the keypoints and intrument entities, the future trajectories are generated by a sequence-to-sequence (seq2seq) translation. To model the diverse motion patterns, we rely on the novel hierarchy, in which the intra-instrument keypoints are modeled in the lower level and the inter-instrument interaction is also modeled by defining root nodes and connecting them in higher level graph topology.

Besides modeling the multiple instruments interaction, object association is another aspect that deeply affecting the tracking performance. Conventional model utilizes trajectories only to understand the dynamics of tracked objects in the past, but apparently, one directional temporal encoding ignoring the future movement results in inconsistent tracking association due to the imperfection of localization error accumulated over time. Tracking-by-prediction is a sophisticated scheme that bridges the gap. Instead of temporal correlation from historic traces, the tracked objects in the next frame are associated with targets via considering their predicted short- and long-term motions. The bi-directional continuity enforces the smoothness and correction of the tracking association in ambiguous scenarios.

The key contribution of the paper focuses on the proposal of surgical instrument tracking-by-prediction with inductive hierarchical spatial-temporal graph network. The novelty of the approach is summarized in twofold:

- We present a novel graph hierarchy to represent the spatial-temporal complexity embedded in the multiple surgical instrument tracking problem. The model explicitly extends the graph predict trajectories of multiple surgical tools in the surgery. The model explicitly

Rui Guo, Xi Liu and Ziheng Wang are machine learning scientists with Intuitive Surgical Inc. Anthony Jarc is leading the operative analytics team in Intuitive Surgical Inc. 5655 Spalding Drive, Peachtree Corners, GA 30092, USA rui.guo, xi.liu, ziheng.wang, anthony.jarc @intusurg.com

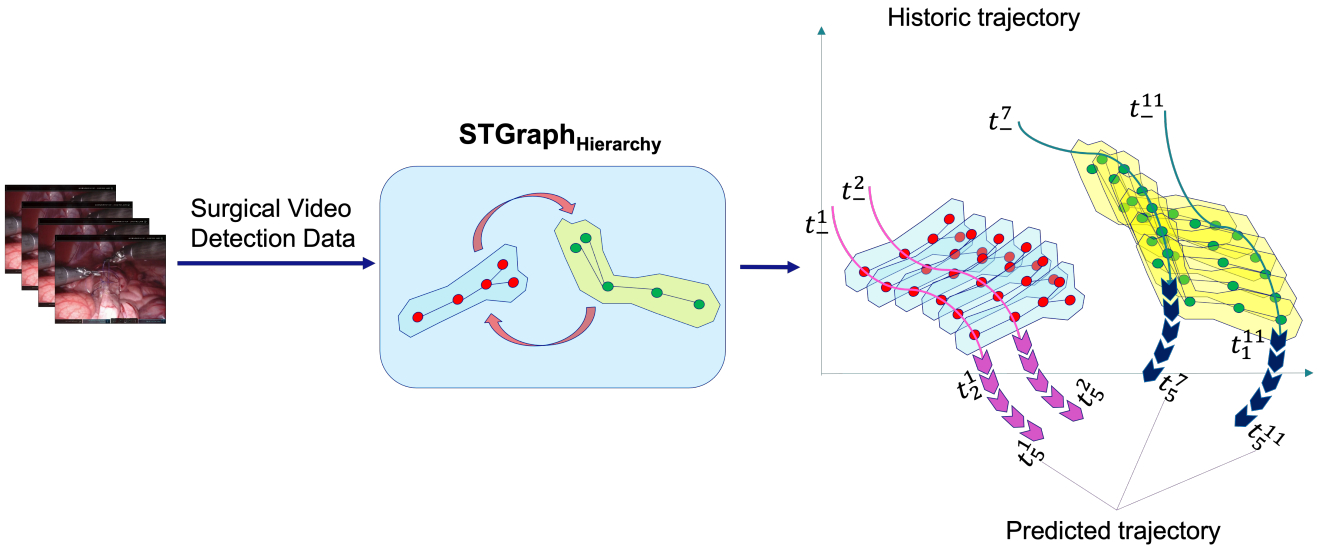

Fig. 1. System pipeline of the proposed tracking-by-prediction. The $STGraph_{hierarchy}$ encodes the positions of detected surgical instrument in a graph hierarchy to predict their motions. Integrated the prediction in the tracking-by-prediction framework further process to generate the optimized data association that completing the tracking task.

encodes the spatial-temporal correlation with emphasis on the instruments' interaction.

