AP-MTL: Attention Pruned Multi-task Learning Model for Real-time Instrument Detection and Segmentation in Robot-assisted Surgery

Mobarakol Islam, Vibashan VS and Hongliang Ren

Abstract—Surgical scene understanding and multi-tasking learning are crucial for image-guided robotic surgery. Training a real-time robotic system for the detection and segmentation of high-resolution images provides a challenging problem with the limited computational resource. The perception drawn can be applied in effective real-time feedback, surgical skill assessment, and human-robot collaborative surgeries to enhance surgical outcomes. For this purpose, we develop a novel end-to-end trainable real-time Multi-Task Learning (MTL) model with weight-shared encoder and task-aware detection and segmentation decoders. Optimization of multiple tasks at the same convergence point is vital and presents a complex problem. Thus, we propose an asynchronous task-aware optimization (ATO) technique to calculate task-oriented gradients and train the decoders independently. Moreover, MTL models are always computationally expensive, which hinder real-time applications. To address this challenge, we introduce a global attention dynamic pruning (GADP) by removing less significant and sparse parameters. We further design a skip squeeze and excitation (SE) module, which suppresses weak features, excites significant features and performs dynamic spatial and channel-wise feature re-calibration. Validating on the robotic instrument segmentation dataset of MICCAI endoscopic vision challenge, our model significantly outperforms state-of-the-art segmentation and detection models, including best-performed models in the challenge.

I. INTRODUCTION

Image-guided robotic surgery is the latest form of development in minimally invasive surgical technology, which increases precision, reliability, and repeatability. A Vision-enabled surgical system like Da Vinci [1] is eminently assisting the surgeon in performing complex surgical manipulations with minimal incisions and recovery time than conventional surgeries. Despite development in surgical technology, sometimes surgeons may lose surgical workflow due to weak tactile feedback or system impairment. Moreover, a compound, surgical scenario with smoke, body fluid, blood, adverse lighting condition, and partial occlusion creates additional challenges in image cognition [2]. Therefore, concurrent detection and segmentation of instruments could enhance surgical outcomes and assist novice surgeons with real-time objective feedback, do skill assessment, and analyze tool movements in the surgical workflow.

In recent years, deep neural network is showing success in the tasks of segmentation [3], detection [4] and visual perception [5] from computer vision to medical imaging. However, most of these models are designed and optimized for the specific task. These restrain the impact of the deep learning model full exploitation, whereas the human brain is capable of doing multiple tasks like classification, tracking, and reasoning in parallel. Therefore, developing a multi-task learning (MTL) model in surgical scene understanding with instrument detection and segmentation can play an important rule to advance the robotic intervention.

A. Related Work

1) Surgical Instrument Segmentation: Recently, CNN models are deploying to achieve state-of-the-art performance in surgical tool tracking. Instrument segmentation of binary, parts and type-wise is done by LinkNet [8] with Jaccard index based loss function [9]. Subsequently, ToolNet [10] with holistically-nested architecture, and dilated residual network with multi-scaled inputs [2] are utilized to segment the surgical tool. Nonetheless, these works lack focusing on the real-time application in terms of both speed and accuracy. A real-time instrument segmentation approach of using auxiliary supervised adversarial learning is proposed in [11]. The work demonstrates high-speed inference with satisfying accuracy in binary instrument segmentation. However, it performs poorly for the type-wise segmentation. In our previous work [6], a novel design of segmentation decoder with spatial and channel squeeze & excitation (scSE) [12] is improved the instrument type segmentation by exciting the significant features.

2) Surgical Instrument Detection: There are few works with CNN on surgical instrument tracking by detection. R-CNN with region proposal networks (RPN) [13] is exploited to detection the instrument in robot-assisted surgery [14]. Subsequently, a residual CNN [15], and RANSAC based CNN [16] are used to detect bounding box with the tooltip. A 3D fully connected network (FCN) similar to UNet [3] is developed to detect articulation joint of the surgical instrument [17]. Nonetheless, more works are needed for real-time application using high-resolution images.

