Self-supervised Sim-to-Real Kinematics Reconstruction for Video-Based Assessment of Intraoperative Suturing Skills

Loc Trinh[1], Tim Chu[1], Zijun Cui[1], Anand Malpani[2], Cherine Yang[3], Istabraq Dalieh[4], Alvin Hui[5], Oscar Gomez[1], Yan Liu[1], and Andrew Hung[3(✉)]

[1] University of Southern California, Los Angeles, USA
{loctrinh,tnchu,zijuncui,gomez,yanliu.cs}@usc.edu
[2] Mimic Technologies Inc., Seattle, USA
malpani.anand.89@gmail.com
[3] Department of Urology, Cedars-Sinai Medical Center, Los Angeles, USA
{cherine.yang,andrew.hung}@cshs.org
[4] Boston University Henry M. Goldman School of Dental Medicine, Boston, USA
idalieh@bu.edu
[5] Western University of Health Sciences, Pomona, USA
alvin.hui@westernu.edu

Abstract. Suturing technical skill scores are strong predictors of patient functional recovery following robot-assisted radical prostatectomy (RARP), but manual assessment of these skills is a time and resource-intensive process. By automating suturing skill scoring through computer vision methods, we can significantly reduce the burden on healthcare professionals and enhance the quality and quantity of educational feedbacks. Although automated skill assessment on simulated virtual reality (VR) environments have been promising, applying vision methods to live ('real') surgical videos has been challenging due to: 1) the lack of kinematic data from the da Vinci® surgical system, a key source of information for determining the movement and trajectory of robotic manipulators and suturing needles, and 2) the lack of training data due to the labor-intensive task of segmenting and scoring individual stitches from live videos. To address these challenges, we developed a self-supervised pre-training paradigm whereby sim-to-real generalizable representations are learned without requiring any live kinematic annotations. Our model is based on a masked autoencoder (MAE), termed as *LiveMAE*. We augment live stitches with VR images during pre-training and require LiveMAE to reconstruct images from both domains while also predicting the corresponding kinematics. This process learns a visual-to-kinematic mapping that seeks to locate the positions and orientations of surgical manipulators and needles, deriving "kinematics" from live videos without requiring supervision. With

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_68.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 708–717, 2023.
https://doi.org/10.1007/978-3-031-43996-4_68

an additional skill-specific finetuning step, LiveMAE surpasses supervised learning approaches across 6 technical skill assessments, ranging from 0.56–0.84 AUC (0.70–0.91 AUPRC), with particular improvements of 35.78% in AUC for *wrist rotation* skills and 8.7% for *needle driving* skills. Mean-squared error for test VR kinematics was as low as 0.045 for each element of the instrument poses. Our contributions provide the foundation to deliver personalized feedback to surgeons training in VR and performing live prostatectomy procedures.

Keywords: Vision transformers · Masked autoencoders · Self-supervised learning · sim-to-real generalization · Suturing skill assessment

1 Introduction

Previous studies have shown that surgeon performance directly affects patient clinical outcomes [1,2,13,14]. In one instance, manually rated suturing technical skill scores were the strongest predictors of patient continence recovery following a robot-assisted radical prostatectomy compared to other objective measures of surgeon performance [3]. Ultimately, the value of skill assessment is not only in its ability to predict surgical outcomes, but also in its function as formative feedback for training surgeons. The need to automate skills assessment is readily apparent, especially since manual assessments by expert raters are subjective, time-consuming, and unscalable [4,5]. View Fig. 1 for problem setup.

Preliminary work has shown favorable results for automated skill assessments on simulated VR environments, demonstrating the benefits of machine learning (ML) methods. ML approaches for automating suturing technical skills leveraged instrument kinematic (motion-tracking) data as the sole input to recurrent networks have been able to achieve effective area-under-ROC-curve (AUC), up to 0.77 for skill assessment in VR sponge suturing exercises [6]. Multi-modality approaches that fused information from both kinematics and video modalities have demonstrated increased performance over uni-modal approaches in both VR sponge and tube suturing exercises, reaching up to 0.95 AUC [7].

