2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

Learning Domain Adaptation with Model Calibration for Surgical Report Generation in Robotic Surgery

Mengya Xu*, Mobarakol Islam*, Chwee Ming Lim and Hongliang Ren

Abstract— **Generating a surgical report in robot-assisted surgery, in the form of natural language expression of surgical scene understanding, can play a significant role in document entry tasks, surgical training, and post-operative analysis. Despite the state-of-the-art accuracy of the deep learning algorithm, the deployment performance often drops when applied to the Target Domain (TD) data. For this purpose, we develop a multi-layer transformer-based model with the gradient reversal adversarial learning to generate a caption for the multi-domain surgical images that can describe the semantic relationship between instruments and surgical Region of Interest (ROI). In the gradient reversal adversarial learning scheme, the gradient multiplies with a negative constant and updates adversarially in backward propagation, discriminating between the source and target domains and emerging domain-invariant features. We also investigate model calibration with label smoothing technique and the effect of a well-calibrated model for the penultimate layer's feature representation and Domain Adaptation (DA). We annotate two robotic surgery datasets of MICCAI robotic scene segmentation and Transoral Robotic Surgery (TORS) with the captions of procedures and empirically show that our proposed method improves the performance in both source and target domain surgical reports generation in the manners of unsupervised, zero-shot, one-shot, and few-shot learning.**

I. Introduction

The popularity of Minimally Invasive Surgeries (MIS) in modern medical treatment has brought new opportunities and challenges for automated surgical scene understanding, which can be used to empower Computer Assisted Interventions (CAI) and is the foundation of intelligent systems, such as robotic vision, surgical training applications, and medical report generation. This kind of intelligent system can remind surgeons not to miss important steps during complex surgery. By integrating with natural language report generation, the intelligent system can also eliminate the frustrating task of document entry task for clinicians, allowing them to focus on patient-centric activities. It can free doctors and nurses from the low-value task of writing reports of surgical procedures that are more suitable for machines. Besides, such a natural language record of the surgical procedure has a detailed post-operative analysis of the surgical intervention.

Automatically generating the description for a given surgical procedure is a complicated problem since it requires an algorithm to complete several computer vision and Natural Language Processing (NLP) tasks, such as object recognition, relationships understanding between vision and text elements, and then generate a sequence of words. Existing studies for automatic report generation focus on medical images, such as radiology and pathology images. In this case, the report is a diagnosis report which will describe whether the body part examined in the imaging technique was normal, abnormal, or potentially abnormal [1]. In our work, the automatically generated report focuses more on understanding and describing the surgical scene. It describes which surgical instrument appears in the surgical scene and the instrument-tissue interaction.

Existing image captioning models cannot generalize well to the Target Domain (TD) images. Learning a predictor when there is a shift in data distribution between training and validation is considered as Domain Adaptation (DA). The cost of generating labels for data is high, which often becomes one of the biggest obstacles preventing the application of the machine learning approach. However, the model with the DA ability can work when the TD lacks labels.

A. Related Work

1) Surgical Scene Understanding: Surgical scene understanding is significant for image-guided robot-assisted surgery. Incorporating scene understanding capability in robotic surgery provides possibilities for future semi-automated or fully automated operations. Our previous works include surgical instrument tracking with task-directed attention [2] [3], and real-time instrument segmentation [4]. After the surgical instrument recognition problem has been solved to a large extent, the next goal is to get out the instrument itself and focus on the relationships between the instruments and the Region of Interest (ROI). Therefore, a graph-based network is introduced in [5] to do deep reasoning for the surgical scene, learn to infer a graph structure of instruments-tissue interaction, and predict the relationship between instruments and tissue. The further intention is that we want to express this prediction of their relationship in natural language, which can have richer scene information and the ability to interact with people.