3) Multi-task learning (MTL): Multi-task learning models in surgical instrument tracking is still an open challenge. There are very few models that are trying to tackle the multiple tasks within a single network. A concurrent instrument segmentation and pose estimation multi-task learning (MTL) method [18] is designed with ResNet-50 [19] as encoder, FCN decoder and a regression block of fully-connected layer. However, the model is not suitable for the real-time application, and parts segmentation performance is not satisfactory. In computer vision, several MTL models such as MaskRCNN [20], Blitznet [21], and Sistu et al. [22]

Authors are with the Medical Mechatronics Lab, Dept. of Biomedical Engineering, National University of Singapore, Singapore-http://bioeng.nus.edu.sg/mm, Corresponding author, Hongliang Ren, ren@nus.edu.sg.
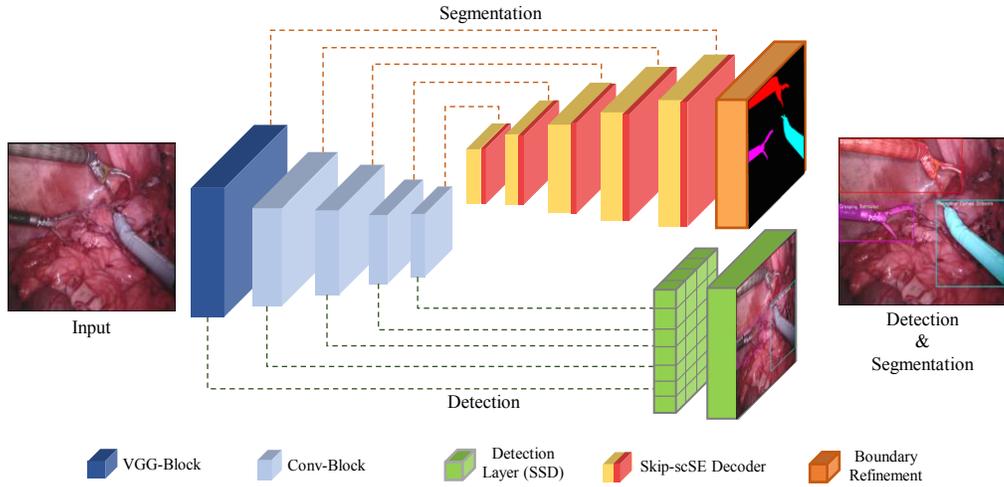
Fig. 1. Network architecture of the proposed AP-MTL model for real-time detection and segmentation. The architecture consists of VGG16 encoder, segmentation decoder like [6], and detection branch like SSD [7].

are developed for joint semantic segmentation and object detection tasks. More recently, UberNet [23] and Islam et al. [6] are tried to optimize MTL model with the multi-phase training approach and tuning. Nonetheless, lower inference speed and optimizing both tasks in a convergence point are still the open challenges in the MTL model.

B. Contributions

In this work, we address the problems of real-time MTL model for robotic instrument detection and segmentation with asynchronous task-aware optimization and attention pruning. Our contributions are summarized as follows:

- We propose a real-time MTL model with a light weight-shared encoder and task-aware spatial decoders for detection and segmentation.
- We introduce a novel way to train proposed MTL model by using asynchronous task-aware optimization (ATO) technique.
- To reduce model computation and singularity, we design an attention-based pruning method known as global attention dynamic pruning for trained models with dynamic learning.
- We develop an innovative design of segmentation decoder with fusing bypass connection in scSE to boost up the model performance.
- We annotate the bounding box for the instruments types with robotic instrument segmentation challenge [24] dataset. Our model achieves impressive results in detection and surpasses the existing state-of-the-art segmentation models and participant methods of the challenge.

II. METHODS

In this work, we construct a real-time multi-task learning model with light-weight encoder of VGG16 [25] and Skip-scSE decoder, SSD decoder [7] for segmentation and detection tasks respectively. To optimize our MTL model

at the same convergence point, we introduce asynchronous task-aware optimization technique (ATO). We also designed a novel skip spatial-channel squeeze and excitation module (Skip-scSE), by integrating skip connection to enhance segmentation prediction. Moreover, to boost real-time computation, reduce singularity, and redundancy in the model, we propose global attention dynamic pruning (GADP) method.

A. Skip-scSE

Spatial and channel squeeze & excitation (scSE) [12], [26] is a parameter learning technique which learns by recalibrates the feature maps to suppress the weak features and signify the meaningful features. However, in deeper layers, it may cause sparsity and singularity due to squeezing of irrelevant features towards zero [27]. In [28], skip connection is used to eliminates the singularity and sparsity in the CNN learning. To resolve this issue, we also integrate skip connection which retains the weak features and enhance the useful features by exciting them. We propose skip-scSE unit, by adding a bypass connection with scSE as in Fig. 2. In skip-scSE, the output of channel excitation $\tilde{x}_{cE}$ and spatial excitation $\tilde{x}_{sE}$ are fused with skip input x_r. Therefore, the output excited feature maps x_{SC} can be formulated as given in equation 1,

$$x_S = \tilde{x}_{cE} + \tilde{x}_{sE} + x_r \qquad (1)$$
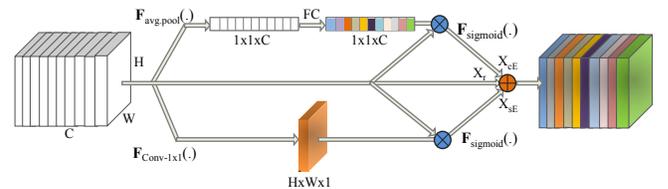

