Point Tracking in Surgery-The 2024 Surgical Tattoos in Infrared (STIR) Challenge

Adam Schmidt^{1,*}, Mert Asim Karaoglu^{2,3}, Soham Sinha⁴, Mingang Jang⁵, Ho-Gun Ha⁵, Kyungmin Jung⁵,

Kyeongmo Gu⁵, Ihsan Ullah⁵, Hyunki Lee⁵, Jonáš Šerých⁶, Michal Neoral⁶, Jiří Matas⁶, Rulin Zhou^{7,9}, Wenlong

He⁹, An Wang⁸, Hongliang Ren^{7,8}, Bruno Silva^{10,11,12}, Sandro Queirós^{10,11}, Estêvão Lima^{10,11}, João L.

Vilaça^{12,13}, Shunsuke Kikuchi^{14,15}, Atsushi Kouno¹⁴, Hiroki Matsuzaki¹⁴, Tongtong Li¹⁶, Yulu Chen¹⁶, Ling Li^{16,17}, Xiang Ma¹⁶, Xiaojian Li^{16,17}, Mona Sheikh Zeinoddin¹⁸, Xu Wang¹⁸, Zafer Tandogdu¹⁹, Greg Shaw¹⁹

Evangelos Mazomenos¹⁸, Danail Stoyanov¹⁸, Yuxin Chen²⁰, Zijian Wu²⁰, Alexander Ladikos², Simon DiMaio¹,

Septimiu E. Salcudean²⁰, Omid Mohareri¹

Abstract-Understanding tissue motion in surgery is crucial to enable applications in downstream tasks such as segmentation, 3D reconstruction, virtual tissue landmarking, autonomous probe-based scanning, and subtask autonomy. Labeled data are essential to enabling algorithms in these downstream tasks since they allow us to quantify and train algorithms. This paper introduces a point tracking challenge to address this, wherein participants can submit their algorithms for quantification. The submitted algorithms are evaluated using a dataset named surgical tattoos in infrared (STIR), with the challenge aptly named the STIR Challenge 2024. The STIR Challenge 2024 comprises two quantitative components: accuracy and efficiency. The accuracy component tests the accuracy of algorithms on in vivo and ex vivo sequences. The efficiency component tests the latency of algorithm inference. The challenge was conducted as a part of MICCAI EndoVis 2024. In this challenge, we had 8 total teams, with 4 teams submitting before and 4 submitting after challenge day. This paper details the STIR Challenge 2024, which serves to move the field towards more accurate and efficient algorithms for spatial understanding in surgery. In this paper we summarize the design, submissions, and results from the challenge. The challenge dataset is available here: https://zenodo.org/records/14803158, and the code for baseline models and metric calculation is available here: https://github.com/athaddius/STIRMetrics

Index Terms—Endoscopy, Point Tracking, Deformable, Tissue Tracking, Challenge

¹ Intuitive Surgical Inc., Sunnyvale, USA, ² ImFusion GmbH, Munich, Germany, ³ Technical University of Munich, Munich, Germany, ⁴ NVIDIA, Santa Clara, USA

⁵ Division of Intelligent Robotics, Daegu Gyeongbuk Institute of Science and Technology (DGIST), South Korea

⁶ CMP Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague

⁷ Shenzhen Research Institute, China, ⁸ The Chinese University of Hong Kong, China, ⁹ Shenzhen University, China

¹⁰ Life and Health Sciences Research Institute (ICVS), School of Medicine, University of Minho, Braga, Portugal, ¹¹ ICVS/3B's - PT Government Associate Laboratory, Braga/Guimarães, Portugal, ¹² 2Ai -School of Technology, IPCA, Barcelos, Portugal 13 LASI - Associate Laboratory of Intelligent Systems, Guimarães, Portugal

¹⁴ Jmees Inc, Japan, ¹⁵ University of California - Los Angeles (UCLA),

USA ¹⁶ School of Management, Hefei University of Technology, Hefei 230009, Optimization and Intelligent Decision-China, ¹⁷ Key Laboratory of Process Optimization and Intelligent Decisionmaking, Ministry of Education, Hefei, China.

¹⁸ Hawkes Institute, University College London, London, UK, ¹⁹ Dept of Urology, University College London Hospitals, UK

²⁰ University of British Columbia

* Corresponding author: adam.schmidt at intusurg.com

I. INTRODUCTION

THE 2024 STIR challenge is designed to help improve tracking and reconstruction methods in surgery. Knowledge of tissue motion and location is critical to enable many tasks in medical computer vision [19]. Improved accuracy of motion estimation is essential to enable automated dexterity [9], autonomous scanning [24], and virtual landmarking. Improved performance here will likely also benefit foundation models, where physical priors can be incorporated into pretraining. This challenge marks the first in kind for point tracking, wherein we use infrared labels to quantify the performance of submitted methods. The data used in the challenge comprises 60 sequences with each sequence including an average of 8 points.