- We re-frame the tracking by aggregating the prediction that to mitigate the data association inaccuracy.

II. Related work

Our proposed keypoint tracking system is the pioneer work in robot-assisted surgical instrument analysis, the most relevant prior arts to the problem are multiple objects tracking with crowd interaction modeling. We conduct the literature review from the following areas that covering the topic.

A. Crowd Objects Motion Modeling And Prediction

Similar to the multiple surgical instruments motion prediction, crowd objects, such as pedestrian interaction and trajectory modeling has gained intensive research attentions. Early researches, such as Social Forces [10], [15], [20], modeling human-human interactions in dynamic scenes by various types of forces. However, as some key parameters are highly based on prior knowledge, such as force definition, they can only be effective in handling attractive and repulsive forces. For the sophisticated scene, in which crowd behavior varies person by person, they failed to perform well in modeling for the crowd. Recent years, Recurrent Neural Network (RNN) has shown great advantages for modeling sequential data [2], [4], [19]. As one variation, long- short term memory (LSTM) received great attention from researchers by offering strong prediction capability. Alahi et al. [1] proposed social-LSTM which applies social pooling after each time step in vanilla LSTM to integrate social features for the group. Gupta et al. [8] improved social pooling to capture global context. These pooling methods use distance measurement among the pedestrians as a criterion to evaluate the importance of the relationship. Further, Xu et al. [28] used LSTMs as motion encoder to handle only temporal information. For spatial context, another encoder was adopted. Xu also introduced attention-like mechanism to propagate social features in his work [28], but it also meets the problem that the "attention" is highly restricted by the softmax activation function. Sadeghian et al. [22] used pure spatial distance in Euclidean metric between agents as a reference to permute these agents rather than attention mechanism, followed by, Ivanovic et al. [12] used Euclidean distance to build a traffic agent graph to guide attention mechanism.

The aggregation of crowd objects interaction modeling with sequential prediction using graph neural network is firstly proposed by [11]. In the framework, pedestrian motion is modeled by an LSTM, and the temporal correlations of interactions is modeled by an extra LSTM. Graph attention network (GAT) is introduced to aggregate hidden states of LSTMs to model the spatial interactions. We follows this work a lot in this paper for spatial-temporal graph encoding.

B. Multiple Objects Tracking-by-Prediction

Crowd modeling using recurrent network is not an isolated research area from detection-prediction-tracking pipeline [14]. With recent advances in sequence modeling with deep learning approaches, a data driven method for crowd motion modeling has been proposed in [24], which was further expanded to capture the entire history from the neighborhood in [6]. By extending the temporal horizon, authors in [5] investigated into modeling long term dependencies such as motion patterns between sequences for the task of trajectory prediction, rather than using dependencies within the moment. However, these methods were crafted for long term prediction of trajectories, which is similar to re-identification, is as opposed to tracking.

Numerous attempts have been made to incorporate the motion modeling approaches for data association to tracking.

In [21] the authors utilize LSTMs as motion and interaction models so as the appearance model to tackle with the occlusions, noisy detection and appearance changes. They utilized two separate LSTMs as their motion and interaction models. Their approach needs heavy processing such as occupancy map generation to obtain interaction pattern. The authors of [16] proposed a LSTM based end-to-end approach for multi-target tracking. However, the built model did not consider interactions among objects and the approach results in erroneous and non realistic trajectory generation. The most closely related work to ours is [6], which proposed a "tracking-by-prediction" paradigm. It compares short-term predictions with each detection for data association. In the work of [29], authors proposed "re-tracking" module, in which parallelized tracking and prediction framework that can be jointly optimized to reduce the accumulated localization error to improve the tracking robustness.