2) Transformer based model from language to vision: The Transformer [6] is proposed for language translation which achieves the state-of-the-art performance. In combining scene understanding and natural language, Convolutional Neural Network (CNN) is used to extract visual features,

* equal contribution

M. Xu and H. Ren are with Dept. of Biomedical Engineering, National University of Singapore, Singapore and NUSRI Suzhou China; Corresponding author: Hongliang Ren, hlren@ieee.org http://labren.org

H. Ren is with Dept. of Electronic Engineering, The Chinese University of Hong Kong

M. Islam is with Dept. of Computing, Imperial College London, UK and was with Dept. of Biomedical Engineering, National University of Singapore, Singapore

C. M. Lim is with Dept. of Otolaryngology-Head and Neck Surgery, Duke-NUS Medical School, Singapore

978-1-7281-9077-8/21/$31.00 ©2021 IEEE

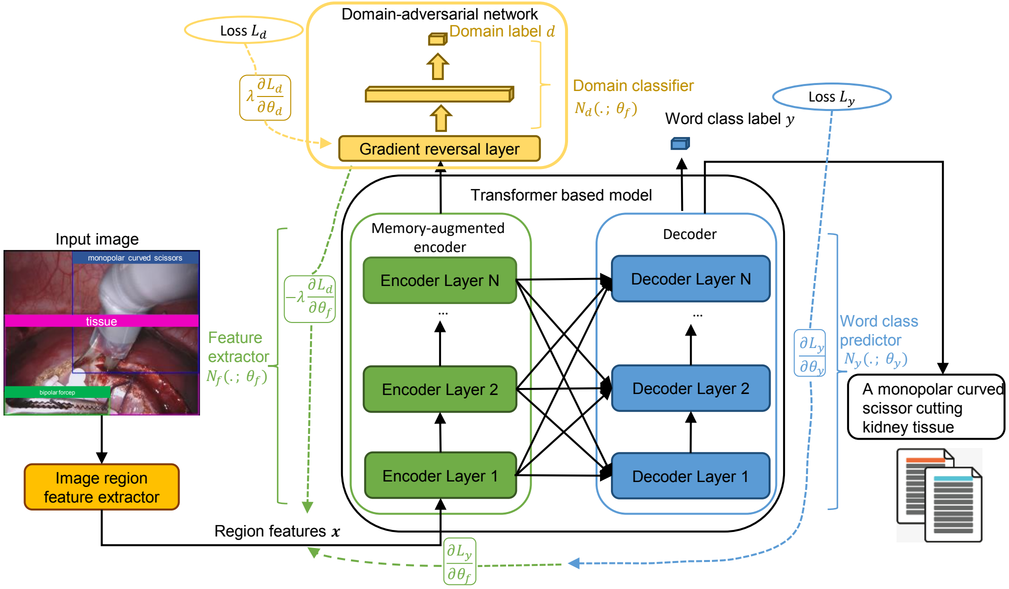


Fig. 1. The architecture of the proposed network. The network comprises a region feature extractor, multi-layer Transformer-like encoder-decoder, and a domain-adversarial network. ResNet18 is used to extract region features in input images. The domain-adversarial network, which consists of a GRL and a domain classifier, is connected with the last encoding layer to distinguish whether the sample is from the SD or the TD. The image encoder incorporated with domain-adversarial neural network takes in the region features and learns the domain-invariant features with an adversarial learning scheme and attempts to understand relationships between regions. The language decoder reads the output of each encoding layer to generate the output sentence word by word. The stacked sentence can form a medical report.

and Recurrent Neural Network (RNN) is widely adopted as the language models, which can be summarized as "CNN + RNN" framework [7]. Although this framework is widely adopted, the RNN-based model is severely limited by its sequential nature and difficult to parallelize characteristics. Some recent researches have used the Transformer model as the language model for scene understanding in which the input sequence can be passed in parallel. [8] still utilized the original architecture of the Transformer and explored the geometric relationships between detected objects. A memory-augmented recurrent Transformer [9] is employed to do video captioning.

3) Label Smoothing: Label Smoothing (LS) is a regularization technique that flattens the true label with a uniform distribution during cross-entropy loss calculation in network training [10]. Recently, there is evidence that LS can prevent over-confidence in prediction and improve the model calibration [11]. Moreover, LS is also used to enhance the prediction uncertainty [12] and the teacher-free knowledge distillation [13]. In our previous work, we present that LS can improve the feature representation in the penultimate layer and effective as an object feature extraction model [5].

4) Domain Adaptation: The model's performance tends to drop when evaluated on a TD dataset, which is different from the dataset used for training. This limitation motivates the research on DA [14] in multiple sclerosis lesion segmentation in one shot way. The method for DA in [15] consists of domain adversarial learning and consistency training. Many methods attempt to match the feature space distribution in the Source Domain (SD) and the TD for unsupervised DA (UDA). To this end, [16] reweigh or select examples of the SD. [17] achieves this matching by modifying and changing the feature representation itself. For supervised DA, the approaches utilize the labeled data from the TD to "fine-tune" the network trained on the SD. In our work, the unsupervised and semi-supervised DA are explored.