In this section, we will first provide a brief clarification of the challenge dataset compared to the original STIR dataset in I-A, followed by a non-exhaustive summary of datasets that we see as useful to tracking in Section I-B. After which, we will describe the dataset format and annotation protocol for the challenge in Section II. Then, we describe the metrics we calculate as part of the challenge in Section III. We then summarize all submissions received in Section IV, and their results in Section V. We provide a discussion of the results and challenge organization in Section VI, and finally conclude in Section VII. For a high-level overview of the challenge, refer to Fig. 1.

A. STIR Challenge Data (STIRC2024) vs STIR Original (STIROrig)

Here, we will explain the differences between the STIR Challenge 2024 and the STIR dataset. The original STIR dataset (STIROrig) is a dataset that is publicly released and usable for test, validation, or training (available at: https: //ieee-dataport.org/open-access/stir-surgical-tattoos-infrared).

This dataset is released as a way to validate, test, design, and evaluate algorithms [18]. STIROrig remains useful for this exact purpose, in addition to being larger than the challenge dataset. The STIR Challenge 2024 dataset (STIRC2024) is a similar dataset that was witheld from the initial STIR dataset release in order to enable proper evaluation without the risk



Fig. 1. This figure describes the STIR Challenge 2024. Participants submit their algorithms in a docker container. The algorithm receives a video and a list of start points from each sequence in the dataset. Participants use their tracker to estimate the motion of a set of points for every frame in a video. Videos are provided in a streaming manner. The final estimates are then compared to the ground truth labels (Section II) at the end of the video. The errors (Section III) are then averaged across all points to obtain the final metrics in 2D or 3D. Latency is also calculated alongside the inference for those who participated in the efficiency component of the challenge.

of participants fine-tuning or overfitting to already released data. STIRC2024 has additional filtering and removal of noisy labels, and can be used for fine grained evaluation and testing.

B. Useful Datasets for Tracking in Surgery

At MICCAI 2022, a similar challenge was organized for the same task of tracking tissue [2]. The primary differences are our use of a point-tracking metric [4], and increased size and diversity of our data. In addition, the STIR dataset is not labeled in a temporally dense manner, while the SurgT dataset is labelled per-frame.

There are many other datasets that can benefit tissue tracking. For a detailed summary of useful datasets in this space, refer to the review [19]. Since that review, additional datasets and meta-datasets have become available. Here is a brief list of data we recommend looking at. Meta-MED [1] is an assembled meta-dataset. This dataset is intended to be used for training and evaluating monocular depth models, but would also be useful for self-supervised training of tracker models. The StereoMIS [8] dataset comprises many stereo sequences and could be used for similar purposes. The SurgVU dataset [25] also serves as a large (hundreds of hours) repository of singleeye video that could be used for self-supervised training.

II. DATASET AND ANNOTATION

The STIR Challenge 2024 dataset consist of sets of stereo video clips collected on a da Vinci Xi system. Each clip

consists of a start IR image I_s and an end IR image I_e , segmentations of the fluorescent ink S_s and S_e , respectively, and the visible light clip V. All frames are of size 1280×1024 pixels. I_s , I_e are in Portable Network Graphic (png) format; V is the action video in MPEG-4 Part 14 (mp4) format; S_s , S_e are binary segmentations of the IR frames (png). This dataset comprises 60 sequences. Their average length is 8.9 seconds, with a standard deviation of 12.1 seconds. The distribution can be seen in Fig. 2. For a histogram of points per video refer to Fig. 3. No clips longer than 4 minutes are included. Summary images of the labels can be found in Fig. 4. There are a total of 496 points over the 60 sequences.

This dataset was created following the same process as STIROrig [18], and more detail can be found there. The labelling process is visually described in Fig. 5. At a high level, points are tattooed with indocyanine green (ICG) ink, to create ground truth labels. The endoscope is switched to fluourescent mode at the start and end of an action to collect the point locations at the start and end of a video. The video for tracking is recorded in white light, and multiple actions can happen within. The data comes from porcine subjects for the *in vivo* cases, and is a mix of different tissue for the *ex vivo* cases.

Segmentations are created by first thresholding the IRchannel of the image. An opening morphological transformation, which consists of erosion followed by dilation, is applied to reduce noise. The resulting segments are then verified by ensuring that if a segment appears in the start image that it



Fig. 2. Temporal distribution of videos. Most clips lie between 0 and 10 seconds, with a few longer clips > 20 seconds. Average clip length is 8.9 seconds.



Fig. 3. Number of labelled points per video. Labels can be seen in Fig. 4.

also appears in the end image. To annotate, we first evaluate visibility of markers over randomly sampled cases, ensuring the tattoos do not provide features that algorithms could track [18]. After this, a user looks through every case and removes noisy segmentation masks that result from specularity. This filtering helps to reduce label noise.