Fig. 2. Our proposed Skip spatial-channel squeeze and excitation model. Skip eliminates the singularity and sparsity in the CNN learning.

B. Network Architecture

Our AP-MTL model forms of a shared-encoder, a segmentation decoder, and a detection block (illustrated in Fig. 1). The encoder and detection block adopt from SSD [7] where VGG16 [25] exploit in the encoder section. The segmentation decoder is designed with proposed skip-scSE by following our previous work [6], as shown in Fig. 3. Multi-scale feature maps are extracted from different stages in the encoder network. Feature network's lower level score maps are concatenated with high-level score maps to increase parameter learning, and then it gets convoluted and excited. Further excited score maps are upsampled with a deconvolution layer. The final semantic score map will be generated after the last upsampling, which is used to output the prediction results.
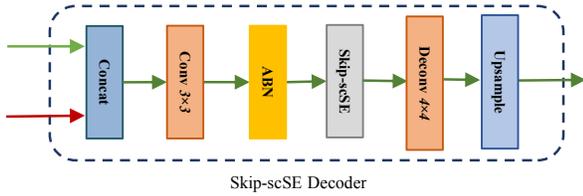

Fig. 3. Our proposed Skip-scSE Decoder for segmentation of surgical iinstruments.

C. Asynchronous Task-aware Optimization (ATO)

The fundamental problem in optimizing an MTL model is obtaining optimal minima at the same convergence point. In prior approaches, optimization of multiple tasks uses a naive weighted sum of losses [20], [21], [29], where the loss weights are uniform, or manually tuned. This kind of optimization can be related as optimizing a single task with added noise, where the noise correlates other task losses. To overcome these problems, we introduce Asynchronous Task-aware Optimization (ATO). In this technique, we optimize the task-aware spatial decoder and detection block by calculating gradient independently. Further, we simultaneously regularize end-to-end MTL model to attain a global convergence point for a few epochs with low learning rate.

In algorithm 1, we do Xavier initialization and construct mini-batches for training. At first, the model is trained with detection loss only by detaching segmentation decoder as detecting bounding box is more sensitive task than pixel-wise segmentation. Then, detach the detection block and train only segmentation decoder on segmentation loss by using encoder features trained for the detection task. Encoder features of detection task are also significant for the segmentation as both tasks are targeted on the same objects. Finally, to generalize the shared weights of the encoder, we opt end-to-end regularization by optimizing encoder and both detection block and segmentation decoder with small learning rate 10^{-5}. This ensures the MTL model to be more generalized and smooth gradient flow throughout the network.

Algorithm 1 Asynchronous Task-aware Optimization

Xavier Initialization
[building minibatch]
$B \Leftarrow \{i_1, ..., i_b\}, i_i \sim Z[1, b]$
[Initializing gradient]
$\mathbf{dW_e} \leftarrow 0, \mathbf{dW_{d1}} \leftarrow 0, ..., \mathbf{dW_{dn}} \leftarrow 0$
[Spatial optimization]
for $dn \in$ spatial **do** :
  [Encoder gradients]
  $\mathbf{dW_e} \leftarrow \mathbf{dW_e} + \sum_t \delta_{dn} \nabla_{W_e} \mathucal{L}_{dn}(f^i_{dn}(W_e, W_{dn}), y^i_{dn})$
  [Spatial Task-specific gradients]
  $\mathbf{dW_{dn}} \leftarrow \mathbf{dW_{dn}} + \sum_t \zeta_{dn} \nabla_{W_{dn}} \mathucal{L}_{dn}(f^i_{dn}(W_e, W_{dn}), y^i_{dn})$
end for
[Spatial regularization]
for $dn \in$ all **do** :
  [Encoder gradients]
  $\mathbf{dW_e} \leftarrow \mathbf{dW_e} + \alpha \sum_t \delta_{dn} \nabla_{W_e} \mathucal{L}_{dn}(f^i_{dn}(W_e, W_{dn}), y^i_{dn})$
  [Spatial Task-specific gradients]
  $\mathbf{dW_{dn}} \leftarrow \mathbf{dW_{dn}} + \alpha \sum_t \zeta_{dn} \nabla_{W_{dn}} \mathucal{L}_{dn}(f^i_{dn}(W_e, W_{dn}), y^i_{dn})$
end for
end =0

