SurgMAE: Masked Autoencoders for Long Surgical Video Analysis

Muhammad Abdullah Jamal and Omid Mohareri
Intuitive Surgical Inc.

Abstract

There has been a growing interest in using deep learning models for processing long surgical videos, in order to automatically detect specific clinical/operational activities and extract metrics that can enable workflow efficiency tools and applications. However, training such models require vast amounts of labeled data which is costly and not scalable. Recently, self-supervised learning has been explored in computer vision community to reduce the burden of the annotation cost. Masked autoencoders (MAE) got the attention in self-supervised paradigm for Vision Transformers (ViTs) by predicting the randomly masked regions given the visible patches of an image or a video clip, and have shown superior performance on benchmark datasets. However, the application of MAE in surgical data remains unexplored. In this paper, we first investigate whether MAE can learn transferrable representations in surgical video domain. We propose SurgMAE, which is a novel architecture with an intelligent masking strategy based on sampling tokens corresponding to high information spatio-temporal regions unlike random and tube masking for MAE. We provide an empirical study of SurgMAE on two large scale long surgical video datasets, and find that our method outperforms several baselines in low data regime. We conduct extensive ablation studies to show the efficacy of our approach and also demonstrate it's superior performance on UCF-101 to prove it's generalizability in non-surgical datasets as well.

1 Introduction

Robotic-assisted surgery (RAS) has been widely adopted for many surgical procedures since it allows surgeons to perform operations with more precision and provides benefits such as fast post-operative recoveries, less blood loss and shorter hospitalization [1]. However, the adaptation of RAS is still not ubiquitous due to barriers such as cost, training, and Operating room (OR) workflow complexities [2]. Many component technologies have been recently introduced to address such issues. Methods such as automatic activity recognition in OR [3, 4], scene understanding and context awareness in the OR [5], and endoscopic video workflow recognition [6] have shown the potential of enabling digital tools that can analyze and improve workflow processes for the surgeon and OR staff. The focus of this paper is automatic surgical activity recognition (SAR) which is the task of detecting activities or phases temporally in long videos. [3, 7] introduced a new dataset called OR-AR consisting of long videos collected from different ORs and analyzed performance of state of the art video action detection models on it. However, these approaches are fully supervised and require clinical data that is manually annotated by medical experts which can impede the scalability of these models. Indeed, we want to have models that can understand the surgical workflow in a scalable fashion either at surgery level or OR level. This requires us to have machine learning models that are data-efficient in nature.

In this paper, we leverage the unlabeled long surgical videos and propose a new self-supervised learning (SSL) approach based on masked autoencoders. SSL learns generic representations on

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

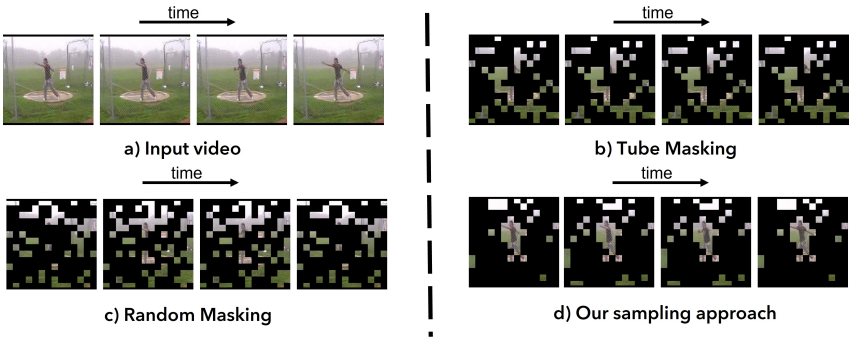

Figure 1: Comparison of different masking strategies. Random and Tube masking sample more uninformative regions from the input video clip while our proposed sampling strategy samples high spatio-temporal region and mostly discard background regions.

unlabeled data that are transferable to various downstream tasks. There are two different approaches used in SSL. In contrastive learning, the model takes different augmentations or views of an image and then uses contrastive loss to pull them together in the embedding space while pushing away the embeddings of different images. On the other hand, masked autoencoders (MAE) [8, 9, 10] take an input (either image or video), patchify it and then pass into patch embedding to generate a set of tokens. A high percentage of tokens are generally dropped and the remaining tokens are passed to the Vision Transformers (ViT) [11]. Then, the tokens embeddings concatenated with learnable masked tokens are passed to the decoders to reconstruct the masked patches. MAEs have recently gained a lot of momentum in SSL paradigm because of less inductive bias and high masking ratio and have consistently been outperforming contrastive learning based approaches. To the best of our knowledge, little or no work has been done in the surgical domain using MAE. We first investigate whether MAE can learn useful representations during pre-training for surgical video datasets. We train MAEs using different masking strategies such as random [8], tube [10] and frame [10] masking. We empirically found that the random masking strategy works best. However, we found that all the tokens are not informative as there are a lot of redundant information in the video and random masking strategy can select tokens from uninformative regions. To tackle this challenge, we propose a new masking strategy that samples tokens from high information spatio-temporal regions. Our masking strategy selects tokens based on the distance in the embedding space while discarding tokens from low information regions (background, redundant frames etc.) as shown in figure 1. We empirically show the efficacy of our new masking strategy on multiple datasets to prove out that it selects tokens from high informative regions, resulting in learning more useful representations for downstream tasks.