III. Approach

The goal of the paper is building up a novel tracking framework for multiple instruments in surgical scenario. We present the approach in this section with a dedicated graph hierarchy, modeling the multiple instruments in an aggregation of spatial and temporal encoding, to be able to predict the trajectories of the tracked keypoints of the instrument. The prediction is then integrated into a tracking-by-prediction pipeline to provide the final high-fidelity tracking output.

A. Surgical Instrument Unified Keypoints Modeling

Instruments used during robotic-assisted surgery on the da Vinci® surgical system (Intuitive Surgical Inc.) have articulated joints to enable fully wristed dexterity. The accurate articulation of the instrument is usually approached by modeling the joints with keypoint representation. Without losing the generality, the keypoint model of the instrument is made up of 5 landmark joints, named tips (tip-1 and tip-2 respectively), clevis, shaft and shaft-end, as illustrated in Fig. 2. The ordered joint pairs are formulating the skeleton of instrument. For monopolar tool such as the permanent cautery hook, we assume the tip-1 and tip-2 are overlayed at the same location.

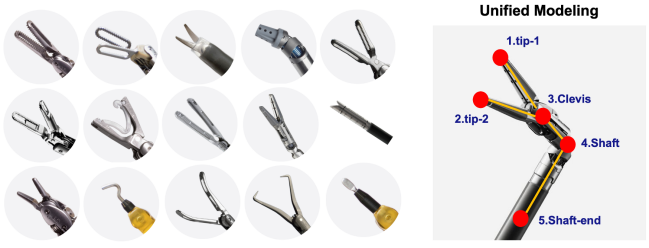

Fig. 2. Various surgical instruments and the unified 5 keypoints modeling

B. Problem Formulation For Multiple Instrument Tracking

We assume the surgery involves N instruments in the scene. Each instrument contains 5 keypoints, denoting them in an order as $p_1, p_2, ..., p_{5N}$. For the i_{th} keypoint at time frame t, the location coordinates are denoted as $S_i^t = (x_i^t, y_i^t)$. The tracking problem targeted in the paper is then formulated as, given a set of trajectories $[TR_1, TR_2, ..., TR_{5N}]$ for N instruments, where $TR_i = S_i^t|_{t=-1}^{-n}$, time interval from $-n$ to -1 represents the past n time steps, how to correctly associate the current observation $O_i^{t=0}$ $(i \in [1, 5N])$ with all trajectories.

C. Hierarchical Graph Modeling

The motion pattern of each keypoint in instrument varies due to the articulation, kinematic freedom and surgical task difference. We use one LSTM for each keypoint to capture the latent motion state and preserve the diverse of motion pattern. For the input of this recurrent layer, the relative position is coverted into embedded vector and fed into the LSTM cell. We term it as $LSTM_{spa}$. For one single time step, the $LSTM_{spa}$ encoding is processed as

$$v_i^t = \Phi(\Delta x_i^t, \Delta y_i^t; W_\varphi) \tag{1}$$

$$r_i^t = LSTM_{spa}(r_i^{t-i}, v_i^t; W_\theta) \tag{2}$$

where, $\Phi(\cdot)$ is an embedding function with weighting parameters W_φ. Δx_i^t, Δy_i^t are relative positions of the keypoint, $\Delta x_i^t = x_i^t - x_i^{t-1}$, $\Delta y_i^t = y_i^t - y_i^{t-1}$ respectively. Given the relative position vector $(\Delta x_i^t, \Delta y_i^t)$, $LSTM_{spa}$ parameterized with W_θ generates hidden state representation r_i^t for spatial encoding.

Using LSTM encoding for each keypoint of the instrument has an obvious drawback: the inter-instrument interaction is totally ignored and thus fails to model the instrument dynamics precisely. As a specific hierarchy, we create a virtual anchor point, geometric center of the 5 keypoints, as a root node to represent the instrument, and re-formulate the surgical scene using a graph. In the graph, keypoints from the same instrument are connected to the root showing their neighboring relationship at the low level. Inter-instrument relation is implemented by connecting their root nodes, as a higher level of the hierarchy. In the new topology, joint-joint interaction and instrument-instrument influence are well modeled, indicating subtle behavioral dependency of the scene. Noted that, we add the virtual root in the keypoint set of the instrument to extend it to 6 points.