Despite recent advances, automated skill assessment in live scenarios is still a difficult task due to two main challenges: 1) the lack of kinematic data from the da Vinci® system, and 2) the lack of training data due to the labor-intensive labeling task. Unlike simulated VR environments where kinematic data can be readily available, current live surgical systems do not output motion-tracking data, which is a key source of information for determining the movement and trajectory of robotic manipulators and suturing needles. Moreover, live surgical videos do not have a clear and painted target area for throwing stitches, unlike VR videos, which makes the task additionally difficult. On the other hand, due to the labor-intensive task of segmenting and scoring individual stitches from each surgical video, the quantity of available and labeled training data is quite low, rendering traditional supervised learning approaches ineffective.

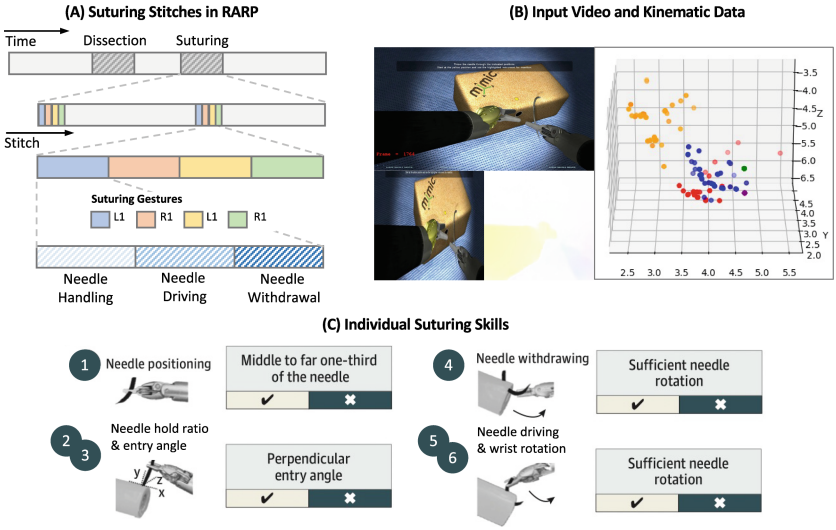

Fig. 1. Suturing skill assessment. (a) The suturing step of RARP is composed of multiple stitches, each of which can also be broken down into three sub-phases (needle handling, driving, and withdrawal). (b) Input video and kinematics data for each VR suturing exercise. Live data do not have kinematics. Colors indicate different instruments such as left/right manipulators and needle/targets. (c) Each sub-phrase can be divided into specific EASE skills [8] and assessed for their quality (low vs high).

To address these challenges, we propose LiveMAE which learns sim-to-real generalizable representations without requiring any live kinematic annotations. Leveraging available video and sensor data from previous VR studies, LiveMAE can map from surgical images to instrument kinematics and derive surrogate "kinematic" automatically by learning to reconstruct images from both VR and live stitches while also predicting the corresponding VR kinematics. This creates a shared encoded representation space between the two visual domains while using available kinematic data from only one domain, the VR domain. Moreover, our pre-training strategy is not skill-specific which brings a bonus in improving data efficiency. LiveMAE enjoys up to six times more training data across the six suturing skills seen in Fig. 1c, especially when we further break down video clips and kinematic sequences into multiple (image, kinematic) pairs.

Overall, our main contributions include:

1. We propose LiveMAE which learns sim-to-real generalizable representations without requiring any live kinematic annotations.
2. We design a pre-training paradigm that increases the number of effective training samples significantly by combining data across suturing skills.
3. We conduct rigorous evaluations to verify the effectiveness of LiveMAE on surgical data collected and labeled across multiple institutions and surgeons. Finetuning on suturing skill assessment tasks yields better performance on 5/6 skills on live surgical videos compared to supervised learning baselines.

2 Methodology

Masked autoencoding is a method for self-supervised pre-training of Vision Transformers (ViTs [12]) on images. It has demonstrated the capability to learn efficient and useful visual representations for downstream tasks such as image classification and segmentation. Our model builds on top of mask autoencoders (MAEs) and we provide a preliminary intro for MAE in Appendix 1.1.