B. Contributions

Our contributions are summarized as follows:

- Propose transformer-based multi-layer encoder-decoder architecture with the gradient reversal adversarial learning scheme to generate the surgical report in robotic surgery.
- Investigate model calibration with label smoothing cross-entropy loss for feature extraction from penultimate layer and UDA.
- Design domain-adaptation in various unsupervised and semi-supervised manners such as zero-shot, one-shot, and few-shot training.
- Annotate robot-assisted surgical datasets with proper captions to generate the surgical report in two different surgical domain datasets of MICCAI robotic scene segmentation challenge and Transoral Robotic Surgery (TORS).

II. METHODS

A. Background

Transformer-based multi-layer encoder-decoder in [18] is a fully-attentive model and shows excellent performance for image captioning tasks. It is a variant of the original Transformer [6] which achieves the state-of-the-art results in machine translation and language understanding. In [18], the encoder module takes regions from images as input and understands relationships between regions. The decoder reads each encoding layer's output to model a probability over words in the vocabulary and generates the output sentence word by word by feeding the predicted word back as input at the next time step. Compared with [6], [18] has made two changes: 1) multi-head attention in image encoder is augmented with memory to understand the semantic relationships between detected input objects; 2) the cross-attention in language decoder is devised to utilize all encoding layers, rather than attending only the last encoding layer.

Despite the excellent accuracy of the deep learning frameworks in vision tasks, it performs poorly in the TD dataset. To solve it, [17] proposes a novel representation learning method called domain adversarial learning for DA. The goal is to learn features that have 1) distinction for the major task during learning on the SD and 2) no distinction regarding the change of domains. H-divergence measures the distance between the source and the target distributions used by [19]. A Gradient Reversal Layer (GRL) with an adversarial learning scheme is used to domain adaption tasks with classification problem [17].

Most recently, it is found that the current deep learning models are poorly calibrated, which drops the generalization capability of the model [11], [20], [21]. LS is showing evidence to improve model calibration and prevent the overconfidence of model learning. [11] presents that LS calibrates the representations learned by the penultimate layer and makes the features of the same class form a tight cluster.

Inspiring by the above works, we design a transformer-based multi-layer encoder-decoder architecture with a GRL domain classifier and label smoothing cross-entropy loss to generate the surgical reports in robotic surgery.

B. GRL Domain Classifier

Domain classifier forms of GRL with a discriminator and trains in an adversarial manner to learn domain invariant features [17]. It trains with adversarial labels that 0 label for the SD and 1 label for the TD. GRL just acts as identity transformation layer, $G(x) = x$ in forwarding propagation and a negative constant multiplies the gradient in backward propagation ($\frac{dG}{dx} = -I$, where I is an identity matrix) and then passes it to the preceding layer. By this way, GRL subtracts the gradients of the main network (our caption generation model) and domain classifier instead of being summed during Stochastic Gradient Descent (SGD) training. On the other hand, this prevents SGD from making features dissimilar across the domains. In this case, the model attempts to learn domain invariant features, making the domain classifier indistinguishable.

We propose two significant improvements of the GRL domain classifier module: (1) we assign the discriminator with 3 classes (0 for source, 1 for target, and 2 for extra class) to enhance learning capacity, which will make the discriminator harder to fool during training; (2) integrate an additional fully-connected layer to boost feature learning, as shown in Fig. 2

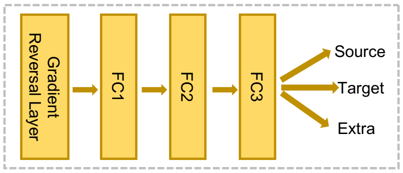


Fig. 2. Proposed GRL domain classifier. It consists of a GRL layer and three fully-connected layers. The output has 3 classes: SD, TD, and the extra one.

C. LS for model calibration and DA

A well-calibrated model estimates confidence probability, which represents true correctness of likelihood and leads to better generalization. LS, where a model is trained on a smoothened version of the true label with Cross-Entropy (CE) loss, shows great effectiveness in improving model calibration [11]. If smoothened label, $T_{LS} = T(1-\epsilon)+\epsilon/K$ then CE loss with LS can be formulated as $CE_{LS} = -\sum_{k=1}^{K} T_{LS} log(P)$ where true label T, smoothing factor ϵ, total number of classes K and predicted probability P. A recent study [5] investigates that LS learns better feature representation in the penultimate layer, which is effective in object feature extraction. In this work, we observe the behavior of LS trained model in the DA task.

