Surgical Video Captioning with Mutual-Modal Concept Alignment

Zhen Chen[1], Qingyu Guo[1], Leo K. T. Yeung[2], Danny T. M. Chan[2], Zhen Lei[1,3], Hongbin Liu[1,3], and Jinqiao Wang[1,3,4,5(✉)]

[1] Centre for Artificial Intelligence and Robotics (CAIR), Hong Kong Institute of Science and Innovation, Chinese Academy of Sciences, Beijing, China
zhen.chen@cair-cas.org.hk, jqwang@nlpr.ia.ac.cn
[2] Department of Surgery, The Chinese University of Hong Kong, Shatin, Hong Kong
[3] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[4] Wuhan AI Research, Wuhan, China
[5] ObjectEye Inc., Beijing, China

Abstract. Automatic surgical video captioning is critical to understanding surgical procedures, and can provide the intra-operative guidance and the post-operative report generation. As the overlap of surgical workflow and vision-language learning, this cross-modal task expects precise text descriptions of complex surgical videos. However, current captioning algorithms neither fully leverage the inherent patterns of surgery, nor coordinate the knowledge of visual and text modalities well. To address these problems, we introduce the surgical concepts into captioning, and propose the Surgical Concept Alignment Network (SCA-Net) to bridge the visual and text modalities via surgical concepts. Specifically, to enable the captioning network to accurately perceive surgical concepts, we first devise the Surgical Concept Learning (SCL) to predict the presence of surgical concepts with the representations of visual and text modalities, respectively. Moreover, to mitigate the semantic gap between visual and text modalities of captioning, we propose the Mutual-Modality Concept Alignment (MC-Align) to mutually coordinate the encoded features with surgical concept representations of the other modality. In this way, the proposed SCA-Net achieves the surgical concept alignment between visual and text modalities, thereby producing more accurate captions with aligned multi-modal knowledge. Extensive experiments on neurosurgery videos and nephrectomy images confirm the effectiveness of our SCA-Net, which outperforms the state-of-the-arts by a large margin. The source code is available at https://github.com/franciszchen/SCA-Net.

Keywords: Neurosurgery · Video caption · Surgical concept

1 Introduction

Automatic surgical video captioning is critical to understanding the surgery with complicated operations, and can produce the natural language description

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 24–34, 2023.
https://doi.org/10.1007/978-3-031-43996-4_3

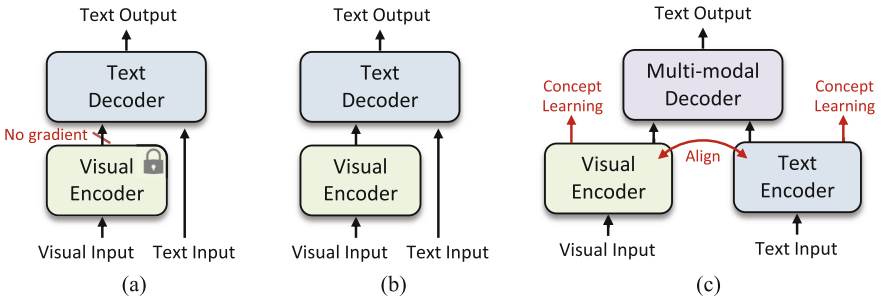


Fig. 1. Differences between existing captioning works (a) and (b), and our method (c). Different from directly mapping from visual input to text output in (a) and (b), we perform the surgical concept learning of two modality-specific encoders and mutually align two modalities for better multi-modal decoding.

with given surgical videos [24,26]. In this way, these techniques can reduce the workload of surgeons with multiple applications, such as providing the intra-operative surgical guidance [17], generating the post-operative surgical report [4], and even training junior surgeons [8].

To generate text descriptions from input videos, existing captioning works [5,9,21,24,26] mostly consist of a visual encoder for visual representations and a text decoder for text generation. Some early works [9,11,21,23] adopted a fixed object detector as the visual encoder to capture object representations for text decoding. This paradigm in Fig. 1(a) requires auxiliary annotations (*e.g.*, bounding box) to pre-train the visual encoder, and cannot adequately train the entire network for captioning. To improve performance with high efficiency in practice, recent works [13,24,25] followed the detector-free strategy, and opened up the joint optimization of visual encoder and text decoder towards captioning, as shown in Fig. 1(b). Despite great progress in this field, these works can be further improved with two limitations of surgical video captioning.