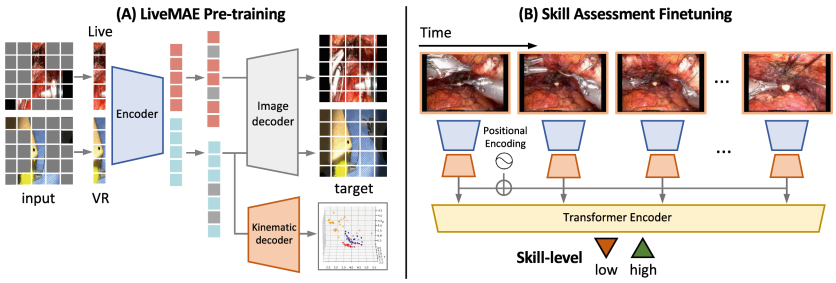

Fig. 2. LiveMAE Overview. (a) Pre-training with a shared encoder between Live and VR images and a kinematic decoder for predicting instrument kinematics. (b) Skill-specific finetuning for suturing skill assessment using pre-trained LiveMAE mapping.

The input to our system contains both VR and live surgical data. VR data for a suturing skill s is defined as $D_s^{VR} = \{(x_i, k_i, y_i)\}_{i=0}^{N_s}$ consisting of segmented video clips $x_i \in \mathbb{R}^{F\times H\times W\times 3}$, aligned kinematic sequence $k_i \in \mathbb{R}^{F\times 70}$, and EASE technical skill score $y_i \in \{0, 1\}$ for non-ideal vs ideal performance. F denotes the number of frames in the video clip. Live data for s is similarly $D_s^L = \{(x_i, y_i)\}_{i=0}^{M_s}$, except there are no aligned kinematics. Kinematic data has 70 features tracking 10 instruments of interest, each pose contains 3 elements for coordinates and 4 elements for quarternions. There are six technical skill labels, see Fig. 1c.

2.1 LiveMAE

Since D_s^L lacks kinematic information that is crucial for suturing skill assessment, we propose LiveMAE to automatically derive "kinematics" from live videos that can be helpful for downstream prediction. Specifically, we aim to learn a mapping $\phi : \mathbb{R}^{H\times W\times 3} \rightarrow \mathbb{R}^{70}$ from images to instrument kinematics using available video and sensor data from D_s^{VR}, and subsequently utilizing that mapping ϕ on live videos. Although the visual style between VR and live surgical videos can differ, this mapping is possible since we know that both simulated VR and live instruments share the exact same dimensions and centered coordinate frames. Our method builds on top of MAE and has three main components: a kinematic decoder, a shared encoder, and an expanded training set.

Kinematic Decoder. For mapping from a surgical image to the instrument kinematics, we propose an additional kinematic output head along with a corresponding self-supervised task of reconstructing kinematics from masked input. See Fig. 2a. The kinematic decoder is also a lightweight series of Transformer blocks that takes in a full sequence of both the (i) encoded visible patches, and (ii) learnable mask tokens. The last layer of the decoder is a linear projection whose number of output channels equals to 70, the dimension of the kinematic data. Similar to the image reconstruction task, which aims to learn visual concepts and semantics by encoding them into a compact representation for reconstruction, we additionally require these representations to contain information regarding possible poses of the surgical instruments. The kinematic decoder also has a reconstruction loss, which computes the mean squared error (MSE) between the reconstructed and original kinematic measurements.

Shared Encoder. To learn sim-to-real generalizable representations that generalize across the different visual styles of VR and live videos, we augment live images with VR videos for pre-training. Since we do not have live kinematics, the reconstruction loss from the kinematic decoder will be set to zero for live samples within a training batch. This creates a shared encoded representation space between the two visual domains such that visual concepts and semantics about manipulators and suturing needles can be shared between them. Moreover, as we simultaneously train the kinematic reconstruction task, we are learning a mapping that can generalize to live videos, since two similar positioning in either VR or live should have similar corresponding kinematics.

Expanded Training Set. Since we have limited surgical data, and the mapping from image to instrument kinematics is not specific to any one suturing skill, we can combine visual and kinematic data across different skills during pre-training. Specifically, we pre-train the model on all data combined across 6 suturing skills to help learn the mapping. In addition, we can further break down video clips and kinematic sequences into $F*(N_s+M_s)$ (image, kinematics) pairs to increase the effective training set size without needing heavy data augmentations. These two key facts provide is a unique advantage over traditional supervised learning, since training each skill assessment task required the full video clip to learn temporal signature along with skill-specific scorings.

