Rethinking RGB-D Fusion for Semantic Segmentation in Surgical Datasets

Muhammad Abdullah Jamal[1] ABDULLAH.JAMAL@INTUSURG.COM
Omid Mohareri[1] OMID.MOHARERI@INTUSURG.COM
[1] *Intuitive Surgical, Sunnyvale CA, United States*

Abstract

Surgical scene understanding is a key technical component for enabling intelligent and context aware systems that can transform various aspects of surgical interventions. In this work, we focus on the semantic segmentation task, propose a simple yet effective multi-modal (RGB and depth) training framework called SurgDepth, and show state-of-the-art (SOTA) results on all publicly available datasets applicable for this task. Unlike previous approaches, which either fine-tune SOTA segmentation models trained on natural images, or encode RGB or RGB-D information using RGB only pretrained backbones, SurgDepth, which is built on top of Vision Transformers (ViTs), is designed to encode both RGB and depth information through a simple fusion mechanism. We conduct extensive experiments on benchmark datasets including EndoVis2022, AutoLapro, LapI2I and EndoVis2017 to verify the efficacy of SurgDepth. Specifically, SurgDepth achieves a new SOTA IoU of **0.86** on EndoVis 2022 SAR-RARP50 challenge and outperforms the current best method by at least **4%**, using a shallow and compute efficient decoder consisting of ConvNeXt blocks.

Keywords: Multi-Modal Learning; RGB-D Fusion; Surgical Instrument Segmentation; Semantic Segmentation

1. Introduction

Intelligent and context aware surgical systems and digital tools have a significant potential to transform minimally invasive procedures by enhancing surgeon and care-team performance, and improving overall safety. Surgical scene parsing is a key component for designing such systems through enabling tasks such as pose estimation Du et al. (2018), tool tracking Sznitman et al. (2014) and phase recognition pha (2019). Applications such as operating room workflow optimization Jin et al. (2021), surgeon skill assessment Reiley and Hager (2009); Zia et al. (2018) and automation of surgical sub-tasks Huang et al. (2023) can be built on top of such technologies.

In this paper, we focus on single frame semantic segmentation of instrument, anatomy and other objects present in surgical scenes. The objective is to assign each pixel a correct semantic label. Most of the earlier work Shvets et al. (2018); Zhao et al. (2020); Jin et al. (2019) follow segmentation models built for non-surgical images such as MaskRCNN He et al. (2018) and UNet Ronneberger et al. (2015), either by directly fine-tuning them or incorporating additional cues such as pose Kurmann et al. (2017), saliency maps Islam et al. (2019), optical flows Jin et al. (2019) and motion flows Zhao et al. (2020). However, unique challenges such as occlusion, variability in lighting, presence of smoke and blood, and diverse instrument and tissue types limit accuracy, generalizability and clinical translation of present methods. Incorporation of 3D geometric information is a promising approach to help enhance the performance of such segmentation algorithms. RGB-D datasets are commonly being used in non-surgical applications such as autonomous driving Huang et al. (2022), robotics Marchal et al. (2020) and SLAM Wang et al. (2023). Existing approaches Wang et al. (2022a); Zhang et al. (2023) have shown state-of-the-art performance on non-surgical benchmark datasets Nathan Silberman and Fergus (2012); Song et al. (2015) through methods that leverage both RGB and depth data. However, to the best of our knowledge, very little or no work has been done on effective utilization of RGB-D data for surgical instrument and tissue segmentation.

This motivates us to present SurgDepth, a simple yet effective RGB-D semantic segmentation framework for endoscopic surgical data. SurgDepth builds the interaction between both data modalities by fusing them using a 3D awareness block. The purpose of this block is to incorporate 3D geometric informa-