D. Network Architecture

The whole network consists of an image region feature extractor, Transformer-based multi-layer encoder-decoder [18], and a GRL domain classifier network. We choose a lightweight feature extraction model of ResNet18 [22], similar as [5], to obtain feature vectors of surgical objects (details in section III-B.1). The GRL domain classifier network consists of a discriminator followed by the GRL as in Fig. 1. The encoder block forms of memory augmented self-attention layer and feed-forward layer and a decoder block consists of self-attention on words and cross attention over the outputs of all the encoder layers, similar to [18]. There are three encoder and decoder blocks stack to encode input features and predict the word class label. The output of the final encoder layer is fed to the proposed GRL domain classifier. Each training iteration calculates losses from both SD and TD, and then caption prediction loss is combined with domain classifier losses to update the model parameters in an adversarial manner. We utilize CE loss with LS as the caption prediction loss and vanilla CE as the domain loss function. Therefore, the model loss can be formulated as

$$L = L_y + L_d, \tag{1}$$

where $L_y = CE_{LS}$ is the caption prediction loss which is CE loss with LS and $L_d = L_S + L_T$ is the domain loss which is the fused losses of the SD loss L_S and the TD loss L_T.

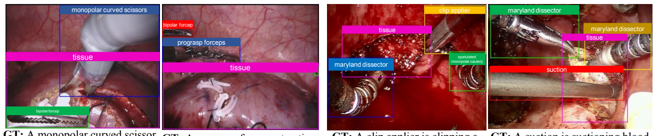
The GRL domain classifier leads to the appearance of features that have a distinction for the captioning task and domain-invariance at the same time. The model attempts to learn a representation allowing the decoder to predict the word class, also weakening the ability of the domain classifier.

III. Experiments

A. Dataset

1) Robotic scene segmentation challenge: The SD dataset is from the robotic instrument segmentation dataset of MICCAI endoscopic vision challenge 2018 [23]. The training set includes 14 robotic nephrectomy operations obtained by the da Vinci X or Xi system. There are 149 frames in each video sequence, and the frame has a dimensionality of 1280*1024. A total of 9 objects appear in the dataset, including 8 Instruments: bipolar forceps, prograsp forceps, monopolar curved scissors, clip applier, suction, ultrasound probe, stapler, and large needle driver respectively and kidney tissue. These surgical instruments have a variety of different semantic relationships and interactions with the tissue. A total of 11 kinds of semantic relationships are identified to generate the natural language description for images. The identified semantic relationships include manipulating, grasping, retracting, cutting, cauterizing, looping, suctioning, clipping, ultrasound sensing, stapling, and suturing. The Sequences 1st, 5th, 16th are chosen for the validation, and the remaining 11 sequences are selected for the training following the previous work [5]. The training and validation data sequences are carefully chosen to ensure that most interactions are presented in both sets.

2) TORS dataset: We have collected the TD dataset of TORS surgery using the Da Vinci Robot on human patients. All 9 patients have oropharynx cancer. Among these 9 patients, there are 8 males and 1 female with ages ranging from 23 to 72 and multiple cancer sites like the tonsil, the base of the tongue, and the posterior pharyngeal wall. We filter out some poor-quality videos that remain still and are not undergoing surgery. A total of 181 frames were selected based on the requirement to have at least one same surgical instrument as the robotic scene segmentation challenge dataset. These 181 frames are cropped and then resized to the same shape with MICCAI Robotic scene segmentation challenge dataset. A total of 5 objects are in the dataset, including tissue, clip applier, suction, spatulated monopolar cautery, maryland dissector. The 5 kinds of relationships comprise manipulating, grasping, cauterizing, suctioning, clipping. We split the dataset based on different patients. The training dataset has 48 frames collected from Patient [1, 2, 3, 4, 5, 6, 7] and the validation dataset has 133 frames collected from Patient [8, 9].