First, existing surgical captioning works [23,24,26] did not fully consider the inherent patterns of surgery to facilitate captioning. Due to the variability of lesions and surgical operations, surgical videos contain complex visual contents, and thus it is difficult to directly learn the mapping from the visual input to the text output. In fact, the same type of surgery has relatively fixed semantic patterns, such as using specific surgical instruments for a certain surgical action. Therefore, we introduce the **surgical concepts** (*e.g.*, surgical instruments, operated targets and surgical actions) from a semantic perspective, and guide the surgical captioning network to perceive these surgical concepts in the input video to generate more accurate surgical descriptions. Second, existing studies [9,24,26] simply processed visual and text modalities in sequential, while ignoring the semantic gap between these two modalities. This restricts the integration of visual and text modality knowledge, thereby damaging the captioning performance. Considering that both visual and text modalities revolve around the same set of surgical concepts, we aim to align the features in the visual and

text modalities with each other through surgical concepts, and achieve more efficient multi-modal fusion for accurate text predictions.

To address these two limitations in surgical video captioning, we propose the Surgical Concept Alignment Network (SCA-Net) to bridge the visual and text modalities through the surgical concepts, as illustrated in Fig. 1(c). Specifically, to enable the SCA-Net to accurately perceive surgical concepts, we first devise the Surgical Concept Learning (SCL) to predict the presence of surgical concepts with the representations of visual and text modalities, respectively. Moreover, to mitigate the semantic gap between visual and text modalities of captioning, we propose the Mutual-Modality Concept Alignment (MC-Align) to mutually coordinate the encoded features with surgical concept representations of the other modality. In this way, the proposed SCA-Net achieves the surgical concept alignment between visual and text modalities, thereby producing more accurate captions with aligned multi-modal knowledge. To the best of our knowledge, this work represents the first effort to introduce the surgical concepts for the surgical video captioning. Extensive experiments are performed on neurosurgery video and nephrectomy image datasets, and demonstrate the effectiveness of our SCA-Net by remarkably outperforming the state-of-the-art captioning works.

2 Surgical Concept Alignment Network

2.1 Overview of SCA-Net

As illustrated in Fig. 2, the Surgical Concept Alignment Network (SCA-Net) follows the advanced captioning architecture [25], and consists of visual and text encoders, and a multi-modal decoder. We implement the visual encoder with VideoSwin [15] to capture the discriminative spatial and temporal representations from input videos, and utilize the Vision Transformer (ViT) [7] with causal mask [6] as the text encoder to exploit text semantics with merely previous text tokens. The multi-modal decoder with ViT structure takes both visual and text tokens as input, and finally generates the caption of the input video. Moreover, to accurately perceive surgical concepts in SCL (Sect. 2.2), the SCA-Net learns from surgical concept labels using separate projection heads after the visual and text encoders. In the MC-Align (Sect. 2.3), the visual and text tokens from two encoders are mutually aligned with the concept representations of the other modality for better multi-modal decoding.

2.2 Surgical Concept Learning

Previous surgical captioning works [23,24,26] generated surgical descriptions directly from input surgical videos. Considering the variability of lesions and surgical operations, these methods may struggle to understand complex visual contents and generate erroneous surgical descriptions, thereby hindering performance to meet clinical requirements. In fact, both the surgical video and surgical caption represent the same surgical semantics in different modalities. Therefore,

we decompose surgical operations into surgical concepts, and guide these two modalities to accurately perceive the presence of surgical concepts, so as to better complete this cross-modal task.

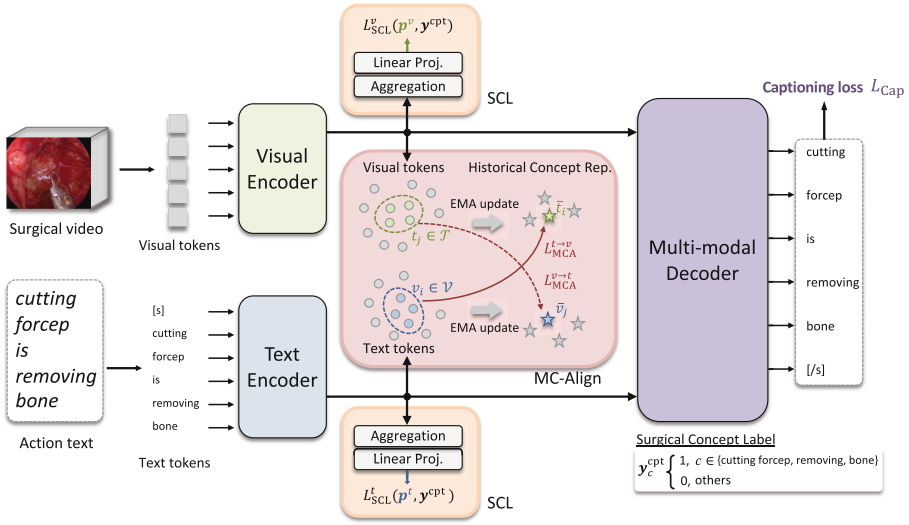


Fig. 2. Surgical Concept Alignment Network (SCA-Net) includes visual and text encoders, and a multi-modal decoder. The SCL supervises two encoders with projection heads by surgical concept labels, and the MC-Align mutually coordinates two modalities with concept representations for better multi-modal decoding.