2 Related Work

Annotating massive amounts of surgical video data temporally and spatially requires manual work from medical experts and is impractical and expensive. To put our work in the context of this application and prior works, we briefly review the surgical video understanding, OR workflow analysis, and self-supervised learning paradigms. Please see the appendix for extended literature review.

Surgical Video Understanding. Datasets such as Cholec80 [12] and Cataract-101 [13] have allowed us to make advancement in surgical phase recognition mainly in laparoscopic and ophthalmological videos. Recent approaches [3, 14] are mainly supervised and consist of two stages.

Self-supervised Learning (SSL). Recently, contrastive learning [15, 16, 17, 18] approaches and masked autoencoder [10, 8, 9, 19, 20] are mainly used in SSL to learn better visual representations from large scale unlabeled datasets.

3 SurgMAE

Our goal is to learn video representations under masked autoencoder paradigm using high spatio-temporal tokens. Given a video clip $\mathbf{V}$ of size T x 3 x H x W where T is the number of frames in the clip, H and W are height and width of the frame and 3 corresponds to the number of channels in

the frame, we first pass it to a 3D convolutional layer with a patch size of 2x16x16 to extract $N = \frac{T}{2}\text{x}\frac{H}{16}\text{x}\frac{W}{16}$ tokens of dimension $\boldsymbol{d}$. Let $\mathbf{X}_i$ be the token embedding of two adjacent frames $(j, j+1)$ of the video clip. We use the euclidean norm between the token embedding $\mathbf{X}_{i+1,k}$ and the token embedding of the previous frame $\mathbf{X}_{i,k}$ over the same 2D position to compute if the token $\mathbf{X}_{i+1,k}$ belongs to high information region or not. Then, we only sample tokens with high distance value d based on the masking ratio as shown in equation 1. The intuition is if the spatial location of object changes in the next frame, then it should be accounted as high information token.

$$\begin{aligned} \mathbf{X}_i &= \texttt{Conv3d}(\mathbf{V}(j, j+1); \theta), \\ d_{i+1,k} &= l_2(\mathbf{X}_{i+1,k}, \mathbf{X}_{i,k}) \end{aligned} \tag{1}$$

Next, we follow MAE [8] to adapt separable positional embedding, one for space, and other for the time. For non-surgical data, we follow [10] for adding positional embeddings into the tokens. After positional embeddings are added, the sampled tokens $\mathbf{X}_v$ are passed to the encoder to extract latent representations $\mathbf{Z}_v$. The latent embeddings are then concatenated with learnable masked tokens $\mathbf{z}_m$. Finally, the positional embeddings are added, and passed to the decoder to reconstruct the masked patches $\hat{\mathbf{V}}$. Following [10, 8], we use mean squared error (MSE) as a loss function between the prediction and the normalized RGB pixel values:

$$\mathcal{L} = \frac{1}{\omega} \sum_{p \in \omega} ||\mathbf{V}(p) - \hat{\mathbf{V}}(p)||_2 \tag{2}$$

where p is the token index, ω is the set of masked tokens. Please see appendix for more ablation study on the effect of different loss functions. We remark that our masking strategy would allow the model to sample tokens based on the spatio-temporal information from each video clip as compared to random masking. As shown in the figure 2, the video frames has different spatio-temporal information and our sampling approach is able to capture the most important cues from them which allow the model to learn useful representations during pre-training and we have empirically shown the superior performance of the downstream tasks in the section 4.1. We present the pseudo-code of our sampler and SurgMAE in the appendix.

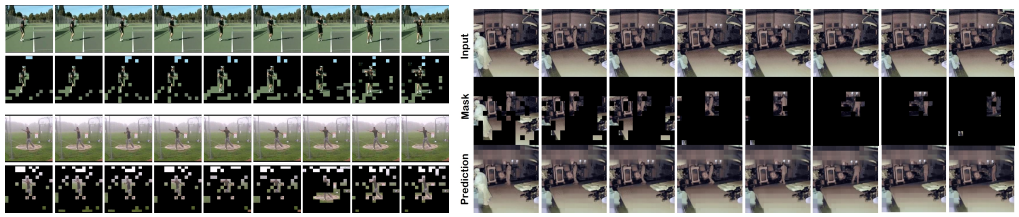

Figure 2: (**Left**) SurgMAE samples high spatio-temporal regions from the input video clip. (**Right**) We show the original input video clip (top), masks (middle) and the reconstruction (bottom row) on OR-AR.