GT: A monopolar curved scissor cutting tissue

GT: A prograsp forcep retracting tissue

GT: A clip applier is clipping a clip

GT: A suction is suctioning blood and a spatulated monopolar cautery is cauterizing tissue

(a) Source domain

(b) Target domain

Fig. 3. Visualization of image-caption pairs from the SD and the TD. There are many unique ground truth captions

3) Domain Shift: Fig. 3 shows example image-caption pairs from the SD and the TD. The SD is about robotic nephrectomy surgery, and the TD is about TORS. Compared with the SD, the TD has a different surgical background, and the surgery is performed on different tissue. Besides, the TD does not share all surgical instruments with the SD. Compared with the SD, the TD has two new instruments: spatulated monopolar cautery and maryland dissector. The distribution shift between the SD and TD can be distinguished from Fig. 4, which shows the t-SNE of examples before DA from a three-dimensional (3D) perspective, which can prove the distribution shift between these two domains.

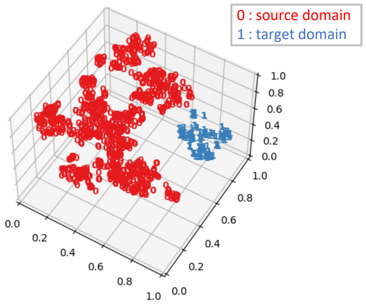


Fig. 4. t-SNE of samples before DA. The "0" indicated by red color are examples from the SD, and the "1" indicated by blue color are examples from the TD. The SD examples do not overlap with examples from the TD because of the distribution shift between these two domains.

4) Annotation: The datasets are annotated with the appropriate caption by an experienced clinical expert in robotic surgery. The caption follows the description pattern of ⟨ object1, predicate, object2 ⟩, which can concisely and accurately convey information of the surgical scene, for instance, "A monopolar curved scissor is cutting tissue."

B. Implementation details

1) Image region feature extractor: We use the ResNet18 classification model to extract the feature vector for surgical objects. We crop the surgical instruments and ROI tissue and train the ResNet18 with CE loss combined with LS by following our previous work [5]. The penultimate layer's feature vector of size 512 is extracted for each surgical object. Fig. 5 shows the extracted features without (w/o) and with (w) LS. The features extracted by the LS model are more distinguishable among the classes.

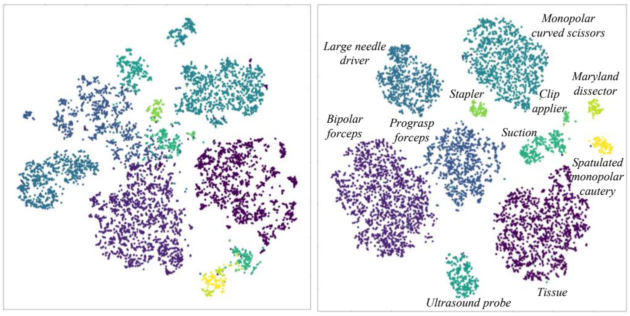
2) Vocabulary and tokenization: All captions will go through a series of processing, including converting them to lower cases, removing punctuation characters, and tokenization using the spacy NLP toolkit. Vocabulary size of 49 is built based on both SD and TD, which include unique words in the captions and special tokens (⟨ unk ⟩, ⟨ pad ⟩, ⟨ bos ⟩, and ⟨ eos ⟩). At test time, the model predicts the next word given the previously predicted words with beam search.

3) Hyper-parameters: Following the experimental setting in [18], when training with cross-entropy loss, the learning rate scheduling strategy of [6] with a warmup equal to 10000 iterations is used. The number of epochs for training is set to 50. The seed is set to a fixed number to make the results

TABLE I

PERFORMANCE OF THE PROPOSED MODEL ON SD AND TD DATASET. EVALUATION METRICS SUCH AS BLEU-N, METEOR, ROUGR, AND CIDER ARE USED TO ASSESS THE SIMILARITY BETWEEN THE GENERATED SENTENCE AND THE GROUND-TRUTH