To emphasize on the hierarchy, a graph attention layer is introduced to the graph representation learning. Attention mechanism allows a learnable weight vector assigned to the neighborhood of nodes that indicates their various contributions to the task objective in the process. In the mathematics, the graph attention layer is taking $LSTM_{spa}$ output and calculating the attention coefficient α_{ij}^t for paired node (i, j) as

$$\alpha_{ij}^t = \frac{\exp(LeakyReLU(a^T[Wr_i^t || Wr_j^t]))}{\sum_{k \in \mathcal{N}_i} \exp(LeakyReLU(a^T[Wr_i^t || Wr_k^t]))} \tag{3}$$

where $\cdot^T$ represents matrix transportation and $||$ is the concatenation operation. W is a shared matrix, to linearly

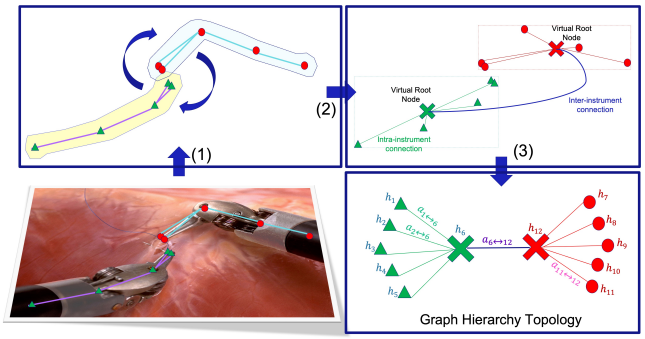

Fig. 3. The graph hierarchy abstraction. Clock-wise from bottom left: (1) The detected keypoints from multiple instruments are grouped and converted into pro-interaction graph representation. (2) The virtual root nodes are created and the connections reflect their neighborhood relationship. (3) Graph hierarchy topology is established.

transform the hidden state vector. $\mathcal{N}_i$ is the neighborhood of node i that is utilized to define the hierarchy. $LeakyReLU$ represents the activation function parameterized with slope a. Once obtained, the graph attention layer outputs the feature vector $\hat{r}_i^t$ for node I by

$$\hat{r}_i^t = \sigma(\sum_{j \in \mathcal{N}_i} \alpha_{ij}^t W r_j^t) \tag{4}$$

where $\sigma(\cdot)$ is a non-linear activation function that aggregating over transformed importance within the node i neighborhood. The hierarchical dynamic modeling is then achieved.

After getting the feature vector from graph attention layer, another LSTM is applied along the temporal dimension that modeling the correlations of the sequence. We term it as $LSTM_{temp}$, and the output is expressed as

$$l_i^t = LSTM_{temp}(l_i^{t-1}, \hat{r}_i^{t-1}; W_{temp}) \tag{5}$$

The intuitive integration of spatial and temporal encodings is concatenating the outputs from $LSTM_{spa}$ and $LSTM_{temp}$ as one comprehensive feature vector h_i^t,

$$h_i^t = [r_i^t || l_i^t] \tag{6}$$

D. Trajectories Prediction

To keep the model consistent and differentiable, we pass the historical trajectory state embedding h_i^t through a LSTM decoder to generate the trajectory prediction. We term the LSTM decoder as $LSTM_{pred}$, and the predicted relative position at one time step is calculated by

$$g_i^{t+1} = LSTM_{pred}(h_i^t, v_i^t; W_{pred}) \tag{7}$$

$$(\Delta x_i^{t+1}, \Delta y_i^{t+1}) = \delta_p(g_i^{t+1}) \tag{8}$$

where W_{pred} denotes weight of $LSTM_{pred}$ and v_i^t could be computed by following embedding equation 1. $\delta(\cdot)$ is as the same as the linear layer used in [1].

The LSTM parameters are learned by minimizing the negative log-likelihood loss ground-truth positions with the predicted Gaussian distribution [24], [1], [9].