In order to compute the ground-truth 3D locations, we complete an epipolar search with normalized cross correlation, using the segmented points as candidates. This enables us to select which segment in the right image corresponds to a given segment in the left image.

The 3D position for a segment is calculated by backprojecting it. Since the left and right eyes of the endoscope do not have the same principal point, with the left at c_x and the right at c'_x , we must backproject with this in mind. We calculate depth using the baseline b, focal length f, and c_x, c'_x from the calibration along with the point x-location in the left and right image (x, x'). The depth, z, is:

$$z = \frac{b * f}{(x - c_x) - (x' - c'_x)}$$

			1	2. 2		, P		
. *		Хų.	Х.					
		•						
					a An an			
	 1. A		50 - A	• • •				а. С
			an An tao	•		•	· · ·	

Fig. 4. Start point labels for all 60 sequences in in the STIR 2024 test dataset. For each sequence, center points are extracted from each segmentation, and passed to each participant's tracker.

A. Data Format

We summarize the dataset of the STIR Challenge 2024 (STIRC2024) here, noting the format is the same as that for STIROrig [18]. STIRC2024 includes a set of 9 collection sessions, named as <%02d>, (02, 03, 04, 05, 06, 07, 08, 09, 11). There are 5 *in vivo* sessions (03, 04, 07, 08, 11) and 4 *ex vivo* sessions (02, 05, 06, 09). Each session includes multiple sequences. An example sequence for one of the *in vivo* cases is shown in Fig. 6.

left

starticg.png (Infrared ICG image of start
frame)

endicg.png (Infrared ICG image of end frame)
segmentation/startim.png,

segmentation/endim.png (Filtered and segmented binary versions of ICG start and end image) frames/<ms>_ms.mp4 (video file)

right

```
starticg.png
endim.png
frames/<ms>_ms.mp4 (video file)
```

calib.json Camera calibration parameters (intrinsics, relative stereo pose translation in metres and axis-angle rotation format)

The video file names include start and end capture times in milliseconds.

III. METRICS

In this challenge, we evaluated the submitted algorithms based on two different metrics: accuracy and efficiency. The accuracy metric is important for clinical verification. The efficiency metric measures the timing latency of the submitted algorithms and evaluates an algorithm's feasibility in running on clinical systems. For the accuracy metric, we used two categories: 2D trackers, and 3D trackers.

A. Accuracy

To evaluate accuracy, we use a metric which manages outliers well via calculating accuracy over multiple thresholds. This metric can be used easily in 2D and 3D. The metric is δ^{avg} , introduced in TAP-Vid [4], which is a non-medical



Fig. 5. Dataset labels and label creation process. The ground truth is collected by using a tattoo needle to label points at the start and end of video frames. After tattooing is completed, multiple sequences can be collected. For each sequence, the camera captures an image in infrared (ground truth start frame), then switches to white light. Actions are performed under white light, and this video is recorded and saved. Then the camera switches back to IR and captures the end frame which is used as the ground truth for each point's motion. Segments are the binary-thresholded IR images; tattooed regions are shown in white. On the right is a figure showing a set of random triplets with the triplet: (IR image, visible light image, segment/GT image) for each point shown.

point tracking challenge. In TAP-Vid [4], the points also have an occlusion score, but here we use data in which the points are unoccluded at the end frame. In our scenes points can be occluded and reappear during a sequence due to camera movement, instrument-tissue occlusion, or tissue-tissue occlusion.

To calculate the metric, δ^{avg} , in our case, each algorithm estimates the position for a point (or multiple points) for each frame in a video in a streaming manner. The final frame location estimate results in a point (or multiple points), \hat{p}_{end} . The calculated finish points, \hat{p}_{end} , are 2 dimensional for 2D trackers, and 3 dimensional for the 3D trackers. The accuracy metric is averaged across all points and thresholds, with each point weighted evenly. To calculate euclidean distance, each point is matched to its nearest point in the end point label.

$$\delta^{avg} = \sum_{i=i}^{M} \delta^{\mathbf{l}_i} / M \tag{1}$$

$$\delta^{\mathbf{l}_i} = \sum_{\hat{p}_{end}} \mathbb{1}(d(\hat{p}_{end}, p_{end}^{nearest}) < \mathbf{l}_i)/N \tag{2}$$

1 is the indicator function, used to count the amount of points under the distance threshold. d() is a function to estimate euclidean distance in the dimension of input (2D/3D). N is the total number of points across all videos, and M is the number of thresholds used. Thus, δ^{l_i} is accuracy at the threshold l_i . For 2D, the thresholds are l = [4, 8, 16, 32, 64] with units as pixels in the full 1024×1280 image. For 3D, the thresholds are l = [2, 4, 8, 16, 32] with units as millimetres.