			BLEU-1↑	BLEU-2↑	BLEU-3↑	BLEU-4↑	METEOR↑	ROUGE↑	CIDEr↑
SD		M^2 Transformer [18]	0.5054	0.4543	0.4055	0.3646	0.4441	0.6355	1.7878
		Ours	0.5228	0.4730	0.4262	0.3861	0.4567	0.6495	2.2598
TD	UDA	M^2 Transformer [18]	0.2302	0.1059	0.0469	0.0267	0.1286	0.2956	0.1305
		Ours	0.2493	0.1150	0.0517	0.0289	0.1390	0.3129	0.1517
	Zero-shot	M^2 Transformer [18]	0.3204	0.2463	0.1923	0.1502	0.2371	0.4413	0.2874
		Ours	0.3118	0.2406	0.185	0.1409	0.2401	0.4336	0.3395
	One-shot	M^2 Transformer [18]	0.3746	0.3285	0.2939	0.2646	0.3449	0.5101	0.6367
		Ours	0.4042	0.372	0.3433	0.3161	0.4066	0.5385	0.8615
	Few-shot	M^2 Transformer [18]	0.4096	0.3803	0.3532	0.3265	0.4203	0.5489	0.9770
		Ours	0.4141	0.3888	0.3637	0.3375	0.4357	0.5538	0.9828



W/O LS

W LS

Fig. 5. The extracted features of the penultimate layer for ResNet18 w and w/o label smoothing (LS). The t-SNE plots show that the extracted features with LS are more distinguishable and tighter clusters for the same class.

reproducible. We train the model using Adam optimizer with a batch size of 50 and a beam size of 5. The network was implemented by PyTorch and trained in the NVIDIA RTX 2080 Ti GPU.

IV. RESULTS AND EVALUATION

A. Evaluation Metrics

We evaluate the approach using four commonly used metrics for image captioning, namely BLEU-n [24], ROUGE [25], METEOR [26], and CIDEr [27]. BLEU-n measures the precision of n-grams between the ground-truth and the predicted sentences. Usually, n = 1, 2, 3, 4. To measure model calibration error, we use well-known metrics like Expected Calibration Error (ECE) [20], Static Calibration Error (SCE) [28], Thresholded Adaptive Calibration Error (TACE), and Brier Score (BS) [28], [29].

B. Caption generation with the SD dataset

Table I represents the overall performance of the proposed approach with validation and domain adaption settings. The validation metrics are calculated with the SD dataset. Our method achieves better performance in almost all the metrics.

C. Caption generation with the TD dataset

Our model's domain adaption performance is evaluated with extensive unsupervised and semi-supervised settings such as zero-shot, one-shot, and few-shot training, as shown in Table I. All TD experiments share the same validation set and just use a different number of training images. In UDA, the model trained on SD is evaluated directly on TD. In zero-shot DA, the trained model is fine-tuned with the frames whose captions cover 85% of the words in the vocabulary from the TD dataset. In one-shot DA, it is fine-tuned with images whose captions cover all the words with the smallest amount of images. Few images whose captions cover all the words are used to fine-tune the model. The proposed model has also outperformed the base model for these experiments.

D. Model Calibration Vs. DA

TABLE II

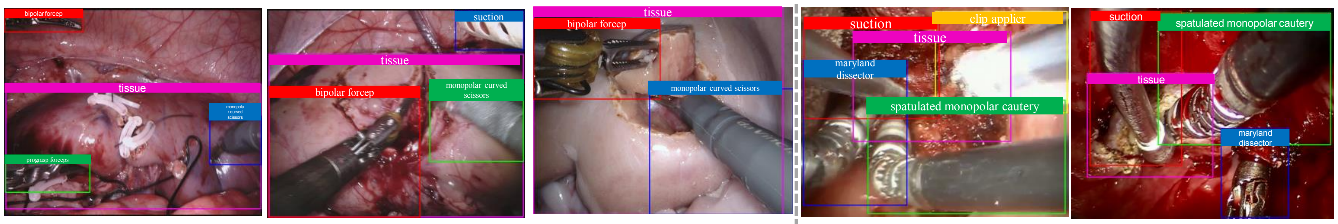
MODEL MISCALIBRATION QUANTIFICATION FOR OUR MODEL W AND W/O LS. EVALUATION METRICS SUCH AS EXPECTED CALIBRATION ERROR (ECE), STATIC CALIBRATION ERROR (SCE), THRESHOLDED ADAPTIVE CALIBRATION ERROR (TACE), BRIER SCORE (BS) ARE USED TO CALCULATE THE CALIBRATION ERROR FOR EACH MODEL.