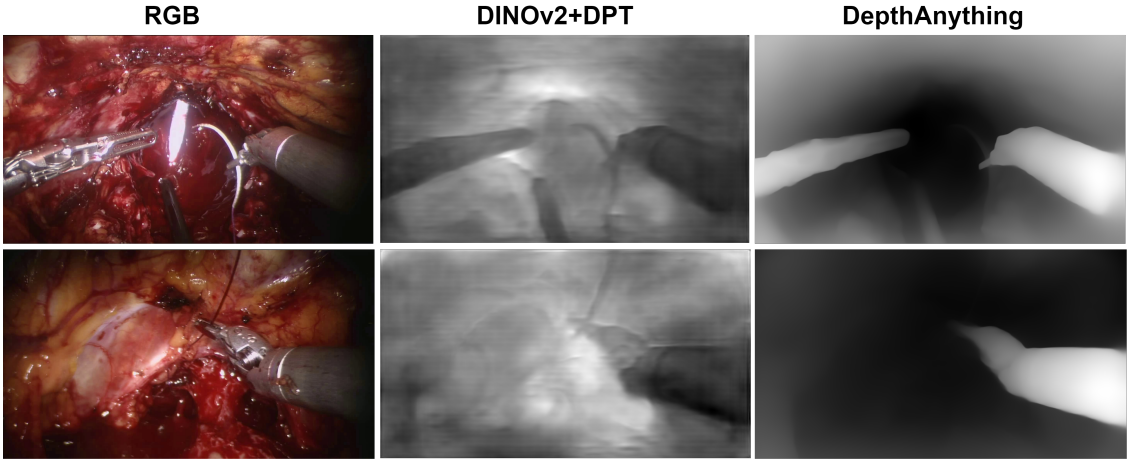

Figure 1: Predicted depth maps using DINOv2+DPT Oquab et al. (2023) and DepthAnything Yang et al. (2024) on SAR-RARP50 examples.

tion from the depth maps to enhance localization of objects and structures. Our fusion module can be plugged in to any Vision Transformer (ViT) Dosovitskiy et al. (2021). Moreover, we propose a shallow decoder based on ConvNeXt Liu et al. (2022) blocks to predict the segmentation map. Through extensive experiments, we found out that such method outperforms transformer based decoders such as Segmenter Strudel et al. (2021). We employ state-of-the-art depth estimation models like DINOv2 Oquab et al. (2023) and DepthAnything Yang et al. (2024) to predict depth maps from RGB only surgical videos. Figure 1 shows some of the predicted depth maps using these models on SAR-RARP50 dataset Psychogyios et al. (2024).

We demonstrate the efficacy of SurgDepth on multiple datasets such as SAR-RARP50 Psychogyios et al. (2024), AutoLapro Wang et al. (2022c), LapI2I Pfeiffer et al. (2019), CholecSeg8k Hong et al. (2020) and EndoVis2017 Allan et al. (2019). By adding a ConvNeXt based decoder, we outperform all competing methods while requiring less computational resources. More specifically, our approach achieves **0.862** IoU on SAR-RARP50 with 98.37M parameters, a new SOTA performance, as compared to Uninades, the best performing method reported in Psychogyios et al. (2024), which achieves 0.829 IoU with 107M parameters.

To summarize, our main contributions are:

1. We propose a new RGB-D training framework called SurgDepth for semantic segmentation in surgical scenes.
2. A new 3D awareness fusion block to encode 3D geometric information from depth maps, and a shallow decoder to produce segmentation maps.
3. Our SurgDepth shows new SOTA performance with less computation cost on five benchmark datasets.
4. We conduct extensive ablation studies to emphasize the importance of the 3D fusion block, demonstrating that it is model-agnostic and can be integrated with any transformer architecture.

2. Related Work

There are a few publicly available datasets Psychogyios et al. (2024); Allan et al. (2019); Wang et al. (2022c) for surgical instrument and tissue segmentation. With the rise of EndoVis challenge, various surgical scene understanding techniques have been explored. In particular, approaches for instrument segmentation can be grouped as semantic Jin et al. (2019); Shvets et al. (2018); Ni et al. (2020); Kamrul Hasan and Linte (2019) and instance segmentation González et al. (2020); Kong et al. (2021); He et al. (2018). Our work targets the semantic segmentation task.

Semantic Segmentation. TernausNet Shvets et al. (2018) used a UNet architecture Ronneberger et al. (2015) on the top of pre-trained VGG encoder for binary instrument segmentation. Kamrul Hasan and Linte (2019) proposed a UNet plus architecture, a

modified encoder-decoder UNet with data augmentation techniques for medical image segmentation. Ni et al. (2020) proposed a progressive alternating attention network (PAANet) which consists of progressive alternating attention dense (PAAD) blocks to construct attention guided map from all scales. MF-TAPNet Jin et al. (2019) incorporates temporal priors by leveraging motion flow to an attention pyramid network. In addition to the above approaches, Islam and et al. (2019); Wang et al. (2021) target real-time semantic segmentation.

Instance Segmentation. Unlike semantic segmentation, which assigns class labels to each pixel, instance segmentation, or mask classification, is an alternative paradigm that assigns class labels to each object instance or binary mask. Most of the earlier work primarily use Mask-RCNN He et al. (2018). ISINet González et al. (2020) builds on the top of Mask-RCNN and proposes a temporal consistency module by taking advantage of the sequential nature of the video data. Kong et al. (2021) re-defined Mask-RCNN by improving region proposal network with anchor optimization. Another line of work in this domain is to develop specialized models Kurmann et al. (2021); Islam et al. (2020). AP-MTL Islam et al. (2020) proposed an encoder-decoder multi-task learning architecture for real-time instance segmentation.

3. SurgDepth

SurgDepth follows a standard encoder-decoder architecture as illustrated in Figure 2. The goal of the encoder is to learn discriminative representations while the decoder is responsible for transforming these features into segmentation maps.