Given a type of surgery, we regard the surgical instruments, surgical actions and the operated targets used in surgical videos as surgical concepts. Considering that both the visual input and the shifted text input contain the same set of surgical concepts, we find out which surgical concepts appear in the input surgical video by parsing the caption label. In this way, the presence of surgical concepts can be represented in a multi-hot surgical concept label $\boldsymbol{y}^{\text{cpt}} \in \{0,1\}^C$, where the surgical concepts that appear in the video are marked as 1 and the rest are marked as 0, and C is the number of possible surgical concepts. For example, the surgical video in Fig. 2 contains the instrument *cutting forcep*, the action *removing* and the target *bone*, and thus the surgical concept label $\boldsymbol{y}^{\text{cpt}}$ represents these surgical concepts in corresponding dimensions.

To guide the visual modality to perceive surgical concepts, we aggregate visual tokens generated by the visual encoder in average, and add a linear layer to predict the surgical concepts of input videos, where the normalized output $\boldsymbol{p}^v \in [0,1]^C$ estimates the probability of each surgical concept. We perform the multi-label classification using binary sigmoid cross-entropy loss, as follows:

$$L^v_{\text{SCL}} = -\sum_{c=1}^{C} \boldsymbol{y}^{\text{cpt}}_c \log\left(\boldsymbol{p}^v_c\right) + \left(1-\boldsymbol{y}^{\text{cpt}}_c\right) \log\left(1-\boldsymbol{p}^v_c\right). \tag{1}$$

In this way, the visual tokens are supervised to contain discriminative semantics related to valid surgical concepts, which can reduce prediction errors in surgical descriptions. For the text modality, we also perform SCL for surgical concept prediction $\boldsymbol{p}_c^t \in [0,1]^C$ and calculate the loss L_{SCL}^t in the same way. By optimizing $L_{\text{SCL}} = L_{\text{SCL}}^v + L_{\text{SCL}}^t$, the SCL enables visual and text encoders to exploit multi-modal features with the perception of surgical concepts, thereby facilitating the SCA-Net towards the captioning task.

2.3 Mutual-Modality Concept Alignment

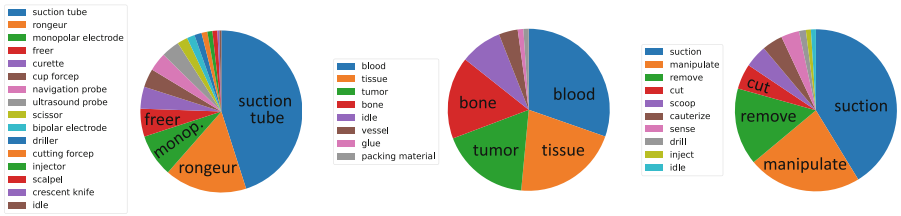
With the help of SCL in Sect. 2.2, our SCA-Net can perceive the shared set of surgical concepts in both visual and text modalities. However, given the differences between two modalities with separate encoders, it is inappropriate for the decoder to directly explore the cross-modal relationship between visual and text tokens [24,26]. To mitigate the semantic gap of two modalities, we devise the MC-Align to bridge these tokens in different modalities through surgical concept representations for better multi-modal decoding, as shown in Fig. 2.

To align these two modalities, we first collect surgical concept representations for each modality. Note that text tokens are separable for surgical concepts, while visual tokens are part of the input video containing multiple surgical concepts. For text modality, we parse the label of each text token and average text tokens of each surgical concept as $\boldsymbol{t}_c$, and update the historical text concept representations $\{\bar{\boldsymbol{t}}_c\}_{c=1}^C$ using Exponential Moving Average (EMA), as $\bar{\boldsymbol{t}}_c \leftarrow \gamma\bar{\boldsymbol{t}}_c + (1-\gamma)\boldsymbol{t}_c$, where the coefficient γ controls the updating for stable training and is empirically set as 0.9. For visual modality, we average visual tokens as the representation of each surgical concept present in the input video (*i.e.*, $\boldsymbol{v}_c$ if surgical concept label $\boldsymbol{y}_c^{\text{cpt}} = 1$), and update the historical visual concept representations $\{\bar{\boldsymbol{v}}_c\}_{c=1}^C$ with EMA, as $\bar{\boldsymbol{v}}_c \leftarrow \gamma\bar{\boldsymbol{v}}_c + (1-\gamma)\boldsymbol{v}_c$. In this way, we obtain the text and visual concept representations with tailored strategies for the alignment.