Ours	TD(UDA)	Calibration Error			
Methods	BLEU-1↑	ECE↓	SCE↓	TACE↓	BS↓
w LS	0.2493	0.1768	0.0493	0.0489	0.5063
w/o LS	0.2302	0.2001	0.0533	0.0531	0.9585

A well-calibrated model produces smoother output distribution, which leads to better generalization and feature learning. In this experiment, we investigate the behavior of a better-calibrated model in the TD validation. As LS can prevent the over-confident prediction and improve probability calibration [11], we observe the TD data prediction performance of our approach trained between w and w/o LS. Table II demonstrates the comparative scores with the TD prediction and calibration errors. The results indicate the superiority of the calibrated model on the DA task.

E. Qualitative Analysis

We visualize some examples predicted by our proposed model for the source and target domain, as shown in Fig. 6. Sometimes there are redundant words at the end of the predicted sentence, which may be caused by exposure bias [30] and word-level training by using the traditional CE loss [30]. This problem will be solved by introducing a Reinforcement Learning training technique to achieve sequence-level training [31] and fine-tune the sequence generation.



GT: A prograsp forcep and a monopolar curved scissor manipulating tissue
Base: A monopolar curved scissor is cutting tissue and a bipolar forcep is retracting tissue with the suture
Ours: A monopolar curved scissor is cutting tissue and a bipolar forcep is manipulating tissue

(1)

GT: A bipolar forcep and a monopolar curved scissor manipulating tissue
Base: A monopolar curved scissor is cutting tissue and a bipolar forcep is retracting tissue
Ours: A monopolar curved scissor is cutting tissue and a bipolar forcep is manipulating tissue

(2)

GT: A monopolar curved scissor is cutting kidney tissue and a bipolar forcep is manipulating kidney tissue
Base: A monopolar curved scissor is cutting tissue and a bipolar forcep is retracting tissue with retracting tissue
Ours: A monopolar curved scissor is cutting tissue and a bipolar forcep is retracting tissue

(3)

Source domain

GT: A suction is suctioning blood and a clip applier is clipping a clip
Ours: A suction is suctioning blood and a spatulated monopolar cautery and a clip manipulating tissue and a and a

(4)

GT: A suction is suctioning blood and a spatulated monopolar cautery is cauterizing tissue
Ours: A suction is suctioning blood and a spatulated monopolar cautery is cauterizing tissue and a cauterizing tissue

(5)

Target domain

Fig. 6. Several image captioning examples generated by the base model and our model. Compared with the base model, the sentences predicted by our model are more reasonable and accurate.

Overall, the generated sentences can describe the content of the surgical images accurately.

V. Ablation Study

A. LS Feature Extraction

Fig. 5 is illustrated the better feature representation extracted from ResNet18 classification model trained with LS. In this section, we present the caption prediction models' performance, train w, and w/o LS extracted features. Fig. 7 demonstrates the significant prediction improvement using LS extracted feature.

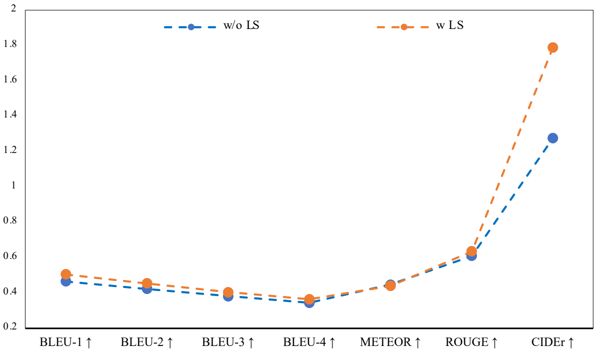


Fig. 7. Caption prediction performance w and w/o LS feature extraction using ResNet18. LS extracted feature boosts the prediction performance.

TABLE III
Ablation Study of the Proposed Model While Integrating LS and GRL for the SD and the TD (UDA) Experiments.

Modules			SD		TD (UDA)	
Base	LS	GRL	BLEU-1	CIDEr	BLEU-1	CIDEr
✓	✓	✓	0.5228	2.2598	0.2493	0.1517
✓	✓	×	0.5161	2.3406	0.2345	0.1314
✓	×	×	0.4659	1.2752	0.2302	0.1305

B. GRL Domain Classifier

Table III shows the effect of each proposed technique in this work. GRL domain classifier boosts the model prediction of both SD and TD datasets for most of the metrics. However, LS produces a significant enhancement in SD prediction. Table IV demonstrates that the extra class improves the model performance by making the model harder to be fooled.