One-step prediction using the proposed framework could be easily extended to a sequence-to-sequence prediction by replacing the input from single step vector into a sequential vector set (to keep the time logic, only the historical position information is utilized in the input vector set). And the output could also be multiple time step position forecast. The trajectory prediction pipeline is illustrated in Fig. 4.

E. Tracking By Prediction

Assuming we have an active trajectory tr with m time step historical position information $tr = [p_{t-m}, p_{t-m+1,...,p_t}]$, applying the trajectory prediction proposed above, it could predict n time step positions fo the same object. The new extended trajectory is denoted as $\overline{tr} = [p_{t-m}, p_{t-m+1}, ..., p_t, \hat{p}_{t+1}, \hat{p}_{t+2}, ..., \hat{p}_{t+n}]$. Considering a hypothesis that we associate current observation o_{obs} with trajectory tr and applying the same prediction process with m-step history to generate n-step positions, we are able to obtain the extended trajectory $\widehat{tr^+} = [p_{t-m+1}, p_{t-m+2}, ..., p_t, o_{obs}, \hat{p}_{obs+1}, \hat{p}_{obs+2}, ..., \hat{p}_{obs+n}]$. The time overlapped portion of trajectories $\overline{tr}$ and $\widehat{tr^+}$ are $\overline{tr}_{sub} = [\hat{p}_{t+2}, ..., \hat{p}_{t+n}]$ and $\widehat{tr^+}_{sub} = [\hat{p}_{obs+1}, ..., \hat{p}_{obs+n-1}]$. If the association hypothesis is established, the difference between $\overline{tr}_{sub}$ and $\widehat{tr^+}_{sub}$ should be minimal. This is the principle of the tracking-by-prediction strategy that we proposed to solve the data association challenge.

We assume each position $p_i^{\overline{tr}}$ in trajectory $\overline{tr}$ is following a Gaussian distribution: $p_i^{\overline{tr}} \sim \mathcal{N}(\mu_i^{\overline{tr}}, \Sigma_i^{\overline{tr}})$. For the same, each position $p_j^{\widehat{tr}^+}$ in trajectory $\widehat{tr}^+$ follows Gaussian distribution: $p_j^{\widehat{tr}^+} \sim \mathcal{N}(\mu_j^{\widehat{tr}^+}, \Sigma_j^{\widehat{tr}^+})$. The distance metric over two distributions in Mahalanosis norm is

$$d(p_i^{\overline{tr}}, p_i^{\widehat{tr}^+}) = \sqrt{(\mu_i^{\overline{tr}} - \mu_i^{\widehat{tr}^+})^T (\Sigma_i^{\overline{tr}} + \Sigma_i^{\widehat{tr}^+})^{-1} (\mu_i^{\overline{tr}} - \mu_i^{\widehat{tr}^+})} \tag{9}$$

Then, the data association under tracking-by-prediction is formulated as solving Hungarian algorithm [29], [27] to select the winning hypothesis by using the distance metric

$$d(\overline{tr}, \widehat{tr}^+) = \frac{1}{|T|} \sum_{k \in T} d(p_k^{\overline{tr}}, p_k^{\widehat{tr}^+}) \tag{10}$$

where T is the set of overlapped tie steps between predicted trajectories $\overline{tr}$ and $\widehat{tr}^+$.

Once the new observation associated with the active trajectory, we post-process the trajectory by applying the Holts-Winters smoothing [29], [13]. From the practice, the smoothing process significantly improve the tracking performance.

IV. Experiments

Our surgical instrument tracking-by-prediction has two core goals: (1) Improve the overall keypoint localization performance in the clinic setting, and (2) Increase the robustness against the tracking error. To this end, our prime