D. Global Attention Dynamic Pruning (GADP)

MTL methods are generally heavy in computation and not applicable in real-time, especially for high-resolution images. Every network contains a sub-network, and most networks' parameter learning is redundant and insignificant. Besides, redundancy and singularity in the learning parameters are the huge hindrances for task-specific decoders in MTL model. Thus, pruning can solve this over-parameterization and singularity problem. In prior works, network are pruned using Taylor expansion ranking [30] or attention statistics [31]. Nonetheless, pruning filters based on local gradient flow or activation is highly redundant, as it is a function of all previous filters. Therefore, we propose a dynamic pruning method based on global understanding of parameters and channel-wise attention [26], [32]. To avoid initial layer bottleneck pruning, we follow proportional pruning in each layer.
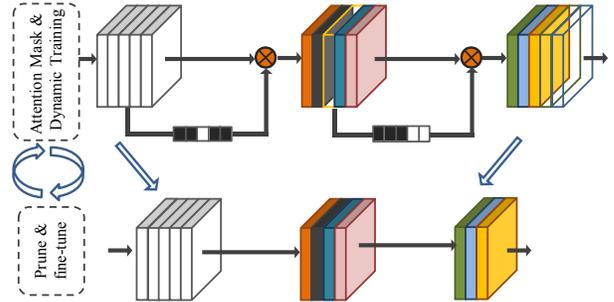

Fig. 4. Overview of our Global Attention Dynamic Pruning (GADP) method. GADP attaches in-between encoder blocks, calculate the rank of the each kernel, and remove low ranked kernels gradually by retraining the model. Attention-based pruning, rank the network parameters accurately which assists to remove redundant and weak parameters.

In algorithm 2, first, we append SE-Attention module to our pre-trained MTL model's encoder. Then, for a few epochs, we train only attention module weights to learn a

Algorithm 2 Global Attention Dynamic Pruning

Input: The pre-trained model $\mathcal{M}$, the training set $\mathcal{D}$, the testing set T, Accuracy, Drop d, attention iter I, finetune iter N.
Ouput: The pruned model $\mathcal{M}'$.
T_m = Accuracy for pre-trained model $\mathcal{M}$ with testing set T
Attach SE-Attention modules to the conv layers of $\mathcal{M}$.
1: **for** attention iteration $i = 1, 2, \ldots, I$ **do**
    Forward propagation with $\mathcal{D}$.
    Update only attention module in backpropagation.
  end for
  Calculate the attention statistics $a_{l,c}$ with $\mathcal{D}$.
  Threshold (t) according to proportional pruning strategy
2: **for** each layer $l = 1, 2, \ldots, L$ **do**
    for each channel $c = 1, 2, \ldots, C_l$ **do**
      if $a_{l,c} < t/C_l$ **then**
        mask the channel c from $\mathcal{M}$.
      end if
    end for
  end for
3: **for** finetune iteration $i = 1, 2, \ldots, N$ **do**
    Forward propagation with $\mathcal{D}$.
    Update only masked model $\mathcal{M}'$ in backpropagation.
  end for
  T_m' = Accuracy for masked model $\mathcal{M}'$ with testing set T
  if $T_m - T_m' < d$ **then**
    goto Step 1:
  else
    pruned model $\mathcal{M}'$ = prune masked channel in $\mathcal{M}'$
    for prune regularize iteration $i = 1, 2, \ldots, N$ **do**
      Forward propagation with $\mathcal{D}$.
      Update pruned model $\mathcal{M}'$ parameters in backpropagation.
    end for
    return pruned model $\mathcal{M}'$
  end =0

---

better global understanding of network filters correlation. Further, a pruning plan is generated with trained attention statistics by setting the mask threshold conditioned upon proportional pruning strategy. Then, we detach the attention modules and dynamically fine-tune the MTL model by masking channel without pruning. Repeat the whole process until our masked model accuracy drops below the specified margin with the required number of parameters to be masked. Finally, prune all masked channels to obtain an efficient weight shared encoder and regularize it for few epochs for better gradient flow and generalization. Overview of our proposed GADP is illustrated in Fig. 4.