TABLE IV
Two vs. Three Outputs of GRL Domain Classifier

GRL Domain Classifier	SD		
	BLEU-1	METOR	CIDEr
Two classes (SD, TD)	0.5231	0.4566	2.1547
Three classes (SD, TD, Extra)	0.5228	0.4567	2.2598

VI. Discussion and Conclusion

We present a multi-layer encoder-decoder transformer-like model incorporating the gradient reversal adversarial learning and LS to generate captions for the SD and the TD surgical images, which can describe the semantic relationship between instruments and surgical ROI. We observe that a well-calibrated model can be beneficial for domain adaption tasks. The experimental results prove that the proposed model can perform better than the base model in the SD and the TD. When these sentences are stacked, a medical report for the surgical procedures can be obtained. The limitation is the small TD dataset. The performance can be further improved if we can use a larger dataset. Future work can be focused on integrating temporal information in surgical report generation.

VII. Acknowledgment

This work was supported by the National Key Research and Development Program, The Ministry of Science and Technology (MOST) of China (No. 2018YFB1307700), Singapore Academic Research Fund under Grant R397000353114 and TAP Grant R-397-000-350-118.

REFERENCES

[1] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[2] M. Islam, V. VS, and H. Ren, "Ap-mtl: Attention pruned multi-task learning model for real-time instrument detection and segmentation in robot-assisted surgery," *arXiv preprint arXiv:2003.04769*, 2020.

[3] M. Islam, Y. Li, and H. Ren, "Learning where to look while tracking instruments in robot-assisted surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 412–420.

[4] 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, 2019.

[5] M. Islam, L. Seenivasan, L. C. Ming, and H. Ren, "Learning and reasoning with the graph structure representation in robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 627–636.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[7] Z. Shi, X. Zhou, X. Qiu, and X. Zhu, "Improving image captioning with better use of captions," *arXiv preprint arXiv:2006.11807*, 2020.

[8] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Advances in Neural Information Processing Systems*, 2019, pp. 11 137–11 147.

[9] J. Lei, L. Wang, Y. Shen, D. Yu, T. L. Berg, and M. Bansal, "Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning," *arXiv preprint arXiv:2005.05402*, 2020.

[10] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2818–2826.

[11] R. Müller, S. Kornblith, and G. E. Hinton, "When does label smoothing help?" in *Advances in Neural Information Processing Systems*, 2019, pp. 4694–4703.

[12] Z. Zhang and M. R. Sabuncu, "Self-distillation as instance-specific label smoothing," *arXiv preprint arXiv:2006.05065*, 2020.

[13] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisit knowledge distillation: a teacher-free framework," *arXiv preprint arXiv:1909.11723*, 2019.

[14] S. Valverde Valverde, M. Salem, M. Cabezas Grebol, D. Pareto, J. C. Vilanova Busquets, L. Ramió i Torrentà, À. Rovira, J. Salvi, A. Oliver i Malagelada, and X. Lladó Bardera, "One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks," *NeuroImage: Clinical, 2019, vol. 21, p. 101638*, 2019.

[15] M. Orbes-Arteaga, T. Varsavsky, C. H. Sudre, Z. Eaton-Rosen, L. J. Haddow, L. Sørensen, M. Nielsen, A. Pai, S. Ourselin, M. Modat, *et al.*, "Multi-domain adaptation in brain mri through paired consistency and adversarial learning," in *Domain Adaptation and Representation Transfer and Medical Image Learning with Less Labels and Imperfect Data*. Springer, 2019, pp. 54–62.

[16] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[18] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 578–10 587.

[19] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *Advances in neural information processing systems*, 2007, pp. 137–144.

[20] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *arXiv preprint arXiv:1706.04599*, 2017.

[21] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *Advances in neural information processing systems*, 2017, pp. 6402–6413.

[22] 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.

[23] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen, *et al.*, "2018 robotic scene segmentation challenge," *arXiv preprint arXiv:2001.11190*, 2020.

[24] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[25] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[26] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[27] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[28] J. Nixon, M. W. Dusenberry, L. Zhang, G. Jerfel, and D. Tran, "Measuring calibration in deep learning." in *CVPR Workshops*, 2019, pp. 38–41.

[29] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov, "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning," *arXiv preprint arXiv:2002.06470*, 2020.

[30] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," *arXiv preprint arXiv:1511.06732*, 2015.

[31] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.