Finetuning of LiveMAE for Skill Assessment After pre-training, we discard the image decoder and only use the pathway from the encoder to the kinematic decoder as our mapping ϕ. See Fig. 2b. We applied ϕ to our live data D_s^L and extract a surrogate kinematic sequence for each video clip. The extracted kinematics are embedded by a linear projection with added positional embeddings and processed with a lightweight sequential DistilBERT model. We append a linear layer on top of the pooled output from DistilBERT for classification. We finetune the last layer of ϕ and the sequential model with the cross-entropy loss using a small learning rate, e.g. 1e−5.

3 Experiments and Results

Datasets. We utilize a previously validated suturing assessment tool (EASE [8]) to evaluate the robotic suturing skill in both VR and live surgery. We collected 156 VR videos and 54 live surgical videos from 43 residents, fellows, and attending urologic surgeons in this 5-center multi-institutional study. VR suturing exercises were completed on the Surgical Science^{TM} Flex VR simulator and live surgical videos of surgeons performing the vesico-urethra anastomosis (VUA) step of a RARP were recorded. Each video was split into stitches, ($n = 3448$) total, and each stitch was segmented into sub-phrases with 6 binary assessment labels (low vs. high skill). See data breakdown and processing in Appendix 1.2.

Table 1. Suturing skill assessments on VR data. Boldfaced denotes best and ± are standard deviations across 5 held-out institutions.

		Repositions		HoldRatio		HoldAngle	
Modality	Model	*AUC*	*AUPRC*	*AUC*	*AUPRC*	*AUC*	*AUPRC*
Kinematics	LSTM	0.808 ± 0.02	0.888 ± 0.03	0.567 ± 0.09	0.859 ± 0.06	0.469 ± 0.06	0.804 ± 0.07
	Transformer	**0.852 ± 0.03**	**0.916 ± 0.04**	**0.652 ±0.07**	**0.895 ±0.05**	0.457 **±0.08**	0.796 ± 0.06
Video	ConvLTSM	0.715 ± 0.06	0.840 ± 0.04	0.587 ± 0.07	0.883 ± 0.05	0.552 ± 0.02	0.831 ± 0.05
	ConvTransformer	0.842 ± 0.02	0.912 ± 0.03	0.580 ± 0.03	0.880 ± 0.05	**0.560 ±0.06**	**0.837 ±0.06**
Video + Kin.	AuxTransformer	0.851 ± 0.02	0.912 ± 0.04	0.597 ± 0.07	0.886 ± 0.05	0.557 ± 0.03	0.842 ± 0.06
		DrivingSmoothness		WristRotation		WristRotationNW	
Modality	Model	*AUC*	*AUPRC*	*AUC*	*AUPRC*	*AUC*	*AUPRC*
Kinematics	LSTM	0.851 ± 0.06	0.953 ± 0.03	0.615 ± 0.07	0.894 ± 0.03	0.724 ± 0.04	0.942 ± 0.03
	Transformer	**0.878 ±0.06**	**0.963 ±0.02**	**0.640 ±0.08**	**0.899 ±0.02**	**0.725 ±0.07**	**0.942 ±0.03**
Video	ConvLTSM	0.851 ± 0.05	0.938 ± 0.03	0.636 ± 0.10	0.897 ± 0.03	0.661 ± 0.03	0.934 ± 0.02
	ConvTransformer	0.858 ± 0.04	0.956 ± 0.02	0.634 ± 0.10	0.895 ± 0.04	0.700 ± 0.06	0.937 ± 0.02
Video + Kin	AuxTransformer	0.868 ± 0.06	0.963 ± 0.02	0.649 ± 0.10	0.898 ± 0.04	0.675 ± 0.05	0.935 ± 0.02

Metrics and Baselines. Across the five institutions, we use 5-fold cross-valid ation to evaluate our model, training and validating on data from 4 institutions while testing on the 5th held-out institution. This allows us to test for generalization on unseen cases across both surgeons and medical centers. We measure and report the mean ± std. dev. for the two metrics: (1) Area-under-the-ROC curve (AUC) and (2) Area-under-the-PR curve (AUPRC) for the 5 test folds.