III. Experiments

A. Dataset

This work is done using robotic instrument segmentation dataset of MICCAI endoscopic vision challenge 2017 [24]. This dataset contains 10 sequences recorded with the resolution of 1920 x 1080 using Da Vinci surgical systems [1]. In each sequence, significant instrument motion and visibility is observed and sampled at a rate of 1 Hz. The training set consist of sequence 1, 2, 3, 5, 6, 8, and cross-validation data consist of sequence 4, 7, and testing data consist of sequence 9, 10. For the training and cross-validation data sequence, we crop the frames into 1280 x 1024 by removing black padding

and annotate bounding box for the instrument with the aid of instrument type segmentation as shown in Fig. 5.
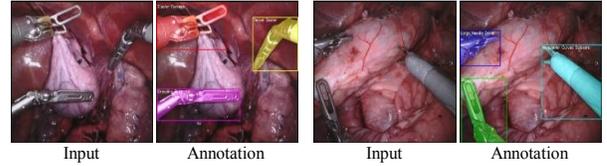

Fig. 5. Dataset visualization of two random frames. We annotate the bounding box on MICCAI robotic instrument segmentation dataset [24].

B. Implementation details

The input images are normalized by subtracting the image mean and dividing standard deviation. While training, we sequentially implemented algorithm 2 and algorithm 1. For other hyper-parameters, we use SGD optimizer with an initial learning rate 0.0001 and 'poly' learning rate with the power of 0.9 to update it. The momentum and weight decay set constant to 0.99 and 10^{-4} respectively.

C. Post-processing

In post-processing, we implemented a novel method for our architecture. Quantitatively we observed that our model can perform well in binary segmentation but under-perform in instrument type segmentation. Thus, we inferred that our model can discriminate instruments and tissues but struggles to differentiate between each instrument types. Moreover, we observed that as detection is conditioned upon IOU greater than 0.5, a major area of predicted bounding box covers the instrument [7]. However, in the case of the overlapping bounding box, the area of intersection is not subjected to post-processing. Therefore for high confidence predicted box, we assign all pixels inside the bounding box to its class in predicted instrument type output. By following this method, instrument type segmentation accuracy improved around 20-25%. Fig. 6 shows the model prediction before and after post-processing.
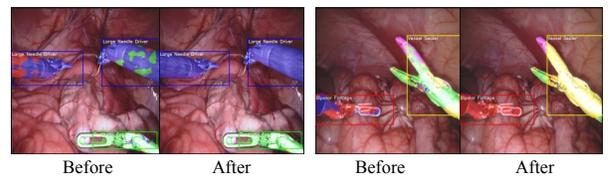

Fig. 6. Post-processing on predicted segmentation using bounding box. It boosts the type wise segmentation accuracy.

IV. Results and Evaluation

We evaluate segmentation accuracy of our model using evaluation metrics like dice coefficient, Hausdorff, sensitivity, specificity and for object detection accuracy by mean average precision (mAP) and real-time performance by frames per second (fps). In Table I, we compare our architecture with the existing state-of-the-art architectures for segmentation, object detection, and multi-tasking. Inference time for all the

TABLE I

EVALUATION SCORE FOR THE CROSS-VALIDATION DATASET. THE EVALUATION METRICS ARE DICE, HAUSDORFF(HAUSD.), MEAN AVERAGE PRECISION(MAP), FRAME PER SECOND (FPS) AND PARAMETERS(PARAM). FPS IS CALCULATED BY SINGLE RTX 2080 TI GPU WITH BATCH SIZE OF 1. THE BEST VALUES OF EACH METRIC ARE BOLDENED AND THE VALUES BETTER THAN OURS ARE UNDERLINED.

Model	Regime	Segmentation				Obj. Detect.	General	
		Binary		Type		mAP	FPS	Param
		Dice	Hausd.	Dice	Hausd.			
Ours	Seg. & Detect.	**0.947**	9.469	**0.704**	10.56	0.392	18	22.4M
MaskRCNN [20]	Seg.& Detect.	0.452	-	0.283	-	0.376	4	-
SSD [7]	Detect.	-	-	-	-	<u>0.406</u>	12	22.6M
YoloV3 [4]	Detect.	-	-	-	-	0.363	12	-
LinkNet [8]	Seg.	0.941	10.39	0.578	11.88	-	111	11.5M
ERFNet [33]	Seg.	0.923	11.05	0.495	11.55	-	44	<u>2.02M</u>
DeepLabv3+ [34]	Seg.	0.947	10.07	0.657	<u>10.47</u>	-	56	59.3M
scSE UNet [12]	Seg.	0.931	10.09	0.542	11.88	-	107	42.7M
TernausNet11 [35]	Seg.	0.920	11.25	0.446	12.29	-	<u>295</u>	34.5M
BiseNet [36]	Seg.	0.942	9.997	0.440	10.80	-	54	90.8M
GCN [37]	Seg.	0.940	10.48	0.612	13.98	-	83	23.9M

models is calculated by using a single NVIDIA RTX 2080 Ti GPU with a batch size of one.