An RGB image and depth map with spatial size of H × W are first processed through modality-specific projection layers consisting of a single convolutional layer. Then, the RGB and depth features are passed to the fusion block which encodes 3D geometrical information. Next, to learn useful representations, the modality-specific features are concatenated and passed to the encoder, which is a ViT in our case. Finally, the features from the encoder are passed to a lightweight decoder to produce the segmentation map of size H × W.

3.1. 3D Awareness Fusion Block

Our fusion block consists of a 3D awareness attention module that incorporates 3D information from the depth maps to enhance localization of the semantic classes as shown in Figure 3. Given the RGB X_i^{rgb} and the depth X_i^{depth} features, we first concatenate the modality-specific features and then down-sample them through an adaptive pooling layer to reduce the computational complexity and generate query (Q) features. The key (K) and value (V) are extracted from the RGB features X_i^{rgb}. This can be formulated as:

$$Q = \text{FC}(\text{AdaptivePool}_{k\times k}(Concat(X_i^{rgb}, X_i^{depth}))),$$
$$K = \text{FC}(X_i^{rgb}), V = \text{FC}(X_i^{rgb}), \quad (1)$$

where AdaptivePool performs adaptive average pooling to downsample the spatial size to $k \times k$, and FC is a fully connected layer. Based on the Q, K, and V, we formulate the attention module as:

$$X_{fusion} = \text{Bilinear}(V \cdot \text{Softmax}(\frac{QK^\top}{\sqrt{C^d}})), \quad (2)$$

where Bilinear$(\cdot)$ is a bilinear upsampling operation (**F.interpolate**() in Pytorch) that converts the spatial size from $k \times k$ to $h \times w$ and C^d represents the dimension of Q, K and V. Finally, the features X_{fusion} are passed to two projection layers (FC) to produce updated RGB features $\hat{X}_i^{rgb}$ and depth features $\hat{X}_i^{depth}$.

3.2. Overall Architecture

RGB features $\hat{X}_i^{rgb}$ and depth features $\hat{X}_i^{depth}$ are concatenated and passed through the ViT encoder to encode RGB-D data. The features from the encoder's last layer are passed to the lightweight decoder to yield segmentation maps. We empirically found that passing only RGB features to the decoder yields higher performance as compared to passing both RGB and depth features (c.f. Table 6). Our lightweight decoder consists of ConvNeXt blocks Liu et al. (2022) and a convolutional layer for producing segmentation maps. Each ConvNext block has one depth-wise convolutional layer with a kernal size of 7 × 7 and two point wise convolutional layers. It follows the inverted bottleneck design where the dimension of the middle point-wise convolution is four times bigger than the input dimension. Please follow Liu et al. (2022) for more details on the block. We also experimented with Segmenter Strudel et al. (2021) as the decoder but

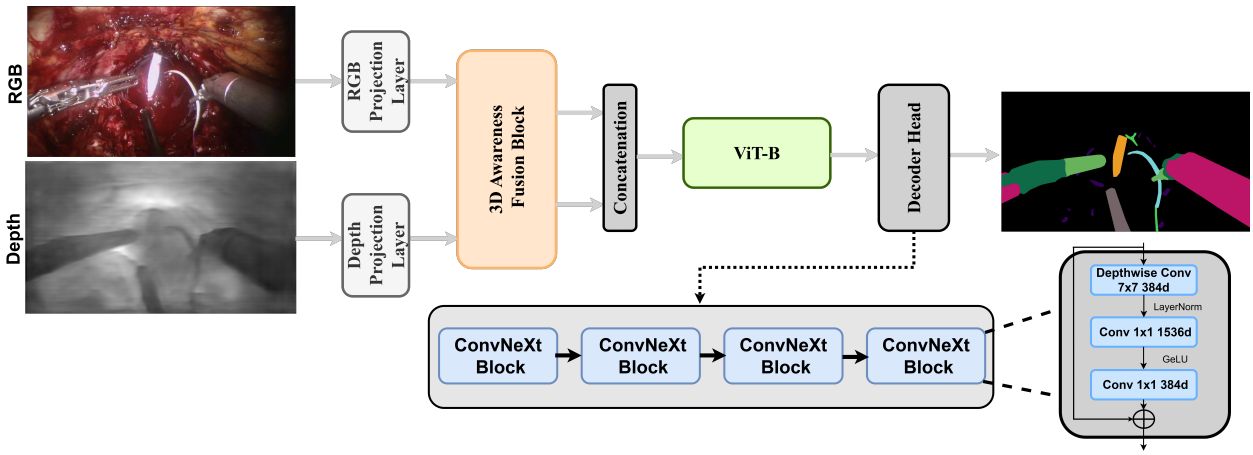

Figure 2: Overall architecture of SurgDepth. First, we encode the 3D geometric information using a 3D awareness fusion block and then encode the concatenated RGB-D in ViT-B. Then, the RGB features are passed to a shallow decoder head to predict the segmentation map.

empirically found that it doesn't outperform our ConvNeXt based decoder and brings about computational overhead.