4 Experiments

Datasets. We evaluate SurgMAE on three datasets: OR-AR [7] and OR-ARv2, Cataract-101 [13], and UCF-101 [21]. OR-AR is a surgical dataset consisting of 820 full videos captured using IR cameras placed in two ORs in a single hospital. OR-ARv2 is an extended version of OR-AR dataset that has videos collected from three new hospitals. This dataset includes 1302 OR videos and has been collected under institutional research board (IRB) approvals. OR-AR [7] and OR-ARv2 consist of 9 temporal workflow phases and contain 61 type of surgeries such as Lobectomy, Colectomy, Umbilical Hernia Repair etc. We split the dataset into 80-20 train-test split. Cataract-101 [13] is a small dataset consisting of 101 surgical videos and 10 surgical phases. UCF-101 [21] is also a relatively small dataset which consists of 9.5k videos in training and 3.5k in testing. Please see the appendix for implementation details.

4.1 Main Results

We compare SurgMAE with the recent masked autoencoder and contrastive learning approaches for video domain. Please see the appendix for more details on the baselines, ablation studies and the

Table 1: **Comparison of SurgMAE with the other state-of the art methods under different data-regime setting on the OR-AR dataset [7]**. We pre-train ViT-B using SurgMAE for 1600 epochs with high masking ratio of 90%.

Methods	Masking	Backbone	Pre-train	5%	10%	20%	100%
MaskFeat [9]	Random	MViT-S	Kinetics-400	62.35	78.88	-	-
MAE [8]	Random	ViT-B	Kinetics-400	64.66	81.48	84.93	94.76
MAE [8]	Random	ViT-B	OR-ARv2	66.58	81.87	84.97	**96.30**
MAE [8]	Frame	ViT-B	OR-ARv2	63.44	78.89	81.45	-
VideoMAE [10]	Tube	ViT-B	OR-ARv2	65.57	81.74	83.89	94.87
SurgMAE	high spatio-temporal sampling	ViT-B	OR-ARv2	**68.91**	**82.14**	**86.29**	95.60
Swin-B+BiGRU [7]	-	Swin-B	Kinetics-400	-	-	-	95.13

results on the Cataract-101 and OR-ARv2. Table 1 shows the comparison of various approaches under different data regimes for OR-AR [7]. We can observe that SurgMAE has a clear advantage over other masking approaches under low-data regime setting. More specifically, SurgMAE achieves 68.91 mAP when fine-tuned using 5% labeled data which shows that it is a more data-efficient learner than the recent sota approaches. Among other approaches, we observe that MAE [8] with random masking performs reasonably well under low-data regime setting and even performs slightly better than SurgMAE when fine-tune using full labeled dataset (96.30% vs 95.60% mAP). Moreover, both random masking and SurgMAE performs better than the fully supervised results obtained using pre-trained Swin transformer-Base [22] (Swin-B) as a backbone which shows the effectiveness of masked autoencoders when pre-training on in-domain large scale dataset.

UCF-101. To empirically test the generalizability of SurgMAE on non-surgical dataset, we run experiments on UCF-101. We report the top-1 accuracy in Table 2. It can be clearly seen that SurgMAE outperforms VideoMAE [10] achieving 92.1% top-1 accuracy compared to 91.2% which shows the efficacy of visible tokens from high information regions for masked autoencoder as shown in the figure 2. For UCF-101, we also sample tokens from the static regions on the top of high spatio-temporal tokens and didn't mask all the remaining tokens for the prediction task. We use a high masking ratio of 80% as compared to 75% ratio used in VideoMAE [10] making it less memory and computational intensive.

Table 2: **Comparison of SurgMAE with the other state-of the art methods on UCF-101 [21]**.

Methods	Masking	Backbone	Pre-train	Frames	Top-1
Scratch	-	ViT-B	UCF-101	16	51.4
OPN [23]	-	VGG	UCF-101	N/A	59.6
VCOP [24]	-	R(2+1)D	UCF-101	N/A	72.4
CoCLR [17]	-	S3D-G	UCF-101	32	81.4
Vi^{2}CLR [25]	-	S3D	UCF-101	32	82.8
CoCLR [17]	-	S3D-G	Kinetics-400	32	87.9
Vi^{2}CLR [25]	-	S3D	Kinetics-400	32	89.1
MoCov3 [16]	-	ViT-B	UCF-101	16	81.7
VideoMAE [10]	Tube	ViT-B	UCF-101	16	91.2
SurgMAE	high spatio-temporal sampling	ViT-B	UCF-101	16	**92.1**

5 Conclusion

In this paper, we investigate masked autoencoder based pre-training techniques for long surgical videos to learn better video representations. We propose SurgMAE, an adapted version of MAE for surgical videos, with a simple and effective token sampling strategy which samples tokens from high information spatio-temporal regions to alleviate the issues with random masking methods commonly used in current MAE approaches. We empirically show that our approach outperforms other masking strategies with ViT-B model on two surgical and one non-surgical (UCF-101) datasets to prove it's superior representation learning capability.

References

[1] Kyle H Sheetz, Jake Claflin, and Justin B Dimick. Trends in the adoption of robotic surgery for common surgical procedures. *JAMA Network Open*, 2020.