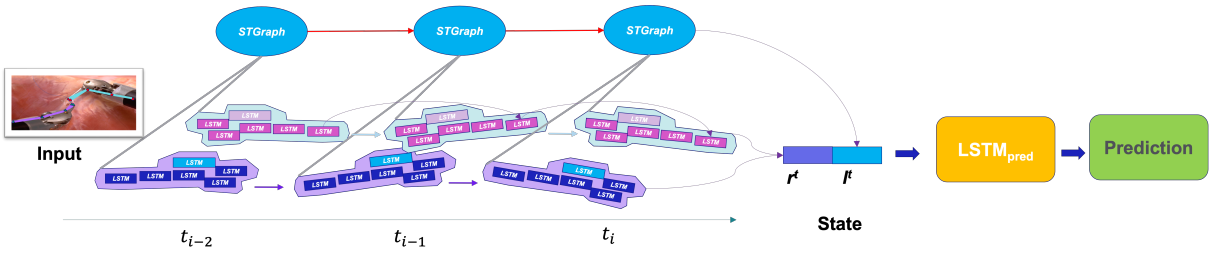

Fig. 4. Trajectory prediction pipeline

evaluation strategy includes making use of realistic surgical instrument keypoint from detection results. Such a practical input stretches the capability of tracking module on the uncertainty handling.

A. Dataset

We collected an endoscopic video dataset from multiple surgeon training sites using da Vinci surgical platform. The whole dataset contains video clips from 44 training cases with 15 different types of robotic instruments. There are total $207,542$ instrument samples annotated with defined 5 keypoint articulation in the clinic profession. The labeled points are subject to the camera view. For most scenarios, multiple instruments show up in one image frame (multi-instrument scenarios taking up to 98%).

We prepared the data in $8:1:1$ split for training, validation and testing. Due to the nature of the tracking problem, the split is conducted at video clip level, the images come from one video clip only exist either in training, validation or testing. Such a splitting strategy keeps the continuity of data and avoids potential data leakage.

For the tracking-by-prediction strategy, it is natural to adjust the temporal length and how long of the prediction horizon of the predictor. We hold the settings that two strategies are adopted, which are $O5||P5$ and $O10||P5$. The former "O" means the total of historical frames used for prediction and "P" represents the prediction horizon.

To avoid the scale ambiguity across surgery videos, all the locations are normalized into image pixel coordinates.

B. Metrics And Baseline

To evaluate the tracking accuracy, we use the standard Average Displacement Error (ADE) per frame over all tracked keypoints following the definition in [29], [25]. If the surgical instrument only partially visible, the ground-truth may have only subset of the entire articulation annotated. For this case, we count the visible keypoints in ADE calculation.

To analyze the localization performance with tracking error, it is necessary to find which keypoints at what time step suffering the tracking error, i.e. identity switches (IDS). IDS happen if two predicted keypoints with the same class label close to the same ground-truth point that confused the global label assignment. It is a typical tracking error when instruments getting too close or overcross each others, the association mis-assigned keypoint from one instrument to another in their trajectories. For this scenario, we still calculate ADE value but condition on IDS. Instead of using ground-truth annotation here, we utilize the keypoint locations as the input from a realistic surgical instrument detection module, which contains IDS as the noise. For those occluded keypoints, the detection module (Hourglass Network [18], [17]) outputs with 5-points set for each instrument with its own auto-interpolation based on the model.

For the baselines, we consider to compare with the commonly used tracking-by-detection model (SORT) [3] and social force model (social-lstm) [1]. Another baseline we chose is using the same spatial-temporal graph encoding but ignore the hierarchy. In the one level spatial-temporal graph encoding, all the keypoints are fully connected to formulate the "neighborhood". The ablation study could help to explain how the hierarchy we proposed help to optimize the interaction modeling in the complicated scene. We term our proposed solution as $STGraph_{Hierarchy}$ and the plain framework without hierarchy as $STGraph_{FC}$. To achieve a fair comparison, we apply trajectory smoothing to all the schemes over their outputs.

C. Implementation Details

The proposed system contains three LSTM modules for different encoding and decoding role. All LSTMs in the implementation only have one hidden layer. For $LSTM_{spa}$ and $LSTM_{temp}$, the latent variable is set to 32. For v_i^t in Eq. 1, the embedding dimension is 16. The two-layer graph attention is built in the model. Batch normalization layer is added before the attention layer. W in Eq. 3 at first layer is of shape 16×16 and 16×32 at second layer. Correspondingly, the dimension of a is set to 32 and 64. We use Adam optimizer over the negative log-likelihood loss with learning rate 0.005 and batchsize 64 in the entire network training phase.