B. Efficiency

The efficiency of an algorithm is measured by its computational latency, assessed across all video frames to derive a latency distribution. While the mean latency provides a general efficiency metric, it fails to capture worst-case behavior. In real-world surgical point tracking, predictability depends on worst-case and tail latencies. Therefore, we evaluate efficiency using the 95th and 99th percentile latencies in addition to the mean. The final efficiency score is the average of these three metrics, offering a comprehensive assessment of both real-time and practical performance. A submission is considered for the efficiency category only if the accuracy of the algorithm is above a certain threshold.

IV. SUBMISSIONS + BASELINES

Here, we summarize submissions to the challenge. The submissions are grouped by those that who were participants for the challenge day (pre-challenge), who were viable for prizes, and those afterwards (post-challenge). We also provide the results from the baseline methods that we provided on our github page, https://github.com/athaddius/STIRMetrics.

A. Baselines

1) MFT: This is the baseline MFT method [16]. This method runs optical flow between a frame and multiple frames at skips into the past. The algorithm selects its optimal trajectory by selecting the highest certainty unoccluded trajectory. RAFT is used as the optical flow method in. To maintain inference efficiency, images are downsampled by a factor of 2 for tracking. Skip factors are the same as those used in the MFT paper $[-\infty, 1, 2, 4, 8, 16, 32]$ The occlusion threshold is set as 0.02. After tracking, the locations are scaled up by 2 to get the coordinates in full resolution.

2) CSRT: This method uses the Channel and Spatial Reliability tracker (CSRT [15]), initialized with a region of interest around each point in the first frame. This tracker uses a correlation based adaptive template matching to track a point across a video. Tracking is performed on half scale images, and upscaled. The region of interest for each point is a 29×29 box with its center as the point location.



Fig. 6. Example *in vivo* sequence (session 04, sequence 10) from the STIR Challenge 2024 test set. The start segmentation (middle, left) is converted into a set of points, 5 in this case, that are passed to the algorithm which tracks in the white light stereo video (top). The algorithm results are compared to the ground truth end points (middle, right).

3) *RAFT*: In the RAFT implementation, we use RAFT [22] off-the-shelf to track points from one frame to the next in a streaming manner. We track on half-scaled images, for efficiency, and multiply the final tracking result by 2 to obtain full-resolution results. RAFT internally iterates multiple times, refining estimates with iteration. We use 12 iterations in inference.

4) RAFT + RAFT Stereo (3D): This is the 3D baseline, which uses RAFT to estimate flow from one frame to the next in the left eye, and finds the 3D position by backprojecting the points using the disparity estimate alongside the camera calibration. The disparity estimate is calculated using RAFT-Stereo [13].

5) Control: The control method estimates 0 motion for every point. This provides a minimum bound of accuracy which is useful for debugging and static scenes. This also allows the organisers and participants to ensure the challenge methods use the correct data. The control method runs alongside submissions. During the challenge, this served as a useful sanity check.

B. Challenge Day

Four teams were able to submit in time for challenge day. 1) Team ICVS_2AI: Team ICVS_2AI proposes an occlusion-aware optical flow-based solution. Targeting to tackle the problem of labeled-data-scarcity in surgical domain, the method employs ARFlow as the optical flow model trained on the SurgT dataset in a self-supervised fashion following previous work [21]. The proposed architecture, which is



Fig. 7. Overview of the solution provided by team ICVS_2AI. The postprocessing block in the baseline method utilized forward-backward occlusion masks.

depicted in Fig. 7, first crops a 512×512 region around the location of the last track location, then estimates the optical flow between the source target frames. To prioritize runtime performance, for the efficiency component, the point tracks are only computed between previous frame t - 1 and current frame t; however, in the 2D accuracy challenge, frame t - 2 is also used as a secondary source frame and the final point track is estimated as the weighted average of the two computations. For the 3D challenge, the estimated 2D point track is lifted to the 3D space by stereo-depth computation applying the flow-model between left-right image pairs.

2) Team MedTrack: Team MedTrack proposes a two-step, hierarchical, long-term tracking method called Dynamic Multi-Frame Point Tracking (DMPTracking) depicted in Fig. 8. In the first step, DMPTracking employs an MFT-based [16] approach to estimate point tracks and their visibilities at a coarse level. Following MFT's [16] original structure, as explained in further detail in Section IV-A1, it computes the optical flow, uncertainty, and visibility scores between



Fig. 8. Overview of the solution provided by team MedTrack: DMPTracking. First, it uses coarse optical flow to provide an initial tracking position Pt for the query points P0, and then refines the tracking results using multi-frame information.



Fig. 9. Overview of the solution provided by team CCG_DGIST that combines sparse feature matching and dense optical flow estimation.

the current frame and a set of geometrically distanced past, template, frames and estimates the track of a point through a temporal chaining mechanism. However, this template-based optical flow mechanism can provide erroneous solutions if there are large brightness changes between the templates and the current frame. DMPTracking utilizes a filtering mechanism that checks the magnitude of the spatial gradient of the optical flow map and replaces the values that are above a threshold with interpolated values from the local neighborhood.