[2] Kenneth Catchpole, Colby E Perkins, Catherine Bresee, M. Jonathon Solnik, Benjamin Sherman, John L Fritch, Bruno Gross, Samantha Jagannathan, Niv Hakami-Majd, Raymund M. Avenido, and Jennifer T. Anger. Safety, efficiency and learning curves in robotic surgery: a human factors analysis. *Surgical Endoscopy*, 2015.

[3] Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. Automatic operating room surgical activity recognition for robot-assisted surgery. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, 2020.

[4] Adam Schmidt, Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. Multi-view surgical video action detection via mixed global view attention. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 2021.

[5] Zhaoshuo Li, Amirreza Shaban, Jean-Gabriel Simard, Dinesh Rabindran, Simon Peter DiMaio, and Omid Mohareri. A robotic 3d perception system for operating room environment awareness. *CoRR*, abs/2003.09487, 2020.

[6] Yanyi Zhang, Ivan Marsic, and Randall S Burd. Real-time medical phase recognition using long-term video understanding and progress gate method. *Medical Image Analysis*, 74:102224, 2021.

[7] Zhuohong He, Ali Mottaghi, Aidean Sharghi, Muhammad Abdullah Jamal, and Omid Mohareri. An empirical study on activity recognition in long surgical videos. In *Proceedings of the 2nd Machine Learning for Health symposium*, Proceedings of Machine Learning Research. PMLR, 2022.

[8] Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked autoencoders as spatiotemporal learners. *arXiv:2205.09113*, 2022.

[9] Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *CVPR*, 2022.

[10] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pages 10078–10093. Curran Associates, Inc., 2022.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[12] Andru Putra Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *CoRR*, abs/1602.03012, 2016.

[13] Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. Cataract-101: video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018*, 2018.

[14] Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23nd International Conference, Shenzhen, China, October 4-8, 2020, Proceedings, Part III*. Springer, 2020.

[15] Rui Qian, Tianjian Meng, Boqing Gong, Ming-Hsuan Yang, Huisheng Wang, Serge J. Belongie, and Yin Cui. Spatiotemporal contrastive video representation learning. *CoRR*, abs/2008.03800, 2020.

[16] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, October 2021.

[17] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. In *Neurips*, 2020.

[18] Tian Pan, Yibing Song, Tianyu Yang, Wenhao Jiang, and Wei Liu. Videomoco: Contrastive video representation learning with temporally adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11205–11214, 2021.

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021.

[20] Peng Gao, Teli Ma, Hongsheng Li, Ziyi Lin, Jifeng Dai, and Yu Qiao. MCMAE: Masked convolution meets masked autoencoders. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[21] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[22] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *CoRR*, abs/2106.13230, 2021.

[23] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequence. In *IEEE International Conference on Computer Vision*, 2017.

[24] Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Ali Diba, Vivek Sharma, Reza Safdari, Dariush Lotfi, Saquib Sarfraz, Rainer Stiefelhagen, and Luc Van Gool. Vi2clr: Video and image for visual contrastive learning of representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1502–1512, October 2021.

[26] Colin Lea, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks: A unified approach to action segmentation. In Gang Hua and Hervé Jégou, editors, *Computer Vision – ECCV 2016 Workshops*.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[28] Abdolrahim Kadkhodamohammadi, Afshin Gangi, Michel de Mathelin, and Nicolas Padoy. A multi-view RGB-D approach for human pose estimation in operating rooms. *CoRR*, abs/1701.07372, 2017.

[29] Vinkle Srivastav, Thibaut Issenhuth, Abdolrahim Kadkhodamohammadi, Michel de Mathelin, Afshin Gangi, and Nicolas Padoy. MVOR: A multi-view RGB-D operating room dataset for 2d and 3d human pose estimation. *CoRR*, abs/1808.08180, 2018.

[30] Muhammad Abdullah Jamal and Omid Mohareri. Multi-modal unsupervised pre-training for surgical operating room workflow analysis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, 2022.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 2019.

[32] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[33] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2020.

[34] Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M. Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders, 2022.

[35] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.

[36] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

[37] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.

Appendices

A Related Work

Annotating large amounts of surgical video data temporally and spatially requires manual work from medical experts and is impractical and expensive. To put our work in the context of this application and prior works, we briefly review the surgical video understanding, OR workflow analysis, and self-supervised learning paradigms.

Surgical Video Understanding. [14] proposed to use temporal convolution neural network (TCN) [26] on top of frame-wise features extracted from ResNet-18 [27]. Surgical activity recognition is not just limited to endoscopic videos, but it has been studied for operating room (OR) workflow analysis. The first large scale dataset (OR-AR) was first introduced in [3] which also proposed a supervised model consisting of I3D and Bi-GRU as backbone and temporal models respectively. The dataset was later extended in [7]. [4] leverages the multiple views of OR-AR dataset, and proposed a new attention module to smartly fuse those views. There are other data-driven based approaches [28] for OR workflow analysis which use multi-view RGBD dataset [29] for clinician detection and human pose estimation. However, there is little or no work done in data-efficient surgical activity recognition. Recently [30] proposed an unsupervised approach based on clustering which fuses multi-modal data collected from the OR.