To understand the benefits of each data modality, we compare LiveMAE against 3 setups: (1) train/test using only kinematics, (2) train/test using only videos, and (3) train using kinematic and video data while testing only on video (no live kinematics). For kinematics-only baselines, we use two sequential models (1) LSTM recurrent model [9], and (2) DistillBERT transformer-based model [10]. For video-only baselines, we used two models based on pre-trained CNNs(3) ConvLSTM and (4) ConvTransformer. Both used pre-trained AlexNet to extract visual and flow features from the penultimate layer for each frame. The features are then flattened as used as input vectors to the sequential model (1) and (2). For a multi-modal baseline, we compare against recent work, AuxTransformer [7], which uses kinematics as privileged data in the form of an auxiliary loss during training. Unlike our method, they have additional kinematic supervision for the live video domain which we do not have.

3.1 Understanding the Benefits of Kinematics

Table 1 presents automated suturing assessment results for each technical skill on VR data from unseen surgeons across the 5 held-out testing institutions. We make 3 key observations: (1) we successfully reproduced assessment performance seen in previous works and showed that sequential models trained on kinematic-only data often achieve the best results (outperforming video and multi-modal on 5/6 skills with high mean AUCs (0.652–0.878) and AUPRC (0.895–0.963). (2) Vision model trained on video-only data can help with skill assessment, especially in certain skills such as *needle hold angle* where the angle between the needle tip and the target tissue (largely responsible for high/low score) is better represented visually, opposed kinematic poses. (3) Lastly, we demonstrated the benefits of using kinematics data as supervisory signals during training, which yields improved performance on video-only baselines where kinematic data are not available during testing, seen with AuxTranformer's numbers. Overall, kinematics provide a wealth of clean motion signals that is essential for skill assessment, which helps to inspire LiveMAE for assessment in live videos.

Table 2. Suturing skill assessment on Live data. Boldfaced denotes best and ± are standard deviations across 5 held-out institutions.

		Repositions		HoldRatio		HoldAngle	
Modality	Model	*AUC*	*AUPRC*	*AUC*	*AUPRC*	*AUC*	*AUPRC*
Video	ConvTransformer	0.822 ± 0.02	0.905 ± 0.02	**0.564 ±0.10**	**0.697 ±0.15**	0.489 ± 0.06	0.813 ± 0.03
Video + Kin.	AuxTransformer	0.831 ± 0.02	0.900 ± 0.01	0.466 ± 0.06	0.631 ± 0.19	0.505 ± 0.02	0.805 ± 0.06
	AuxTransformer-FT	0.828 ± 0.02	0.897 ± 0.01	0.472 ± 0.06	0.630 ± 0.19	0.499 ± 0.05	0.790 ± 0.07
	(Ours) LiveMAE	0.832 ± 0.03	0.911 ± 0.03	0.430 ± 0.05	0.5930 ± 0.20	**0.550 ±0.10**	**0.844 ±0.07**
	(Ours) LiveMAE-FT	**0.837 ±0.01**	**0.912 ±0.02**	0.474 ± 0.03	0.610 ± 0.20	0.489 ± 0.04	0.822 ± 0.04
		DrivingSmoothness		WristRotation		WristRotationNW	
Modality	Model	*AUC*	*AUPRC*	*AUC*	*AUPRC*	*AUC*	*AUPRC*
Video	ConvTransformer	0.667 ± 0.07	0.894 ± 0.06	0.435 ± 0.06	0.649 ± 0.05	0.445 ± 0.01	0.702 ± 0.05
Video + Kin.	AuxTransformer	0.502 ± 0.03	0.830 ± 0.07	0.519 ± 0.04	0.708 ± 0.04	0.519 ± 0.04	0.708 ± 0.04
	AuxTransformer-FT	0.483 ± 0.04	0.810 ± 0.10	0.517 ± 0.04	0.707 ± 0.04	0.520 ± 0.08	0.753 ± 0.06
	(Ours) LiveMAE	0.683 ± 0.08	0.878 ± 0.08	0.543 ± 0.13	0.721 ± 0.10	0.486 ± 0.12	0.723 ± 0.12
	(Ours) LiveMAE-FT	**0.725 ±0.12**	**0.903 ±0.06**	**0.562 ±0.08**	**0.733 ±0.08**	**0.634 ±0.06**	**0.826 ±0.01**