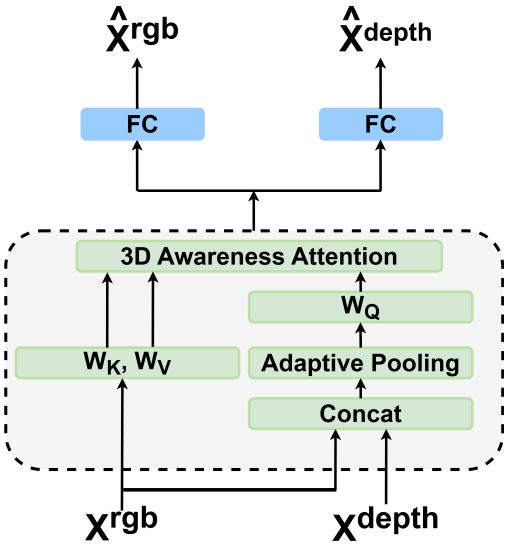

Figure 3: 3D awareness fusion block.

4. Experiments

We evaluate the performance of our architecture on multiple benchmark datasets.

4.1. Datasets and Evaluation.

SAR-RARP50 Challenge Psychogyios et al. (2024). This dataset contains 50 videos collected from Robot-Assisted Radical Prostatectomy procedures. It consists of 12998 training frames from 40 different videos and 3252 test frames from 10 other videos and nine semantic classes. Please follow Psychogyios et al. (2024) for more details on the dataset.

AutoLapro Wang et al. (2022c). This dataset has 21 laparoscopic hysterectomy videos recorded with 25 fps and a resolution of 1920×1080 pixels. For segmentation task, the dataset has 1800 frames annotated with nine semantic classes which constitutes 5936 annotations. We use the official train/validation/test splits provided in the dataset.

LapI2I Pfeiffer et al. (2019). The Laparoscopic Image-to-Image (I2I) Translation dataset contains 20,000 synthetic frames annotated with seven classes including liver, fat, diaphragm, tool shaft, tool tip and gallbladder. Synthetic frames are generated using 3D laparoscopic simulations created from CT scans of ten patients. The dataset doesn't provide any official split for train/val/test. We use data from 7 patients in the training set and the rest for testing.

CholecSeg8k Hong et al. (2020). This dataset consists of 8,080 laparoscopic cholecystectomy frames with a resolution of 854 × 480 from 17 video clips in Choloec80. Each image is annotated with 13 classes (abdominal wall,liver, gastrointestinal tract,

fat, grasper etc.) at the pixel-level. We use 13 videos in the training set and set aside 4 videos for testing.

EndoVis 2017 Allan et al. (2019). EndoVis 2017 challenge dataset consists of 10 video sequences of abdominal porcine procedures recorded by a daVinci Xi system. Each video contains 300 frames with a resolution of 1280×1024. The frames are annotated with six instrument classes and an ultrasound probe. For fair comparison, we follow Shvets et al. (2018) and use 4-fold cross-validation from 1800 frames (8 x 225). Each fold consists of 1350 and 450 frames for training and validation respectively.

Evaluation. We use mean intersection over union (IoU) as our evaluation metric Strudel et al. (2021); Cheng et al. (2022).

4.2. Implementation Details

SurgDepth is compatible with any Vision Transformer (ViT) as an encoder in our framework. In this work, we use ViT-B as an encoder, and four ConvNeXt blocks and one convolutional layer to produce segmentation maps. First, we increase the dimensionality D of the encoder's output with a linear layer and then reshape it to the size of $H/4$ x $W/4$ x $D/8$. We use Adaptive pool of 7x7 in 3D fusion block. The output dimension of encoder is 768 and projected layer dim is 6144. The output dimension of the decoder dim is 384 and finally, the final conv2d dim is number of classes + 1. Before calculating the loss on the segmentation map, we upsample the resolution of the map using bilinear interpolation. We initialize our ViT-B encoder with ImageNet Deng et al. (2009) unsupervised pre-trained weights He et al. (2021). During training, we use the input resolution of 480x640 across all the datasets. We train the model for 50 epochs with a learning rate of 1e-4 and adamw optimizer. Moreover, we apply random gaussian blurring, random horizontal flip and colorJitter which randomly change the brightness, contrast, saturation and hue of a training image. Lastly, we train our models using the batch size of 2 on 4 NVIDIA A100 GPUs.