Self-supervised Learning. Self-supervised learning based approaches for learning good video representations have been studied in the literature. Recently, contrastive learning based approaches [15, 16, 17, 18] has been proposed to learn better visual representation. These approaches generally require larger batch sizes, extra memory component and data augmentations.

Masked visual modeling. Masked visual modeling leverages the idea from masked language modeling used in bidirectional encoder (BERT) [31] and Generative Pre-Training (GPT) [32]. iGPT [33] follows GPT to process the pixel in sequential manner which shows that the masked pixel prediction can be performed. Recently, Vision Transformers (ViT) [11] have been designed to convert the patches into tokens to learn visual representations. Following the success of ViTs, several masked autoencoder based self-supervised approaches [8, 9, 19, 20] have been proposed. Masked Image Modeling [19] is a big success and an alternate approach to contrastive learning to learn useful image representations. Similarly, several masked autoencoder based approaches [10, 8, 34] have been designed for video domain to learn spatio-temporal representations. MAE [8] uses asymmetric encoder-decoder ViT with random masking during pre-training. VideoMAE [10] proposes tube masking strategy while AdaMAE [34] proposes a sampling network to sample tokens from high spatio-temporal regions which is trained end-to-end with ViT encoder using reinforcement learning. MaskFeat [9] instead of predicting masked patches, predicts the features of the masked tokens. These approaches have shown to use high masking ratio (75% to 95%) during pretraining as opposed to 60% masking ratio in image domain.

B Architecture

Table 3 refers to asymmetric encoder-decoder architecture. Each token is represented by a 768 embedding dimension. Next we sample (1 - r) x 1568 tokens as the high spatio-temporal sampling tokens and then pass them into ViT encoder which consists of 12 multi-head self-attention blocks (MHA). Then, the output of ViT encoder is concatenated with masked token representations and passed through the projection (MLP) layer which brings down the dimension from 768 to 512. These representations are then passed through the ViT decoder which consists of 4 MHA with the dimension of 512. Finally, these are passed to the MLP layer to increase the dimension from 512 to 1536 which is essentially the total number of pixels in the input video clip. The output is reshaped to the original shape for the computation of the reconstruction loss.

C Pseudocode for SurgMAE

We present the pseudo-code of our sampler and SurgMAE in the Algorithm 1 and 2 respectively.

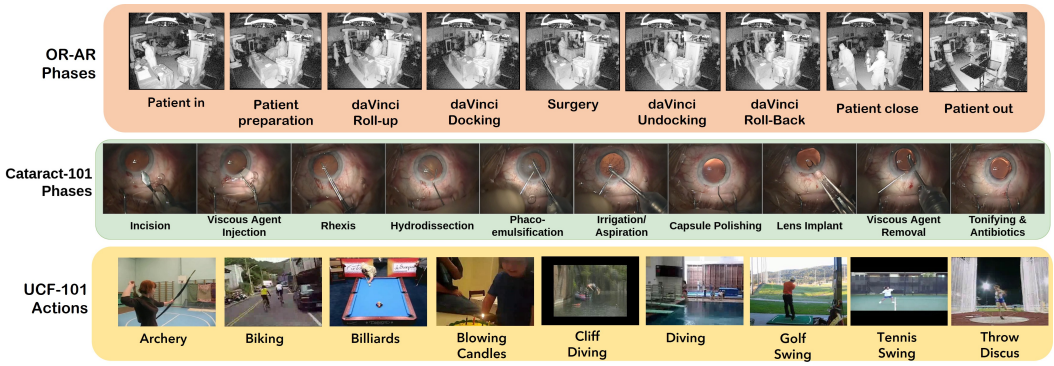

Figure 3: Some of the examples from the datasets with their activity or phase labels.

Algorithm 1 Pseudo-code for `SpatioTemporalSampler`.

Inputs: Tokenized Video $\mathbf{X} \in \mathbb{R}^{N \times d}$, Masking ratio: $r \in (0, 1)$
Outputs: Mask indices $\boldsymbol{I}_p$, Mask: $\boldsymbol{M}$
$d = \texttt{EuclideanNorm}(\mathbf{X})$
$N_{\text{v}} = \texttt{int}(N \times (1 - r))$
$\boldsymbol{I}_v = d.\texttt{sample}(N_{\text{vis}})$
$\boldsymbol{I}_p = \boldsymbol{U} - \boldsymbol{I}_v$
$\boldsymbol{M} = \texttt{GetMask}(\boldsymbol{I}_v, \boldsymbol{I}_p)$

Algorithm 2 Pseudo-code for our SurgMAE.