In the second step, DMPTracking utilizes a CoTrackerbased [11] structure to refine the coarse-solution. As an endto-end learning based approach, CoTracker [11] processes a video sequence in chunks with a window size of 8 frames and in each step slides the window by a stride of 4 frames. In certain cases, such as occlusions spanning multiple windows or long sequences, this window-based processing can cause failures in tracking. To prevent this, the proposed architecture alters the structure of the sliding window, and instead fixes the first and last frames of it to be, respectively, the initial frame of the entire sequence and the current frame. The remaining 6 frames are chosen from previously observed frames.

3) Team Jmees: Team Jmees employs MFT [16] for both 2D and 3D tracking tasks. For 2D tracking, the model processes input frames at 1/4 of their original size to enhance computational efficiency. In the 3D tracking task, points are tracked independently in the left and right frames and subsequently lifted into 3D space using the disparity computed between them.

4) Team CCG_DGIST: Team CCG_DGIST proposes a joint sparse keypoint matching and dense optical approach, depicted in Fig. 9. The optical flow model consecutively processes frames to estimate and store the point tracks. Concurrently, the sparse keypoint matching method is utilized to estimate homographic transformations of tracked points between the initial and target frames. The final track update is selected based on a decision algorithm. In more detail, the optical flow based decision is selected if a matched keypoint exists within a certain threshold of the optical flow estimation. In the case of detection of a drift of the optical flow estimation is selected as a correction. Finally, if the displacement between the keypoint and the optical flow based method is above a certain threshold the keypoint-based

estimation is selected.

C. Post-Challenge Day

1) Team CTUPrague: Team CTUPrague participates with their MFT [16]-extending method, MFTIQ [20]. Similar to MFT, MFTIQ uses the optical-flow chaining structure and replaces the implicit occlusion and uncertainty estimation of the optical flow model with an independent network that aggregates the warped feature maps with a feature similarity cost map to compute the quality and occlusion scores. This renders further adaptability of the architecture with various flow estimation backends. In this challenge, their submissions utilize two separate flow models: SEA-RAFT [23] and ROMA [7].

2) Team UBC_RCL: Team UBC_RCL participates with A-MFST [3]. Extending the flow-chaining architecture of MFT [16], A-MFST replaces the RAFT [22]-based optical flow backend with SENDD [17] and the certainty and occlusion estimation with backward-forward flow consistency. This consistency score is also used for frame selection to prioritize more reliable template curation. Instead of storing input images as templates, as in MFT [16], A-MFST caches previously computed features to reduce redundant computations in the flow-estimation.

3) Team SRV: Team SRV participates with a frame-toframe tracking method, SEA-RAFT [23]. Focusing on efficiency improvements, SEA-RAFT extends over the prior work RAFT [22] and introduces a more efficient architecture. Coupled with a more robust training strategy, it additionally shows accuracy improvements.

4) Team CUHK: Team CUHK participates with TAP-Endo, a semi-supervised method extending MFT [16]. The proposed architecture utilizes SEA-RAFT [23] to replace MFT's original optical flow component. The architecture uses a three step training strategy to improve generalizibility to endoscopic scenes. Their method is shown in Fig. 10. In the first step, a pretrained SEA-RAFT with frozen weights are appended with a gated attention module and trained in supervised fashion on synthetic datasets to predict the uncertainty and occlusion. In the second step, a part of the SEA-RAFT backbone is unfrozen and refined using endoscopic datasets [2], [18] in a



Fig. 10. Overview of the solution provided by team CUHK: TAP-Endo. It consists of a three step training structure: (1) supervised training of the uncertainty and occlusion block; (2) fine-tuning of the optical flow model using endoscopic data in a self-supervised fashion; (3) fine-tuning of the uncertainty and occlusion block leveraging pseudo-labels acquired using a set of state-of-the-art point tracking methods.

 $\begin{array}{c} \text{TABLE I} \\ \text{2D Tracking accuracy comparison. } \delta^{l_i} \text{ indicates the tracking} \\ \text{accuracy where } l_i \text{ is the threshold defined in pixels.} \end{array}$