A. Evaluation with Cross-validation

From Table I, our model outperforms state-of-the-art multi-tasking architecture like MaskRCNN [20] in both segmentation and object detection. Additionally in instrument type segmentation, our model surpasses single task state-of-the-art architectures such as LinkNet [8], scSE UNet [12] and ERFNet [33] with a large margin. In object detection, our model produces competitive results compared to single-task state-of-the-art network such as SSD [7]and perform better than YoloV3 [4]. In comparison with MTL model like MaskRCNN [20], our model outperforms type-wise segmentation accuracy. Also, produce a competitive result in object detection with outperforming real-time computation. Further, from the above results, multi-task learning is more generalized and scalable and produce better results than single-task learning [22]. Fig. 7 demonstrates the type segmentation prediction with our AP-MTL model.
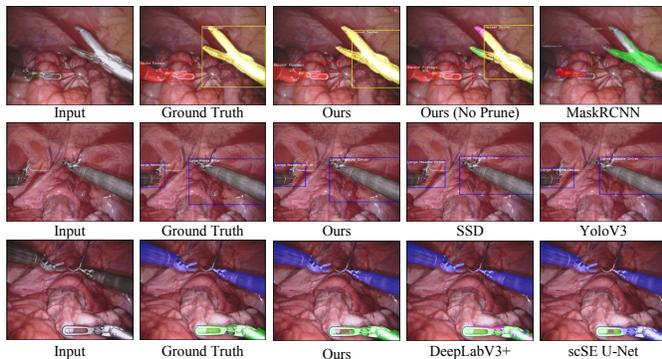

Fig. 7. The comparison of the type-wise segmentation and detection prediction. Predictions of our model contain lower false positive, false negative and competitive bounding box with SSD [7].

TABLE II

COMPARISON OF QUANTITATIVE RESULTS FOR THE BINARY AND TYPE BASE SEGMENTATION TASKS OF PROPOSED MODEL AND REPORTED PERFORMANCE IN [24]. THE BEST VALUES OF EACH METRIC ARE BOLDENED. THE VALUES BETTER THAN OURS ARE UNDERLINED.

Model	Binary Seg.		Type Seg.	
	Dataset9	Dataset10	Dataset9	Dataset10
Ours	**0.883**	0.892	0.350	**0.795**
MIT	0.865	0.905	<u>0.357</u>	0.609
SIAT	0.839	0.899	0.315	0.791
TUM	0.877	<u>0.909</u>	-	-
Delhi	0.626	0.715	-	-
UCL	0.808	0.869	0.272	0.583
NCT	0.789	0.899	0.247	0.552
UB	0.855	0.917	0.106	0.709
BIT	0.236	0.403	-	-
UA	0.539	0.689	0.040	0.715
UW	0.377	0.603	-	-

B. Evaluation with Challenge

Table II contains a mean intersection-over-union (IoU) performance comparison between our method and challenge methods for binary and type segmentation in MICCAI 2017 robotic instrument segmentation challenge [24]. Our model prediction of sequences 9 and 10 is evaluated by challenge portal[1]. We compare the prediction of our model with the binary and instrument type segmentation of the participated models in the challenge [24]. From the table, our model achieves competitive performance with all other participants' models for both binary and type segmentation task. From Fig. 8, our segmentation predictions of a less false positive and false negative than other challenge models.

V. ABLATION STUDY

A. Module and Methods

To evaluate our method and integration of the modules, we calculate the performance and complexity of the model, as shown in Table III. Proposed optimization approach, ATO

[1]https://endovissub2017-roboticinstrumentsegmentation.grand-challenge.org/
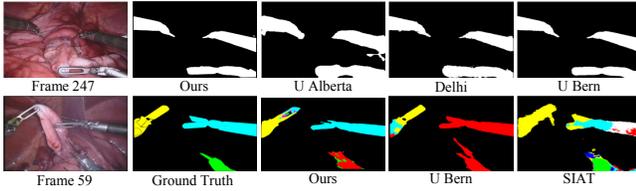
Fig. 8. Qualitative results comparison of frames from different dataset for binary and type segmentation with corresponding results from the methods of robotic instrument segmentation challenge [24] and proposed AP-MTL.