Input: Dataset $\mathcal{T}_r = \{\boldsymbol{V}_i : |i = 1, 2, 3, \cdots, |\mathcal{T}_r|\}$, Masking ratio: $r \in (0, 1)$
for $\boldsymbol{V}_i \in \mathcal{D}$ **do**
  $\boldsymbol{X}_i = \texttt{Tokenizer}(\boldsymbol{V}_i)$
  $\boldsymbol{X}_i = \boldsymbol{X}_i + \texttt{PosEmbed}$
  $\boldsymbol{M} = \texttt{SpatioTemporalSampler}(\boldsymbol{X}_i, r)$
  $\boldsymbol{X}_v = \boldsymbol{X}_i[\sim \boldsymbol{M}]$
  $\boldsymbol{Z}_v = \texttt{ViT-Base}(X_v)$
  $\boldsymbol{Z}_v = \boldsymbol{Z}_v + \texttt{PosEmbed}[\sim \boldsymbol{M}]$
  $\boldsymbol{z}_m = z_m + \texttt{PosEmbed}[\boldsymbol{M}]$
  $\boldsymbol{Z} = \boldsymbol{Z}_v \oplus \boldsymbol{z}_m$
  $\widehat{\boldsymbol{V}} = \texttt{Decoder}(\boldsymbol{Z})$
  $\widehat{\boldsymbol{V}}_m = \widehat{\boldsymbol{V}}[\boldsymbol{M}]$
  $\boldsymbol{V}_m = \boldsymbol{V}[\boldsymbol{M}]$
  $\mathcal{L} = \|\widehat{\boldsymbol{V}}_m - \boldsymbol{V}_m\|_2$
  $\mathcal{L}.backward()$
end for

Table 3: Architecture used in SurgMAE. MHA denotes Multi-head self-attention and MLP denotes Multilayer perceptron. We used joint space-time attention following [10].

Stage	ViT-B	Output shape
Input	stride 4 x 1 x 1	3 x 16 x 224 x 224
Tokenization	stride 2 x 16 x 16, emb dim 768	1568 x 768
Masking	High spatio-temporal sampling masking ratio r	[(1 - r) x 1568] x 768
Encoder	[MHA(768)] x 12	[(1 - r) x 1568] x 768
Projection	MLP(512) concat masked tokens	1568 x 512
Decoder	[MHA(512)] x 4	[(1 - r) x 1568] x 512
Projector	MLP(1536)	1568 x 1536
Reshape	from 1536 to 3 x 2 x 16 x 16	3 x 16 x 224 x 224

D Implementation Details

We use Vision transformer-Base (ViT-B) with joint-space time attention as the backbone following prior work [10]. We set the input number of frames to 16 and sampling rate of 4.0, and set the patch size of 2x16x16 which generates 1568 tokens for an input video clip with a size of 224x224. We follow [10, 8] for pre-training (see Table 4) and fine-tuning (see Table 5) settings. We conduct our experiments on 8 NVIDIA A100 GPUs.

Evaluation. In order to evaluate the pre-trained models for surgical datasets, we fine-tune ViT-B model on video clips similar to [3], and then extract features for full videos from ViT-B to train Bidirectional Gated Recurrent Unit (Bi-GRU) to detect surgical activities in surgical video datasets. For UCF-101, we only perform end-to-end fine-tuning of the backbone. We use mean average precision (mAP) and top-1 accuracy as evaluation metrics.

Table 4: Pre-training setting on OR-ARv2, UCF-101 and Cataract-101 datasets.

Configuration	OR-ARv2	UCF-101	Cataract-101
Optimizer	Adamw		
Optimizer betas	{0.9, 0.95}		
Base learning rate	1e-4	1e-3	1e-4
Weight decay	5e-2		
Learning rate schedule	cosine decay		
gradient clipping	0.02	None	0.02
Warmup epochs	40		
Epochs	1600	3200	800
Flip augmentation	True	True	False
Augmentation	MultiScaleCrop		
Num of Frames	16		
sampling rate	4.0		

Table 5: Fine-tune setting on OR-AR, UCF-101 and Cataract-101 datasets.

Configuration	OR-AR	UCF-101	Cataract-101
Optimizer	Adamw		
Optimizer betas	{0.9, 0.95}		
Base learning rate	6e-4	1e-3	6e-4
Weight decay	5e-2		
Learning rate schedule	cosine decay		
Warmup epochs	5		
Epochs	100		
Flip augmentation	True	True	False
Mixup	None	0.8	None
CutMix	None	1.0	None
drop path	0.1	0.2	0.1
drop out	0.0	0.5	0.0
Layer-wise lr decay	0.65	0.70	0.65
Temporal Model learning rate	1e-3	None	1e-3
Temporal Model Epochs	25	None	25

E Baselines

We directly compare SurgMAE to the recent masked autoencoders that includes MaskFeat [9], VideoMAE [10], MAE [8]. We also compare with SwAV [35] and SimCLR [36] for Cataract-101 experiments. Moreover, we add MoCov3 [16] as one of the baselines for UCF-101 experiments following VideoMAE [10].

F Ablation Study

We perform in-depth ablation studies of SurgMAE on OR-AR [7] dataset. We pre-trained ViT-B on the OR-ARv2 dataset and then fine-tune it under low-data regime setting (5% labeled data) for evaluation. We report these studies in Table 6.