Method	$\delta^{\mathbf{l}_i} \uparrow$							
	$\mathbf{l}_i = 4$	$\mathbf{l}_i = 8$	$l_i = 16$	$l_i = 32$	$l_i = 64$	Avg.		
Baselines								
RAFT	07.26	19.56	39.92	64.92	80.44	42.42		
CSRT	22.78	47.38	67.14	74.80	81.05	58.63		
MFT	42.54	69.36	86.49	93.35 96.37		77.62		
MICCAI Submissions								
ICVS_2AI	25.40	51.01	74.19	88.71	92.54	66.37		
JMEES	26.00	54.44	77.42	91.73	95.36	68.99		
CCG_DGIS	ST36.09	63.51	83.87	90.93	94.36	73.75		
MedTrack	38.91	67.54	86.69	93.15	95.97	76.45		
Post MICCAI Submissions								
SRV	07.86	18.15	31.65	47.78	66.73	34.44		
UBC_RCL	20.57	42.34	66.73	77.42	85.89	58.59		
MFTIQ-	42.34	68.95	85.89	92.14	94.76	76.82		
SEARAFT								
MFTIQ-	44.36	69.56	85.89	91.13	95.16	77.22		
ROMA								
CUHK	40.93	68.95	87.50	93.15	96.37	77.38		

self-supervised fashion [14]. In the final step, the occlusion and uncertainty heads are fine-tuned using pseudo-labels generated by a set of state-of-the-art point tracking methods [10], [16], [12], [6].

V. RESULTS

This section will detail and analyze the results for each challenge participant.

A. Accuracy

For the 2D methods, Table I provides the overarching summary. Of the ranked challenge submissions (non-post-challenge), Team Medtrack came in first, with a $\delta^{avg} = 76.45$. Team CCG_DGIST was second with a $\delta^{avg} = 73.75$, and

 $\begin{array}{c} \text{TABLE II} \\ \text{3D Tracking accuracy comparison. } \delta^{l_i} \text{ indicates the tracking} \\ \text{accuracy where } l_i \text{ is the threshold defined in pixels.} \end{array}$

ACCURACT WHERE I ₁ IS THE THRESHOLD DEFINED IN FIXELS.								
Method	$\delta^{\mathbf{l}_i}\uparrow$							
	$\mathbf{l}_i = 2$	$\mathbf{l}_i = 4$	$\mathbf{l}_i = 8$	$l_i = 16$	$l_i = 32$	Avg.		
Baselines								
RAFT- Stereo	13.94	36.16	60.40	79.80	91.51	56.36		
MICCAI Submis- sions								
JMEES ICVS_2AI	25.70 27.88	45.40 55.15	65.31 75.96	81.16 91.11	91.01 97.58	61.71 69.54		
200 180 160 140 (m) 120 100 80 60 40 20 0 —	ł	• mea	n • 95th • 3	99th	144.6	3ms		
0	1	2	3 Tost Sogur	4	5			

Fig. 11. Efficiency Result of ICVS_2AI's method.

Team Jmees came in third with a $\delta^{avg} = 68.99$. The overall best performing method was MFT [16].

For the 3D methods, refer to the summary Table II. Of the ranked challenge submissions, Team ICVS_2AI came in first with a $\delta^{avg} = 69.54$, and Team Jmees came in second with a $\delta^{avg} = 61.71$.

B. Efficiency

In this section we summarize the efficiency results. Only the ICVS_2AI team participated in the efficiency component. To assess their algorithm's efficiency, we evaluated selected test cases from our dataset, measuring the mean, 95th percentile, and 99th percentile latencies on an NVIDIA A100 GPU. The final latency score averaged across our test cases was 144.63 ms, corresponding to 7 FPS, which is within the acceptable range for many surgical point-tracking applications. Fig. 11 shows the latencies of the ICVS_2AI submission for 5 test sequences. The maximum latency remains below 200 ms, a critical threshold in this inaugural efficiency evaluation in this year's challenge. Broader participation in the efficiency component would provide deeper insights into the deployability of competing algorithms.

VI. DISCUSSION

In this section, we will summarize the numerical results from the challenge in Section VI-A, by takeaways from organizing the challenge in Section VI-B, and finish with suggestions for future algorithms and areas we believe could be useful in Section VI-C.

A. Accuracy and Efficiency

Here we discuss the 2D results, followed by the 3D results and efficiency.

The most accurate method in the 2D component of the challenge was the baseline MFT [16] method. In terms of best fine-grained performance (at $\delta = 4$), the baseline MFT also achieved the highest accuracy. This could be due to the intrinsic memory of MFT which allows it to search back to the first frame. Methods that encountered difficulty over sequences in this challenge were those which do not support longer window recovery (RAFT (baseline), SEA-RAFT (Team SRV)). Frame-wise methods (or methods with short windows) still attained reasonable performance when integrating occlusion masks (Team ICVS_2AI).

Regarding long-term methods, CCG DGIST addressed long term tracking by having two branches to decide between an optical flow method and a homography transformation to redetect points after they may be occluded. To deal with longer occlusion, Team Medtrack uses MFT along with CoTracker. They alter the sliding window of CoTracker to include the initial frame to help performance under long time spans. In Medtrack's submission, MFT is used to calculate initial positions. The optical flow for MFT is also filtered to remove outliers in non-smooth regions using the spatial gradients of the flow map. The candidates from MFT are then used as initial points to seed CoTracker. Team Medtrack attained the second highest score among the methods submitted for 2D accuracy tracking. The TAP-Endo method submitted by Team CUHK extends SEA-RAFT [23] with an occlusion and uncertainty module and fine-tunes it using a semi-supervised strategy. They integrate their method into the MFT architecture for robust point tracking with a longer temporal context allowing them to achieve the highest accuracy of all the submissions for the task of 2D tracking.