produces a considerable impact on our MTL model for both segmentation and detection tasks. It fixes the common challenge of MTL model to attain optimal minima at a convergence point (at same epoch). We monitor that MTL model without ATO converges the two tasks into different epochs. Without ATO, the convergence epoch of segmentation consists of poor detection accuracy in Table III. Moreover, ATO achieves the convergence of both tasks into the same epoch with the best accuracy for both tasks of segmentation and detection. Skip-scSE and GADP enhance the segmentation and detection accuracy with a significant margin, respectively. High-resolution images require high computational resource with larger learning parameters and remarkably reduce the prediction FPS (frame per second). Thus, GADP removes redundant parameters and boost fps. Moreover, MTL regularization smoothes gradient flow throughout the model, which increase generalization and optimized prediction. From Table IV, we can infer that our proposed attention model requires same learnable parameters but outperforms scSE [12].

TABLE III

OUR MODEL PERFORMANCE AND COMPLEXITY WHILE INTEGRATING DIFFERENT PROPOSED METHODS AND MODULES. REGU. DENOTES REGULARIZATION.

Modules and Methods				Seg.	Det.	FPS
ATO	Skip-scSE	GADP	Regu.	Dice	mAP	
✓	✓	✓	✓	0.704	0.392	**18**
✓	✓	✓	✗	0.685	0.392	**18**
✓	✓	✗	✗	**0.714**	**0.406**	10
✓	✗	✗	✗	0.695	**0.406**	11
✗	✗	✗	✗	0.610	0.379	11

TABLE IV

EXPERIMENTAL COMPARISON BETWEEN NO ATTENTION, SCSE [12] AND OUR ATTENTION PROPOSED MODEL. FROM RESULTS, WE CAN INFER THAT OUR ATTENTION MODULE ENHANCES SEGMENTATION PREDICTION BY A SIGNIFICANT MARGIN.

Modules	Seg.	Det.	FPS
	Dice	mAP	
No Attention	0.658	0.406	11
scSE	0.695	0.406	11
Skip-scSE	**0.714**	0.406	10

B. Single-task and Multi-task

From the comparison Table V between single-task and multi-task, the MTL model improves the performance of both

tasks comparing to single-task models for segmentation and detection. Our proposed MTL optimization technique, ATO, is playing crucial rule behind the performance improvement. Moreover, MTL regularization refines the learning parameters for both tasks and attains optimal convergence point.

TABLE V

PERFORMANCE COMPARISON BETWEEN SINGLE-TASK, MULTI-TASK MODEL AND REGULARIZATION. OUR MULTI-TASK MODEL ACHIEVES HIGHER PERFORMANCE THAN THE MODELS OF THE INDIVIDUAL TASKS.

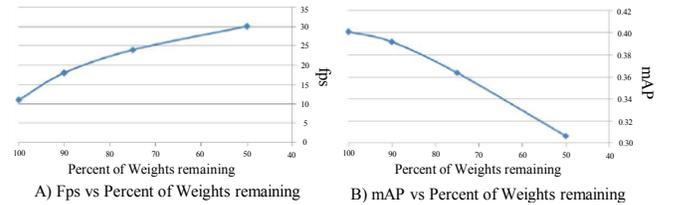| Task			Segmentation		Detection	FPS
Seg.	Det.	Regu.	Binary	Type	mAP	
			Dice	Dice		
✓	✗	✗	0.939	0.680		**56**
✗	✓	✗	-	-	**0.406**	12
✓	✓	✗	0.941	0.685	0.392	18
✓	✓	✓	**0.947**	**0.704**	0.392	18

C. Prune ration

Fig. 9. Graphical visualization of A) fps vs Percent of weights remaining B) mAP vs Percent of weights remaining. Total network parameters with 100 % weights remaining is 22.6 M.

Variation of model performance with pruning is studied and illustrated in Fig. 9. As we increase prune ration, mAP reduces by a small margin with significant improvement in fps. Attention-based pruning aids removal of redundant filters with global receptive field perception. Therefore, noise variants in the model can be reduced by pruning, and detection can be reinforced.

VI. CONCLUSION

In this paper, we present attention pruned multi-task learning model (AP-MTL) for real-time instrument localization and tracking in endoscopic surgery. We introduce attention based dynamic pruning technique to eliminate sparsity and singularity in MTL model's shared weight encoder. We present a novel method to optimize the task-aware MTL model to obtain the same optimal convergence point for multi-tasks. To enhance segmentation, we design a novel decoder using skip-scSE to reduce sparsity and redundancy in the decoder. Also, we introduced a novel post-processing method which exploits the benefit of multi-tasking for increasing segmentation accuracy. Further, our model outperforms most of the state-of-the-art architecture by a significant margin. In future work, we extend our research to segment defected tissues along with instruments and learning to exploit the surgical scene representation, compositionality, and reasoning.