Masking ratio. Table 6a shows the performance of SurgMAE on different masking ratios. It shows that SurgMAE performs well with a high masking ratio which makes the pre-training fast and less memory intensive. Surprisingly, masking ratio of 95% achieves 63.37% mAP which is in line with the fact that SurgMAE samples

Table 6: Ablation studies on **OR-AR [7]** under low-data (5% labeled data) regime setting. We use ViT-B as a backbone for all the experiments.

(a) **Different Masking Ratio**. SurgMAE works well with high masking ratio. Models are trained for 400 epochs

ratio	mAP
0.95	63.37
0.90	**64.97**
0.85	62.26
0.80	61.01

(b) **Decoder Depth**. SurgMAE performs the best with 4 blocks of decoder. Models are trained for 800 epochs with masking ratio of 90%.

blocks	mAP
1	61.35
2	65.84
4	**67.86**
8	62.75

(c) **Mask sampling**. SurgMAE outperforms random, frame and tube masking.

case	mAP
random	66.58
tube	65.57
frame	63.44
SurgMAE	**68.91**

(d) **Pre-training epochs**. Better performance achieves during fine-tuning when pre-trains for more epochs.

epochs	mAP
400	64.97
600	65.89
800	67.86
1600	**68.91**

(e) **Loss function**. SurgMAE performs best with MSE loss and normalization. Models are trained for 800 epochs with masking ratio of 90%.

case	mAP
MSE (w / norm)	**67.86**
MSE (wout / norm)	63.41
L1 (w / norm)	64.06
L1 (wout / norm)	62.14

more high information spatio-temporal tokens and requires fewer tokens during pre-training to achieve a reasonable performance. We observe a drop in fine-tuning performance when pre-training the model using lower masking ratio. We hypothesize that lower masking ratios sample more redundant patches which results in poor generalization.

Decoder design. Table 6b shows the performance of SurgMAE on different blocks of the decoder. We observe that the performance increases when increasing the depth of the decoder from 1 block to 4 blocks. However, with much deeper decoder, we see a performance degradation which is in accordance with the observation made in recent masked autoencoder approaches [8, 10].

Masking strategy. Table 6c compares the performance of SurgMAE with the recent masking strategies. We observe that the random masking outperforms tube masking with high masking ratio. However, frame masking which masks out future or past frames performs poorly compared to tube or random masking. Same observation has been made in [8]. **SurgMAE** which samples high spatio-temporal tokens yields the best performance with high masking ratio (68.91% with 90% ratio).

Pre-training epochs. Next, we show the impact of pre-training epochs on the fine-tuning performance in Table 6d. We observe an increase in the performance (64.97% mAP to 67.86% mAP) when number of epochs goes from 400 to 800. If we further pre-train the model with 1600 epochs, we achieve our best performance, but it comes with a cost of more pre-training time.

Reconstruction target. Table 6e compares the performance of SurgMAE using different loss functions. We observe that MSE loss performs better compared to L1 loss. We also observe that per-patch normalized pixels yields better results compared to using raw pixel value which is on par with observation made in recent masked autoencoder approaches [8, 10].

G Additional Results on Cataract-101 and OR-ARv2

Table 7 compares the performance of SurgMAE on OR-ARv2 on full labeled dataset. We see the same observation that SurgMAE yields the best performance (93.11% mAP) compared to other masking strategies which empirically verify the effectiveness of sampling high information spatio-temporal tokens during pre-training. Moreover, similar to what we observe in OR-AR results, MAE [8] with random masking performs better than the tube and frame masking.

We report the mAP of Bi-GRU for various approaches in Table 8. We carefully follow the training practices to pre-train ViT-B for SimCLR and SwAV to avoid collapse issue. We observe that Cataract-101, being a relatively small dataset, is more challenging to pre-train vision transformers which is on par with the observation found in VideoMAE [10]. Nevertheless, SurgMAE still outperforms other masking approaches which makes it more data-efficient approach for self-supervised pre-training. We also find out that by fine-tuning from a ViT-B

Table 7: **Comparison of SurgMAE with the other state-of the art methods on the OR-ARv2 dataset**.

Methods	Masking	Backbone	Pre-train	mAP
MAE [8]	Random	ViT-B	Kinetics-400	92.06
MAE [8]	Random	ViT-B	OR-ARv2	92.70
MAE [8]	Frame	ViT-B	OR-ARv2	91.87
VideoMAE [10]	Tube	ViT-B	OR-ARv2	92.36
SurgMAE	high spatio-temporal sampling	ViT-B	OR-ARv2	**93.11**

pre-trained on large-scale dataset (kinetics-400 [37]) yields the best performance which is generally a standard practice for such small datasets.

Table 8: **Comparison of SurgMAE with the other state-of the art methods on Cataract-101 [13]**. We pre-train ViT-B using SurgMAE for 800 epochs with high masking ratio of 90%. For evaluation, we report mean average precision of Bi-GRU during fine-tuning stage.