3.2 Evaluation of LiveMAE on Live Videos

Quantitative Results. Table 2 presents automated suturing assessment results for each technical skill on Live data across the 5 held-out institutions. We make 3 key observations: (1) Skill assessment on live stitch using LiveMAE or LiveMAE-finetuned often achieves the best results (outperforming supervised baselines and AuxTransformer with mean AUCs (0.550–0.837) and AUPRC (0.733–0.912) with particular improvements of 35.78% in AUC for wrist rotation skills and 8.7% for needle driving skill. (2) LiveMAE can learn generalizable representations from VR to Live using its shared encoder and kinematic mapping, achieving reasonable performance even without fine-tuning in the *needle repositioning, hold angle*

and *wrist rotation* skills. (3) Clinically, we observe that VR data can directly help with live skill assessment, especially in certain skills such as *wrist rotation* and *wrist rotation withdrawal* (+35.78% increase in AUC), where medical students confirmed that the rotation motions (largely responsible for high/low score) are more pronounced in VR suturing videos and less so in Live videos due to how manipulators are visualized in the simulation. Hence training with VR data can help to teach LiveMAE of the desired assessment procedure that is not as clear in Live data and supervised training paradigm. Overall, LiveMAE contributes positively to the task of automated skill assessment, especially in live scenarios where it is not possible to obtain kinematics from the da Vinci® surgical system.

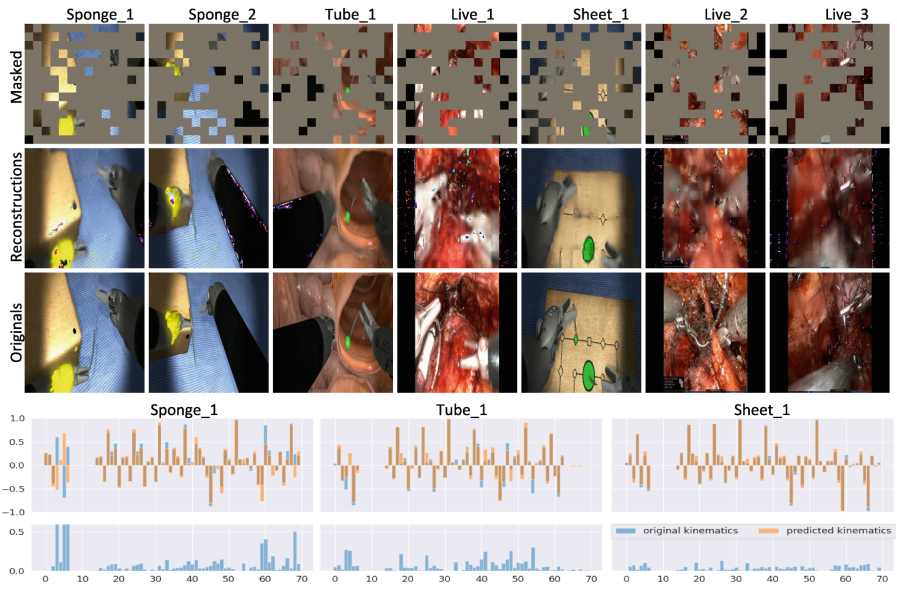

Fig. 3. Visualizations of reconstructed images and kinematics. Images for different exercises {sponge, tube, sheet} and live videos are presented. (a) Masked inputs and reconstructed images vs original images for held-out VR and live samples. (b) Predicted and original 70 kinematic features for the 4 VR samples. The bottom row plots the absolute difference. MSE for held-out VR kinematics are 0.045 ± 0.001.

Qualitative results. We visualized both reconstructed images and kinematics from masked inputs in Fig. 3. Top row of Fig. 3 a shows the 75% masked image where only 1/4 of the visible patches are input into the model. The block patterns are input patches to LiveMAE that were not masked. The middle row shows the image's visual reconstruction vs. the original images (last row). We observe that LiveMAE can pick out and reconstruct the positioning of the manipulators quite well. It also does a good job at reconstructing the target tissue, especially in *Tube1* and *Sheet1*. However, we also observe very small reconstruction artifacts