4.3. Results on SAR-RARP50 challenge

In Table 1, we compare SurgDepth for SAR-RARP50 challenge with the methods that are reported in the Psychogyios et al. (2024). SurgDepth is able to surpass all competitors using the ConvNeXt based decoder. We also compare with the approaches proposed for natural images such as MAE He et al. (2021) which pre-trains ViTs using ImageNet-1k dataset under a masked autoencoding paradigm. As it can be seen in the table, SurgDepth achieves 0.862 IoU, a new state-of-the-art performance on SAR-RARP50 dataset. Compared to best performing method (Uninades) reported in the paper Psychogyios et al. (2024), our approach is computationally more efficient as it requires lesser image resolution and ConvNeXt blocks instead of Mask2Former as segmentation head. Unindaes model has 107M parameters while SurgDepth with ConvNeXt decoder has only 98.37M parameters. Since Uninades' code is not open source, so we only report the number of trainable parameters based on the architecture they used for this dataset. Although using off-the-shelf depth estimation model increases the total number of parameters and inference time, but we want to show impact of encoding the depth information. One can use depth sensors like RGB-D for example, to extract accurate depth maps and which will in-turn reduce the inference time and total number of parameters. Moreover, we ran the baseline Mask2Former with ViT-B and compared it with our method, both using Mask2Former as the segmentation head. ViT-B + Mask2Former achieves an IoU of 0.832, while SurgDepth with Mask2Former achieves an IoU of 0.857 using DINOv2-generated depth maps Oquab et al. (2023).

4.4. Results on AutoLapro

We compare SurgDepth for AutoLapro dataset with the method reported in Wang et al. (2022c). To the best of our knowledge, there hasn't been a lot of work reporting performance on AutoLapro so we mainly reported the results on the methods proposed for natural images. We can see from the Table 2 that SurgDepth with ConvNeXt decoder outperforms all the baselines. More specifically, it achieves 78.0 IoU setting a new SOTA on AutoLapro dataset.

4.5. Results on LapI2I and CholecSeg8k

LapI2I dataset consists of annotations for instruments as well as tissue. We directly compare our approach with recent segmentation models for non-surgical images in Table 3. SurgDepth outperforms Mask2Former and MaskRCNN by **4.6** IoU and **8.0** IoU indicating the importance of 3D information during the training. Compared with transformer model pretrained under masked autoencoding paradigm, our approach is superior to them, achieving a new state-of-the-art of **98.1** IoU. We can draw about the same observations as

Table 1: Comparison of SurgDepth with the other approaches on **SAR-RARP50 Semantic Segmentation**.

Methods	Backbone	Segmenter Head	Input size	Pre-train	IoU
Hi-Lab 2022	Swin-B	SegFormer Ensemble	512x512	N/A	0.817
Summer Lab - AI	Swin-L	UperNet	422x750	ImageNet-22K	0.816
AIA - Noobs	EfficientNetB4	UNet ++	480x640	ImageNet-1K	0.789
Uninades	Swin-B	Mask2Former	750x1333	COCO + EndoVis 17 + EndoVis 18	0.829
MAE He et al. (2021)	ViT-B	ConvNeXt	480x640	ImageNet-1K	0.835
MAE He et al. (2021)	ViT-B	ConvNeXt	480x640	ImageNet-1K + SAR-RARP50	0.809
SurgDepth w/ DINOv2	ViT-B	ConvNeXt	480x640	ImageNet-1K	**0.862**
SurgDepth w/ DINOv2	ViT-B	Segmenter	480x640	ImageNet-1K	0.854
SurgDepth w/ DepthAnything	ViT-B	ConvNeXt	480x640	ImageNet-1K	**0.858**
SurgDepth w/ DepthAnything	ViT-B	Segmenter	480x640	ImageNet-1K	0.851

Table 2: Quantitative comparison on **AutoLapro Semantic Segmentation**.

Methods	Architecture	Input size	Pre-train	IoU
MaskRCNN He et al. (2018)	ResNet-50	480x640	COCO	67.8
YOLACT	ResNet-50	480x640	ImageNet-1K	65.2
YolaactEdge	ResNet-50	480x640	ImageNet-1K	64.4
MAE He et al. (2021)	ViT-B + ConvNeXt	480x640	ImageNet-1K	76.8
SurgDepth w/ DINOv2	ViT-B + ConvNeXt	480x640	ImageNet-1K	**78.0**
SurgDepth w/ DINOv2	ViT-B + Segmenter	480x640	ImageNet-1K	76.3
SurgDepth w/ DepthAnything	ViT-B + ConvNeXt	480x640	ImageNet-1K	**77.2**
SurgDepth w/ DepthAnything	ViT-B + Segmenter	480x640	ImageNet-1K	76.1

above for the CholecSeg8K from Table 3 (last column).