Interestingly, Team MFTIQ [20] (ROMA/SEA-RAFT) and TAP-Endo both achieved competitive performance but did not surpass the baseline MFT. TAP-Endo (Team CUHK) employed a domain-adaptation strategy, yet still fell short of outperforming the baseline. This outcome may indicate that existing backbone optical flow architectures and refinement strategies, generalize suboptimally or inconsistently to the unique challenges posed by the endoscopic domain. We expect further exploration and optimization in this area could significantly enhance performance and robustness in future.

Notably, most of these methods do not train a long-term tracking or occlusion management method. ICVS trains an optical flow on SurgT, and UBC_RCL trains an optical flow method on image pairs. Although Medtrack uses CoTracker with the start frame as the first frame in each sliding window, the CoTracker model is not trained under these surgical scenarios. It could be expected that more focus on the occlusion management and re-detection could improve performance in future submissions.

In terms of 3D methods, Team ICVS_2AI had the most accurate submission. Team ICVS_2AI tracks in 2D and uses this 2D track along with estimated depth to obtain 3D locations. Team Jmees tracks a point in each eye, and uses

the disparity between these points to backproject. In terms of future work, a method could be envisioned which either: tracks directly in 3D, or communicates between both the left and right tracks via a transformer-like model or a simpler classical consensus filtering. In the future, methods that filter or use both frames should be able produce better 3D results in addition to improving tracking in the 2D frame space. We believe this since more information is available in the stereo data. As a simple example, a point may be occluded in one eye but not in the other.

Finally, for the efficiency component, although there was only one submission, the emphasis on efficiency in the submissions helped to ensure that the methods are clinically feasible. This component alongside the constraint that methods must run in a streaming manner, in which they are unable to see future frames, helps to focus the challenge on algorithms that could be usable in a surgical system.

B. Challenge Takeaways

We have a few takeaways from the challenge organization as a whole, and we will summarize them here. For future iterations, we will make the docker submission process more clear in order to better enable quick evaluation without having to debug submissions with challenge teams. Some teams ran into memory issues with validation on the STIROrig dataset since the dataloader we provide loads videos fully into memory. Providing a simpler and more lightweight evaluation code structure should help to fix this.

In terms of 3D participation and efficiency, we saw lower participation, and believe this is due to the ease of use for implementing methods within these frameworks. In the future, we will look to provide more detailed documentation for every component of the challenge.

C. Algorithmic Directions

For future methods, here is a brief list of ideas that can be focused on:

- Using pseudolabelling, synthetic data augmentation, student teacher models applied to surgical scenarios (ie. CoTracker3 [10], Bootstap [5])
- Pretraining models to use surgical features (Maskedautoencoders, self-supervision, etc.)
- Using stereo data to improve tracking, rather than tracking in a single eye (left/right).
- Training efficient models for long-term tracking, and relocalization.

VII. CONCLUSION

In this paper, we summarized and discussed the design, data, participation, and results from the 2024 STIR Challenge, which was organized as a part of EndoVis at MICCAI 2024. We expect this challenge, and the publicly released test dataset will serve as a high-quality resource for methods to test, compare, and iterate on algorithms for tissue tracking and other applications. The field of image guidance in surgery, and many other applications depend on accurate methods, and see this work as a key step in continuing to enable surgical applications.