in darker/black regions. This can be attributed to the training data, which sometimes contain all black borders that were not cropped out, yielding the confusion between black borders in live videos and black manipulators in the VR videos. In Fig. 3b, we plot in the top row the original and predicted kinematics of the VR samples in blue and orange, respectively. The bottom row plots their absolute difference. LiveMAE does well in predicting kinematics from unseen samples, especially in *Sheet1* where it gets both positioning and orientations correctly for all instruments of interest, off by at most 0.2. In *Sponge1* and *Tube1*, we notice it does a poor job at estimating poses for some of the instruments, namely the orientation of the needle and target positions (index 4–7, 60–70) in *Sponge1* and the needle orientation (index 4–7) in *Tube1*. This can happen in cases where it is hard to see and recognize the needle in the scene, making it difficult to estimate the exact needle orientation, which may explain LiveMAE's poorer performance for the skill *Needle hold ratio.* and presents a promising direction for future work in diving deeper into CV models to segment out instruments of interest since they can be easily ignored.

4 Conclusion

Self-supervised learning methods, as utilized in our work, showed that video-based evaluation of suturing technical skills in live surgical videos is achievable with robust performance across multiple institutions. Although current work is limited to using VR data from one setup, namely Surgical Science™ Flex VR, our approach is independent from that system and can be applied on top of other surgical simulation systems with synchronized kinematics and video recordings. Future work will expand on the applications we demonstrated to determine whether it is possible to have a fully autonomous process, or semi-autonomously with a "human-in-the-loop".

Acknowledgements. This study is supported in part by the National Cancer Institute under Award Number 1RO1CA251579-01A1.

References

1. Birkmeyer, J.D., et al.: Surgical skill and complication rates after bariatric surgery. N. Engl. J. Med. **369**(15), 1434–1442 (2013). https://doi.org/10.1056/NEJMsa130062
2. Hung, A.J., et al.: A deep-learning model using automated performance metrics and clinical features to predict urinary continence recovery after robot-assisted radical prostatectomy. BJU Int. **124**(3), 487–495 (2019). https://doi.org/10.1111/bju.14735
3. Trinh, L., et al.: Survival analysis using surgeon skill metrics and patient factors to predict urinary continence recovery after robot-assisted radical prostatectomy. Eur. Urol. Focus. **S2405–4569**(21), 00107–00113 (2021). https://doi.org/10.1016/j.euf.2021.04.001

4. Chen, J., et al.: Objective assessment of robotic surgical technical skill: a systematic review. J. Urol. **201**(3), 461–469 (2019). https://doi.org/10.1016/j.juro.2018.06.078
5. Lendvay, T.S., White, L., Kowalewski, T.: Crowdsourcing to assess surgical skill. JAMA Surg. **150**(11), 1086–1087 (2015). https://doi.org/10.1001/jamasurg.2015.2405
6. Hung, A.J., et al.: Road to automating robotic suturing skills assessment: battling mislabeling of the ground truth. Surgery **S0039–6060**(21), 00784–00794 (2021). https://doi.org/10.1016/j.surg.2021.08.014
7. Hung, A.J., Bao, R., Sunmola, I.O., Huang, D.A., Nguyen, J.H., Anandkumar, A.: Capturing fine-grained details for video-based automation of suturing skills assessment. Int. J. Comput. Assist. Radiol. Surg. **18**(3), 545–552 (2023). Epub 2022 Oct 25. PMID: 36282465; PMCID: PMC9975072. https://doi.org/10.1007/s11548-022-02778-x
8. Sanford, D.I., et al.: Technical skill impacts the success of sequential robotic suturing substeps. J. Endourol. **36**(2), 273–278 (2022). PMID: 34779231; PMCID: PMC8861914. https://doi.org/10.1089/end.2021.0417
9. Graves, A.: Long short-term memory. In: Supervised Sequence Labelling with Recurrent Neural Networks. Studies in Computational Intelligence, vol. 385, pp. 37–45. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-24797-2_4
10. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
11. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
12. Dosovitskiy A, et al.:. An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Balvardi, S., et al.: The association between video-based assessment of intraoperative technical performance and patient outcomes: a systematic review. Surg. Endosc. **36**(11), 7938–7948 (2022). Epub 2022 May 12. PMID: 35556166. https://doi.org/10.1007/s00464-022-09296-6
14. Fecso, A.B., Szasz, P., Kerezov, G., Grantcharov, T.P.: The effect of technical performance on patient outcomes in surgery: a systematic review. Ann Surg. **265**(3), 492–501 (2017). PMID: 27537534. https://doi.org/10.1097/SLA.0000000000001959