4.6. Results on EndoVis 2017

Table 4 shows the comparison of SurgDepth with the competing methods on EndoVis 2017 dataset. We report both the challenge IoU and standard IoU. As our approach is a single frame based method, we mainly list the performance of baselines that require single frame in the input. We can see from the table that SurgDepth consistently outperforms all the approaches across both metrics. More specifically, it improves over ISINet by **15%** challenge IoU and **1.45%** IoU showing the importance of encoding 3D geometric information for semantic segmentation. We want to emphasize that ISINet is an instance-based method for instrument segmentation and we only report the result without the temporal consistency module which takes advantage of the sequential nature of the video data and the instrument motion. With temporal consistency module, ISINet achieves 55.62 challenge IoU which is still lower than SurgDepth but it achieves a higher IoU (52.2), showing that for this dataset, apart from 3D geometric information, other factors like motion, optical flow and temporal information of the instruments are also important.

4.7. Ablation Study

We perform ablation studies to evaluate the impact of each component in our approach.

Input features to the decoder. In Table 6, we show that using only RGB features as input to the decoder brings the best performance on SAR-RARP50 dataset as compared to both RGB and depth features while saving the computational cost.

Number of ConvNext blocks in decoder. Table 5 shows the performance of SurgDepth on SAR-RARP50 challenge by varying the number of ConvNeXt blocks in the decoder. We observe the performance boost when we increase the number of blocks from 1 to 4. However, we see a degradation in the performance when we use a much deeper decoder.

Effectiveness of 3D fusion block. In Table 7, We report the results on the SAR-RARP50 dataset after removing the 3D awareness block from the model. The findings show that simply concatenating or adding RGB and depth features results in poorer performance compared to our fusion block. This clearly demonstrates that the fusion block enhances 3D awareness and helps the model capture semantic objects more effectively.

Table 3: Results on **LapI2I** and **CholecSeg8K** Semantic Segmentation.

Methods	Architecture	Input size	Pre-train	LapI2I IoU	CholecSeg8K IoU
MaskRCNN He et al. (2018)	ResNet-50	480x640	COCO	90.1	47.3
Mask2Former Cheng et al. (2022)	ResNet-50	480x640	COCO	93.5	48.9
MAE He et al. (2021)	ViT-B + ConvNeXt	480x640	ImageNet-1K	97.4	54.1
SurgDepth w/ DINOv2	ViT-B + ConvNeXt	480x640	ImageNet-1K	**98.1**	**55.6**
SurgDepth w/ DINOv2	ViT-B + Segmenter	480x640	ImageNet-1K	96.8	54.4

Table 4: Comparison of SurgDepth with the other approaches on **EndoVis 2017**.

Methods	Architecture	Pre-train	Challenge IoU	IoU
MaskRCNN He et al. (2018)	ResNet-50	COCO	45.65	41.77
Mask2Former Cheng et al. (2022)	ResNet-50	COCO	40.39	39.84
MAE He et al. (2021)	ViT-B + ConvNeXt	ImageNet-1K	55.28	44.87
CascadeRCNN Cai and Vasconcelos (2017)	ResNet-50	COCO	49.03	39.9
UNetPlus Kamrul Hasan and Linte (2019)	UNet	None	36.14	13.14
PlainNet Jin et al. (2019)	UNet	None	36.45	13.28
TernausNet-11 Shvets et al. (2018)	UNet11	None	35.27	12.67
MF-TAPNET Jin et al. (2019)	UNet	None	37.35	13.49
ISINet González et al. (2020)*	ResNet-50	None	53.55	49.57
SurgDepth w/ DINOv2	ViT-B + ConvNeXt	ImageNet-1K	**61.93**	**50.29**
SurgDepth w/ DINOv2 (224 × 224)	ViT-B + ConvNeXt	ImageNet-1K	57.67	48.51

Table 5: SurgDepth performs the best with 4 ConvNeXt blocks in the decoder.

Blocks	**IoU**
1	0.843
2	0.851
4	**0.862**
8	0.856

Table 6: SurgDepth performs the best when only RGB features are passed to the decoder.

Deocder Input	**#Params**	**IoU**
RGB	98.37M	**0.862**
RGB+Depth	103.1M	0.853

Table 7: Ablation results without 3D awareness block.

Model	**IoU**
3D awareness block	**0.862**
Concatenation	0.850
Addition	0.847

Table 8: Different fusion manner after 3D awareness block.

Fusion manner	**IoU**
Concatenation	**0.862**
Addition	0.858
Hadamard	0.859

Fusion manner after 3D awareness block. In the model architecture, we concatenate the RGB and depth features after 3D fusion block. In Table 8, we report the performance of different fusion manner on the SAR-RARP50 dataset. We observe that by naively concatenating the features brings the best performance as compared to addition and hadamard product.

Variants of Vision Transformer. In the model architecture, we use ViT-B as the encoder. In Table 11, we show that our approach can also work with other variants of vision transformer such as ViT-L.

Effect of fusion depth features with K and V. In Table 10, we show the effect of fusing depth features with K and V in the 3D awareness block on the SAR-RARP50 dataset. We find that using only RGB features with K and V brings the best performance and fusing depth with K or V doesn't bring any improvement but can increase the computational cost.