REFERENCES

- [1] Budd, C., Vercauteren, T.: Transferring Relative Monocular Depth to Surgical Vision with Temporal Consistency. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024: 27th International Conference, Marrakesh, Morocco, October 6–10, 2024, Proceedings, Part VI. pp. 692–702. Springer-Verlag, Berlin, Heidelberg (Oct 2024)
- [2] Cartucho, J., Weld, A., Tukra, S., Xu, H., Matsuzaki, H., Ishikawa, T., Kwon, M., Jang, Y.E., Kim, K.J., Lee, G., Bai, B., Kahrs, L.A., Boecking, L., Allmendinger, S., Müller, L., Zhang, Y., Jin, Y., Bano, S., Vasconcelos, F., Reiter, W., Hajek, J., Silva, B., Lima, E., Vilaça, J.L., Queirós, S., Giannarou, S.: SurgT challenge: Benchmark of soft-tissue trackers for robotic surgery. Medical Image Analysis **91**, 102985 (Jan 2024). https://doi.org/10.1016/j.media.2023.102985
- [3] Chen, Y., Wu, Z., Schmidt, A., Salcudean, S.E.: A-mfst: Adaptive multiflow sparse tracker for real-time tissue tracking under occlusion. arXiv preprint arXiv:2410.19996 (2024)
- [4] Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., Yang, Y.: TAP-Vid: A Benchmark for Tracking Any Point in a Video. In: Advances in Neural Information Processing Systems. vol. 35, pp. 13610–13626 (Dec 2022)
- [5] Doersch, C., Luc, P., Yang, Y., Gokay, D., Koppula, S., Gupta, A., Heyward, J., Rocco, I., Goroshin, R., Carreira, J., et al.: Bootstap: Bootstrapped training for tracking-any-point. In: Proceedings of the Asian Conference on Computer Vision. pp. 3257–3274 (2024)
- [6] Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J., Zisserman, A.: Tapir: Tracking any point with per-frame initialization and temporal refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10061–10072 (2023)
- [7] Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: Roma: Robust dense feature matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19790– 19800 (2024)
- [8] Hayoz, M., Hahne, C., Gallardo, M., Candinas, D., Kurmann, T., Allan, M., Sznitman, R.: Learning how to robustly estimate camera pose in endoscopic videos. International Journal of Computer Assisted Radiology and Surgery (2023). https://doi.org/10.1007/s11548-023-02919-w
- [9] Kam, M., Wei, S., Opfermann, J., Saeidi, H., Hsieh, M., Wang, K., Kang, J., Krieger, A.: Autonomous System for Vaginal Cuff Closure via Model-Based Planning and Markerless Tracking Techniques. IEEE Robotics and Automation Letters 8(7) (2023)
- [10] Karaev, N., Makarov, I., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. arXiv preprint arXiv:2410.11831 (2024)
- [11] Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: European Conference on Computer Vision. pp. 18–35. Springer (2025)
- [12] Li, H., Zhang, H., Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, L.: Taptr: Tracking any point with transformers as detection. In: European Conference on Computer Vision. pp. 57–75. Springer (2024)
- [13] Lipson, L., Teed, Z., Deng, J.: Raft-stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021)
- [14] Liu, L., Zhang, J., He, R., Liu, Y., Wang, Y., Tai, Y., Luo, D., Wang, C., Li, J., Huang, F.: Learning by analogy: Reliable supervision from transformations for unsupervised optical flow estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6489–6498 (2020)
- [15] Lukezic, A., Vojir, T., Zajc, L.C., Matas, J., Kristan, M.: Discriminative Correlation Filter with Channel and Spatial Reliability. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4847–4856. IEEE, Honolulu, HI (Jul 2017)
- [16] Neoral, M., Šerých, J., Matas, J.: Mft: Long-term tracking of every pixel. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6837–6847 (2024)
- [17] Schmidt, A., Mohareri, O., DiMaio, S., Salcudean, S.E.: Sendd: Sparse efficient neural depth and deformation for tissue tracking. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 238–248. Springer (2023)
- [18] Schmidt, A., Mohareri, O., DiMaio, S., Salcudean, S.E.: Surgical Tattoos in Infrared: A Dataset for Quantifying Tissue Tracking and Mapping. IEEE Transactions on Medical Imaging (2024). https://doi.org/10.48550/arXiv.2309.16782
- [19] Schmidt, A., Mohareri, O., DiMaio, S., Yip, M.C., Salcudean, S.E.: Tracking and mapping in medical computer vision:

A review. Medical Image Analysis **94**, 103131 (May 2024). https://doi.org/10.1016/j.media.2024.103131

- [20] Serych, J., Neoral, M., Matas, J.: Mftiq: Multi-flow tracker with independent matching quality estimation. arXiv preprint arXiv:2411.09551 (2024)
- [21] Silva, B., Queirós, S., Fernández-Rodríguez, M., Oliveira, B., Torres, H.R., Morais, P., Buschle, L.R., Correia-Pinto, J., Lima, E., Vilaça, J.L.: Evaluating unsupervised optical flow for keypoint tracking in laparoscopic videos. In: Medical Imaging 2024: Image-Guided Procedures, Robotic Interventions, and Modeling. vol. 12928, pp. 419–427. SPIE (2024)
- [22] Teed, Z., Deng, J.: Raft: Recurrent all-pairs field transforms for optical flow. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 402– 419. Springer (2020)
- [23] Wang, Y., Lipson, L., Deng, J.: Sea-raft: Simple, efficient, accurate raft for optical flow. In: European Conference on Computer Vision. pp. 36– 54. Springer (2024)
- [24] Zhan, J., Cartucho, J., Giannarou, S.: Autonomous tissue scanning under free-form motion for intraoperative tissue characterisation. In: 2020 IEEE international conference on robotics and automation (ICRA). pp. 11147–11154. IEEE (2020)
- [25] Zia, A., Berniker, M., Nespolo, R., Perreault, C., Wang, Z., Mueller, B., Schmidt, R., Bhattacharyya, K., Liu, X., Jarc, A.: Surgical Visual Understanding (SurgVU) Dataset (Jan 2025). https://doi.org/10.48550/arXiv.2501.09209