Methods	Masking	Backbone	Pre-train	mAP
SwAV [35]	-	ViT-B	Cataracts	83.61
SimCLR [36]	-	ViT-B	Cataracts	83.20
MAE [8]	Random	ViT-B	Cataracts	86.43
VideoMAE [10]	Tube	ViT-B	Cataracts	85.05
SurgMAE	high spatio-temporal sampling	ViT-B	Cataracts	**87.78**
MAE [8]	Random	ViT-B	Kinetics-400	**92.85**

H Mask Visualization

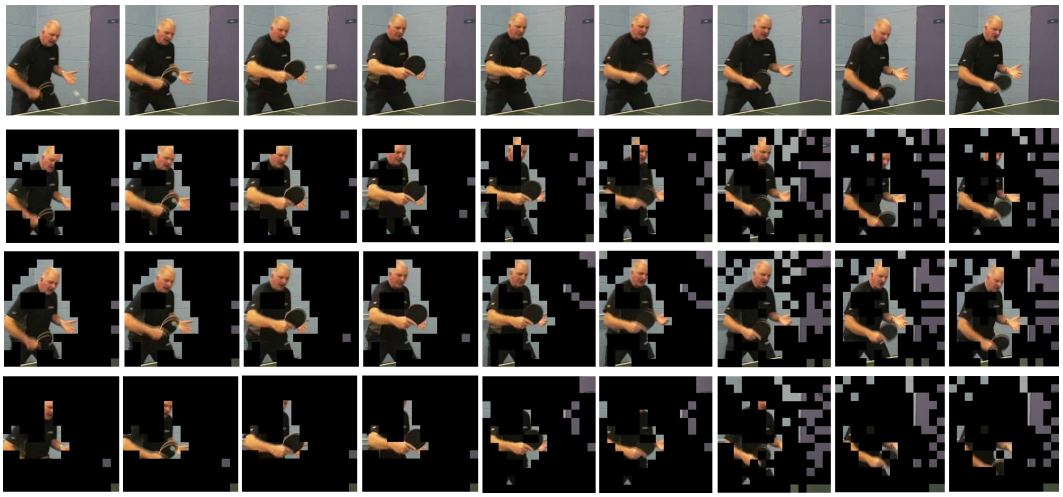

Figure 4: SurgMAE samples high spatio-temporal regions from the input video clip. The first row shows the frames of the video clip while the remaining rows show the sampling of important regions using our approach by varying the masking ratio.

I Reconstruction Visualization

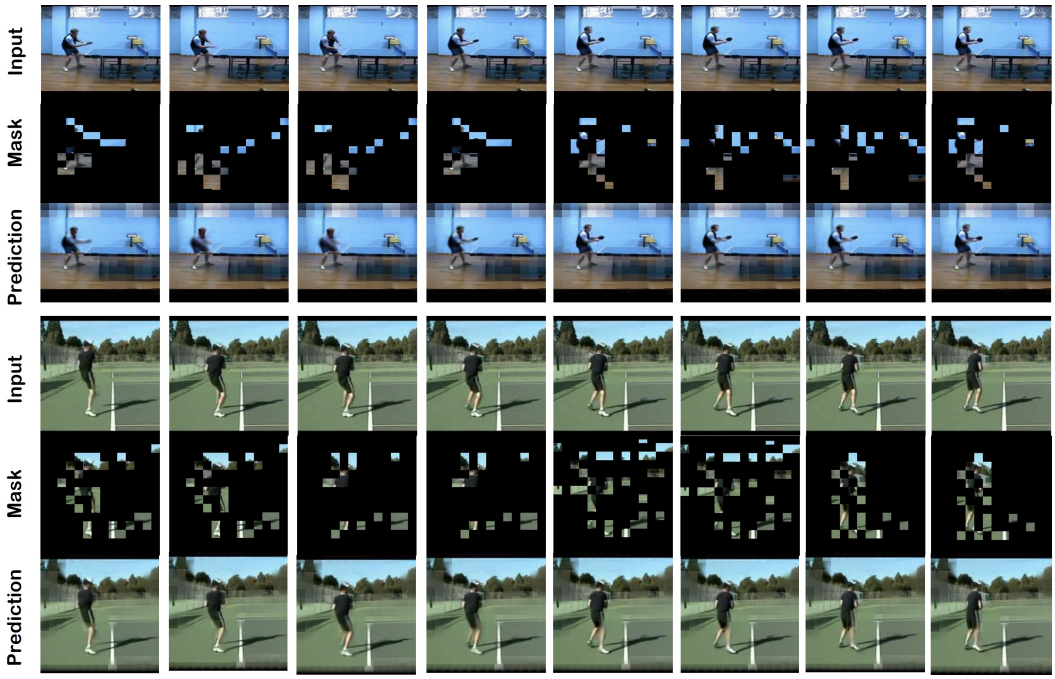

Figure 5: More reconstruction results using SurgMAE on UCF-101.