Table 9: Comparison of SurgDepth with other methods on the EndoVis 2017 dataset for instance segmentation.

Method	Ch_IoU	ISI_IoU	BF	PF	LND	VS	GR	MCS	UP	mc_IoU
TernausNet-11 Shvets et al. (2018)	35.27	12.67	13.45	12.39	20.51	5.97	1.08	1.00	16.76	10.17
MF-TAPNet Jin et al. (2019)	37.35	13.49	16.39	14.11	19.01	8.11	0.31	4.09	13.40	10.77
ISINet González et al. (2020)	55.62	52.20	38.70	38.50	50.09	27.43	2.01	28.72	12.56	28.96
TraSeTR Zhao et al. (2022)	60.40	65.20	45.20	56.70	55.80	38.90	11.40	31.3	18.20	36.79
S3Net Baby et al. (2023)	72.54	71.99	**75.08**	54.32	61.84	35.5	**27.47**	43.23	**28.38**	46.55
MATIS Ayobi et al. (2023)	71.36	66.28	68.37	53.26	53.55	31.89	27.34	21.34	26.53	41.09
CRIS Wang et al. (2022b)	69.94	67.83	54.87	50.21	68.33	50.12	0.00	43.97	0.00	38.21
CLIPSeg Lüddecke and Ecker (2022)	70.15	65.02	51.29	42.27	49.56	30.12	9.96	30.69	20.05	33.42
TP-SIS (448) Zhou et al. (2023)	77.79	76.45	69.57	68.91	89.88	82.60	0.00	72.53	0.00	54.78
SurgDepth + TP-SIS (448)	78.34	77.07	70.12	69.37	90.29	83.44	0.00	73.18	0.00	55.20

Table 10: Effect of fusing depth features with Q, K and V on the SAR-RARP50 dataset.

Fusion	**IoU**
only Q	0.862
Q,K,V	0.862

Table 11: Performance of variants of ViT on SAR-RARP50 Semantic Segmentation.

Encoder	**IoU**
ViT-B	0.862
ViT-L	0.865

4.8. Results on Instrument Instance Segmentation

Although, the problem we're aiming to address in this work is semantic segmentation of the entire surgical scene and we want to reiterate that our focus is not surgical instrument instance segmentation. However, our 3D awareness fusion block is model-agnostic and can be incorporated with any transformer based approach. For instrument instance segmentation, we pick TP-SIS Zhou et al. (2023), and add 3D fusion block before passing the features to the CLIP Radford et al. (2021) image encoder. Table 9 shows the performance of different methods on EndoVis 2017. We can see that our 3D awareness fusion block further boosts the original TP-SIS approach Zhou et al. (2023). These results are interesting because it shows us that adding a new modality such as depth by encoding 3D geometric information with the text modality enhances the surgical instrument segmentation. This suggests that future research could benefit from exploring the integration of such modalities. Besides, compared with the other competing methods, our results establish a new state of the art for the EndoVis 2017 dataset.

5. Conclusion

We propose a novel RGB-D training framework called SurgDepth for semantic segmentation in surgical videos. SurgDepth consists of a novel 3D awareness attention block which builds interaction between RGB and depth by incorporating 3D geometric information from the depth maps. The method can be used with any type of Vision Transformer (ViT). Moreover, it can act as a building block for architectures that take video or stereo data as inputs. Our experiments demonstrate that SurgDepth achieves new state-of-the-art performance on five benchmark datasets with less computational cost, thanks to a shallow decoder consisting of ConvNeXt blocks.

6. Limitations

The depth estimation models we used are trained on natural images, which may not be fully accurate for generating depth maps in surgical images. Fine-tuning these models on surgical data could improve depth map extraction, or using RGB-D cameras or active sensors might provide more accurate depth information. Another limitation is the lack of temporal information in our approach, which we believe could further enhance the performance of surgical segmentation.

7. Broader Impact

Surgical segmentation tool can enhance precision and safety in procedures by accurately identifying anatomical structures and tracking instruments, reducing the

risk of errors and unintended tissue damage. This technology can also support personalized surgical planning, aids in surgeon training, and drives the development of smarter, AI-driven surgical tools that improve patient care.

References

Surgical workflow and skill analysis (2019). https://endovissub-workflowandskill.grand-challenge.org/, 2019.

Max Allan, Alex Shvets, and et al. 2017 robotic instrument segmentation challenge, 2019.

Nicolás Ayobi, Alejandra Pérez-Rondón, Santiago Rodríguez, and Pablo Arbeláez. Matis: Masked-attention transformers for surgical instrument segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 2023.

Britty Baby, Daksh Thapar, Mustafa Chasmai, Tamajit Banerjee, Kunal Dargan, Ashish Suri, Subhashis Banerjee, and Chetan Arora. From forks to forceps: A new framework for instance segmentation of surgical instruments, 2023.

Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection, 2017.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Alexey Dosovitskiy, Lucas Beyer, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Xiaofei Du, Thomas Kurmann, and et al. Articulated multi-instrument 2-d pose estimation using fully convolutional networks. *IEEE Transactions on Medical Imaging*, 2018.

Cristina González, Laura Bravo-Sánchez, and Pablo Arbelaez. Isinet: An instance-based approach for surgical instrument segmentation. 2020.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn, 2018.

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

W. Y. Hong, C. L. Kao, and et al. Cholecseg8k: A semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80, 2020.

Keli Huang, Botian Shi, Xiang Li, Xin Li, Siyuan Huang, and Yikang Li. Multi-modal sensor fusion for auto driving perception: A survey, 2022.

Tao Huang, Kai Chen, and et al. Demonstration-guided reinforcement learning with efficient exploration for task automation of surgical robot, 2023.

Mobarakol Islam and et al. Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning. *IEEE Robotics and Automation Letters*, 2019.

Mobarakol Islam, Yueyuan Li, and Hongliang Ren. Learning where to look while tracking instruments in robot-assisted surgery, 2019.

Mobarakol Islam, V. S. Vibashan, and Hongliang Ren. Ap-mtl: Attention pruned multi-task learning model for real-time instrument detection and segmentation in robot-assisted surgery. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

Yueming Jin, Keyun Cheng, Qi Dou, and Pheng-Ann Heng. Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video, 2019.

Yueming Jin, Yonghao Long, Cheng Chen, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. Temporal memory relation network for workflow recognition from surgical video, 2021.

S. M. Kamrul Hasan and Cristian A. Linte. U-netplus: A modified encoder-decoder u-net architecture for semantic and instance segmentation of surgical instruments from laparoscopic images. 2019.

Xiaowen Kong, Yueming Jin, and et al. Accurate instance segmentation of surgical instruments in robotic surgery: model refinement and cross-dataset evaluation. *International journal of computer assisted radiology and surgery*, 2021.

Thomas Kurmann, Marquez Neila, and et al. *Simultaneous Recognition and Pose Estimation of Instruments in Minimally Invasive Surgery.* 2017.

Thomas Kurmann, Pablo Márquez-Neila, and et al. Mask then classify: multi-instance segmentation for surgical instruments. *International Journal of Computer Assisted Radiology and Surgery*, 2021.

Zhuang Liu, Hanzi Mao, and et al. A convnet for the 2020s., 2022.

Timo Lüddecke and Alexander S. Ecker. Image segmentation using text and image prompts, 2022.

Nicolas Marchal, Charlotte Moraldo, Hermann Blum, Roland Siegwart, Cesar Cadena, and Abel Gawel. Learning densities in feature space for reliable segmentation of indoor scenes. *IEEE Robotics and Automation Letters*, 2020.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Zhen-Liang Ni, Gui-Bin Bian, and et al. Pyramid attention aggregation network for semantic segmentation of surgical instruments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, and et al. Dinov2: Learning robust visual features without supervision, 2023.

Micha Pfeiffer, Isabel Funke, and et al. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation, 2019.

Dimitrios Psychogyios, Emanuele Colleoni, and et al. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Carol E. Reiley and Gregory D. Hager. Task versus subtask surgical skill evaluation of robotic minimally invasive surgery. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, 2009.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

Alexey A. Shvets, Alexander Rakhlin, Alexandr A. Kalinin, and Vladimir I. Iglovikov. Automatic instrument segmentation in robot-assisted surgery using deep learning. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015.

Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation, 2021.

Raphael Sznitman, Carlos Becker, and Pascal Fua. Fast part-based classification for instrument detection in minimally invasive surgery. In *MICCAI*, 2014.

Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam, 2023.

Jiacheng Wang, Yueming Jin, and et al. *Efficient Global-Local Memory for Real-Time Instrument Segmentation of Robotic Surgical Video*. 2021.

Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers, 2022a.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation, 2022b.

Ziyi Wang, Bo Lu, and et al. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *MICCAI*, 2022c.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation, 2023.

Zixu Zhao, Yueming Jin, Xiaojie Gao, Qi Dou, and Pheng-Ann Heng. Learning motion flows for semi-supervised instrument segmentation from robotic surgical video, 2020.

Zixu Zhao, Yueming Jin, and Pheng-Ann Heng. Trasetr: Track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery, 2022.

Zijian Zhou, Oluwatosin Alabi, Meng Wei, Tom Vercauteren, and Miaojing Shi. Text promptable surgical instrument segmentation with vision-language models. *Advances in Neural Information Processing Systems*, 2023.

Aneeq Zia, Andrew Hung, Irfan Essa, and Anthony Jarc. Surgical activity recognition in robot-assisted radical prostatectomy using deep learning, 2018.