REFERENCES

[1] J. C.-Y. Ngu, C. B.-S. Tsang, and D. C.-S. Koh, "The da vinci xi: a review of its capabilities, versatility, and potential role in robotic colorectal surgery," *Robotic Surgery: Research and Reviews*, vol. 4, pp. 77–85, 2017.

[2] D. Pakhomov, V. Premachandran, M. Allan, M. Azizian, and N. Navab, "Deep residual learning for instrument segmentation in robotic surgery," *arXiv preprint arXiv:1703.08580*, 2017.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[4] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[5] J. D. Seo, "Visualizing uncertainty and saliency maps of deep convolutional neural networks for medical imaging applications," *arXiv preprint arXiv:1907.02940*, 2019.

[6] M. Islam, Y. Li, and H. Ren, "Learning where to look while tracking instruments in robot-assisted surgery," *arXiv preprint arXiv:1907.00214*, 2019.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[8] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[9] A. Shvets, A. Rakhlin, A. A. Kalinin, and V. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," *arXiv preprint arXiv:1803.01207*, 2018.

[10] L. C. García-Peraza-Herrera, W. Li, L. Fidon, C. Gruijthuijsen, A. Devreker, G. Attilakos, J. Deprest, E. Vander Poorten, D. Stoyanov, T. Vercauteren, *et al.*, "Toolnet: holistically-nested real-time segmentation of robotic surgical tools," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5717–5722.

[11] M. Islam, D. A. Atputharuban, R. Ramesh, and H. Ren, "Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 2188–2195, April 2019.

[12] A. G. Roy, N. Navab, and C. Wachinger, "Concurrent spatial and channel 'squeeze & excitation'in fully convolutional networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 421–429.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[14] D. Sarikaya, J. J. Corso, and K. A. Guru, "Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection," *IEEE transactions on medical imaging*, vol. 36, no. 7, pp. 1542–1549, 2017.

[15] Z. Chen, Z. Zhao, and X. Cheng, "Surgical instruments tracking based on deep learning with lines detection and spatio-temporal context," in *2017 Chinese Automation Congress (CAC)*. IEEE, 2017, pp. 2711–2714.

[16] Z. Zhao, S. Voros, Y. Weng, F. Chang, and R. Li, "Tracking-by-detection of surgical instruments in minimally invasive surgery via the convolutional neural network deep learning-based method," *Computer Assisted Surgery*, vol. 22, no. sup1, pp. 26–35, 2017.

[17] E. Colleoni, S. Moccia, X. Du, E. De Momi, and D. Stoyanov, "Deep learning based robotic tool detection and articulation estimation with spatio-temporal layers," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2714–2721, 2019.

[18] I. Laina, N. Rieke, C. Rupprecht, J. P. Vizcaíno, A. Eslami, F. Tombari, and N. Navab, "Concurrent segmentation and localization for tracking of surgical instruments," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 664–672.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[21] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid, "Blitznet: A real-time deep network for scene understanding," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4154–4162.

[22] G. Sistu, I. Leang, and S. Yogamani, "Real-time joint object detection and semantic segmentation network for automated driving," *arXiv preprint arXiv:1901.03912*, 2019.

[23] I. Kokkinos, "Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6129–6138.

[24] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt, *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[27] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *2017 International Conference on 3D Vision (3DV)*. IEEE, 2017, pp. 11–20.

[28] A. E. Orhan and X. Pitkow, "Skip connections eliminate singularities," *arXiv preprint arXiv:1701.09175*, 2017.

[29] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[30] P. Molchanov, S. Tyree, T. Karras, T. Aila, and J. Kautz, "Pruning convolutional neural networks for resource efficient inference," *arXiv preprint arXiv:1611.06440*, 2016.

[31] K. Yamamoto and K. Maeno, "Pcas: Pruning channels with attention statistics for deep network compression," *arXiv preprint arXiv:1806.05382*, 2018.

[32] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[33] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.

[34] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[35] V. Iglovikov and A. Shvets, "Ternausnet: U-net with vgg11 encoder pre-trained on imagenet for image segmentation," *arXiv preprint arXiv:1801.05746*, 2018.

[36] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 325–341.

[37] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters–improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4353–4361.