D. Tracking Evaluation And Analysis

In order to evaluate the system performance, we conduct experiments first with human annotated sequences as input. The annotation is considered error- and noise-free. The main focus is testing the effectiveness of the tracking-by-prediction pipeline. Our solution received 3.083 on ADE, outperforms SORT by 81.1%. The only constrain enforced by SORT kernel is the smoothness/continuity based on short-term dynamics modeled by Kalman filter. Obviously, the motion pattern involved in surgery are much more diverse. Some of the trajectory distributions are non-Gaussian.

TABLE I

TRACKING PERFORMANCE WITH HUMAN ANNOTATED SEQUENCES AND DETECTION RESULTS.

Input	Method	ADE ↓(O5‖P5)	ADE ↓(O10‖P5)
Human Annotation Sequences (GT)	SORT	16.352	
	Social-LSTM	14.458	11.775
	STGraph_FC	7.090	5.461
	STGraph_Hierarchy	4.372	**3.083**
Keypoint Detection From Hourglass Network (IDS = 1713)	SORT	36.579	
	Social-LSTM	28.710	24.401
	STGraph_FC	16.188	11.435
	STGraph_Hierarchy	8.035	6.859

$STGraph_{Hierarchy}$ outperforms $STGraph_{FC}$ by 43.5% on ADE. Lack of the instrument level modeling, $STGraph_{FC}$ has the limitation to learn a instrument dynamic pattern considering the heterogeneous motion freedoms of different keypoints.

In the more practical setting, tracking error is a non-negligible factor. Rather than feeding the system with error-free positions, we challenged them by injecting the detection results, which contains large amount of uncertainties, such as IDS and position shift (inaccurate localization). We use the tracking evaluation method in [26] to identify IDS per frame. There are totally 1713 detected IDS in the testing sequences. Under such a condition, we conduct the tracking-by-prediction schemes again. The quantitative results are illustrated in Table. I. $STGraph_{Hierarchy}$ achieved 6.859 on ADE, outperforms the competitors by a large margin. This significance advocates effectiveness of our proposal in improving both of the tracking accuracy and robustness towards tracking errors.

For the final demonstration of tracking performance, please refer to the submitted video for a better illustration.

E. Ablation Study

The predictor in the system plays a crucial role in enhancing tracking performance by forecasting the behavior of each tracklet. It provides an extended horizon for correcting associations and implementing post-processing smoothing. To ensure a fair comparison, we assess the prediction capability of different predictors, and the results are shown in Table II. It is clearly demonstrated that the error in Social-LSTM has been propagated in the tracking, causes even worse overall localization accuracy. The study highlights how the tracking-by-prediction approach helps solve the problem by integrating an accurate predictor with the tracking task.

The second ablation presents various prediction strategies and evaluates their accuracy in terms of Average Displacement Error (ADE). It is evident that incorporating more historical information results in better predictions. It is also noted that predicting 1-step movement using the sequence input has a better accuracy. However, the limited prediction horizon worsen the data association, and results in poor overall tracking performance.

TABLE II

PREDICTION PERFORMANCE IN DIFFERENT PREDICTORS.

Predictors	ADE ↓(O5‖P5)	ADE ↓(O10‖P5)
Social-LSTM	21.432	16.535
STGraph_FC	17.990	13.414
STGraph_Hierarchy	8.519	7.805

TABLE III

PREDICTION PERFORMANCE IN DIFFERENT STRATEGIES WITH HIERARCHICAL STGRAPH.

Strategies	ADE ↓
O1‖P1	9.034
O5‖P1	8.194
O10‖P1	6.011
O5‖P5	8.519
O10‖P5	7.805

V. CONCLUSIONS

In this paper, we presented a novel tracking paradigm that integrating the trajectory prediction to constrain the data association error that propagated from the false detection. As the core component, the prediction module composed of spatial-temporal graph attention that aims to model the multi-instrument interaction in a hierarchy, benefited the overall representation learning that resulted in significant performance improvement in forecasting the instrument dynamics. Facilitated with the tracking-by-prediction scheme, the proposal strengthened to associate the observation with active trajectories in a high accuracy. Experimental results over a large multi-instrument dataset revealed the effectiveness and robustness of the work.

REFERENCES

[1] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social lstm: Human trajectory prediction in crowded spaces," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 961–971.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 3464–3468.

[4] J. Chorowski, D. Bahdanau, K. Cho, and Y. Bengio, "End-to-end continuous speech recognition using attention-based recurrent nn: First results," *arXiv preprint arXiv:1412.1602*, 2014.

[5] T. Fernando, S. Denman, A. McFadyen, S. Sridharan, and C. Fookes, "Tree memory networks for modelling long-term temporal dependencies," *Neurocomputing*, vol. 304, pp. 64–81, 2018.

[6] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Tracking by prediction: A deep generative model for mutli-person localisation and tracking," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1122–1132.

[7] P. Gao, B. Reily, R. Guo, H. Lu, Q. Zhu, and H. Zhang, "Asynchronous collaborative localization by integrating spatiotemporal graph learning with model-based estimation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 1695–1701.

[8] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social gan: Socially acceptable trajectories with generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2255–2264.

[9] I. Hasan, F. Setti, T. Tsesmelis, A. Del Bue, F. Galasso, and M. Cristani, "Mx-lstm: mixing tracklets and vislets to jointly forecast trajectories and head poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6067–6076.

[10] D. Helbing and P. Molnar, "Social force model for pedestrian dynamics," *Physical review E*, vol. 51, no. 5, p. 4282, 1995.

[11] Y. Huang, H. Bi, Z. Li, T. Mao, and Z. Wang, "Stgat: Modeling spatial-temporal interactions for human trajectory prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6272–6281.

[12] I. Ivanović and J. Jović, "Sensitivity of street network capacity under the rain impact: case study of belgrade," *Transport*, vol. 33, no. 2, pp. 470–477, 2018.

[13] P. S. Kalekar *et al.*, "Time series forecasting using holt-winters exponential smoothing," *Kanwal Rekhi school of information Technology*, vol. 4329008, no. 13, pp. 1–13, 2004.

[14] N. Liang, G. Wu, W. Kang, Z. Wang, and D. D. Feng, "Real-time long-term tracking with prediction-detection-correction," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2289–2302, 2018.

[15] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *2009 IEEE conference on computer vision and pattern recognition*. IEEE, 2009, pp. 935–942.

[16] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler, "Online multi-target tracking using recurrent neural networks," in *Thirty-First AAAI conference on artificial intelligence*, 2017.

[17] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *Advances in neural information processing systems*, vol. 30, 2017.

[18] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proceedings of 14th European Conference on Computer Vision, Part VIII 14*. Springer, 2016, pp. 483–499.

[19] O. Olabiyi, E. Martinson, V. Chintalapudi, and R. Guo, "Driver action prediction using deep (bidirectional) recurrent neural network," *arXiv preprint arXiv:1706.02257*, 2017.

[20] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *2009 IEEE 12th international conference on computer vision*. IEEE, 2009, pp. 261–268.

[21] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 300–311.

[22] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "Sophie: An attentive gan for predicting paths compliant to social and physical constraints," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 1349–1358.

[23] J. Su, P. A. Beling, R. Guo, and K. Han, "Graph convolution networks for probabilistic modeling of driving acceleration," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–8.

[24] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *2018 IEEE international Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4601–4607.

[25] X. Weng, B. Ivanovic, K. Kitani, and M. Pavone, "Whose track is it anyway? improving robustness to tracking errors with affinity-based trajectory prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6573–6582.

[26] X. Weng, J. Wang, D. Held, and K. Kitani, "3d multi-object tracking: A baseline and new evaluation metrics," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 10 359–10 366.

[27] M. Wright, "Speeding up the hungarian algorithm," *Computers And Operations Research*, vol. 17, no. 1, pp. 95–96, 1990.

[28] Y. Xu, Z. Piao, and S. Gao, "Encoding crowd interaction with deep neural network for pedestrian trajectory prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5275–5284.

[29] R. Yu and Z. Zhou, "Towards robust human trajectory prediction in raw videos," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8059–8066.