Then, we mutually align visual and text concept representations with corresponding historical ones in another modality. For visual-to-text alignment, visual concept representations are expected to be similar to corresponding text concept representations, while differing from other text concept representations as possible. Thus, we calculate the alignment objective $L_{\text{MCA}}^{v\to t}$ with regard to surgical concepts [10], and the visual encoder can be optimized with the gradients of visual concept representations in backward, thereby gradually aligning visual modality to text modality. Similarly, text concept representations are also aligned to the historical visual ones, as text-to-visual alignment $L_{\text{MCA}}^{t\to v}$. The MC-Align is summarized as follows:

$$L_{\text{MCA}} = \underbrace{-\sum_{\boldsymbol{v}_i\in\mathcal{V}} \log \frac{\exp\left(\boldsymbol{v}_i \cdot \bar{\boldsymbol{t}}_i\right)}{\sum_{c=1}^C \exp\left(\boldsymbol{v}_i \cdot \bar{\boldsymbol{t}}_c\right)}}_{L_{\text{MCA}}^{v\to t}} \underbrace{-\sum_{\boldsymbol{t}_j\in\mathcal{T}} \log \frac{\exp\left(\boldsymbol{t}_j \cdot \bar{\boldsymbol{v}}_j\right)}{\sum_{c=1}^C \exp\left(\boldsymbol{t}_j \cdot \bar{\boldsymbol{v}}_c\right)}}_{L_{\text{MCA}}^{t\to v}}, \tag{2}$$

where $\mathcal{V}$ and $\mathcal{T}$ denote all visual and text representations respectively, and $\cdot$ is the inner product of vectors. In this way, the MC-Align aligns visual and text representations with each other modality according to the surgical concept, thus benefiting multi-modal decoding for captioning.



(a) Surgical instruments (b) Operated targets (c) Surgical actions

Fig. 3. Surgical concepts and proportions in neurosurgery video captioning dataset.

2.4 Optimization

For the surgical captioning task, we adopt standard captioning loss L_{Cap} to optimize the cross-entropy of each predicted word based on previous words $y_{<t}$ and input video x, as follows:

$$L_{\text{Cap}} = -\sum_{t=1}^{T} \log p\left(y_t \mid y_{<t}, x\right), \tag{3}$$

where T is the length of caption prediction. Overall, the final objective of SCA-Net is summarized as $L = L_{\text{Cap}} + \lambda_1 L_{\text{SCL}} + \lambda_2 L_{\text{MCA}}$, where loss coefficients λ_1 and λ_2 control the trade-off of SCL and MC-Align. By optimizing this final objective L, the proposed SCA-Net can achieve multi-modal concept alignment, and generate superior descriptions for the surgical video captioning.

3 Experiment

3.1 Dataset and Implementation Details

Neurosurgery Video Captioning Dataset. To evaluate the effectiveness of surgical video captioning, we collect a large-scale dataset with 41 surgical videos of endonasal skull base neurosurgery. These surgical videos are recorded at the Prince of Wales Hospital, Chinese University of Hong Kong, where surgeons remove pituitary tumors through the endonasal corridor to the skull base. After necessary data cleaning, we divide these surgical videos with resolution of $1,920 \times 1,080$ into $11,004$ thirty-second video clips with clear surgical purposes. These video clips are annotated under Tool-Tissue Interaction (TTI) principle [18], and include a total of 16 instruments, 8 targets, and 10 surgical actions. The annotation preprocessing follows [26] using NLTK [16] toolkit. The proportion of surgical concepts is illustrated in Fig. 3. We split these video clips at patient-level, where the video clips of 31 patients are used for training and the rest of 10 patients are utilized for test.

EndoVis Image Captioning Dataset. We further compare our method with state-of-the-arts on the public EndoVis-2018 Image Captioning Dataset [1,23].

Table 1. Comparison on neurosurgery video captioning dataset. Best and second best results are **highlighted** and underlined.

Method	BLEU@4	METEOR	SPICE	ROUGE	CIDEr
VideoSwin + Self-Seq [21]	34.4	24.8	39.7	53.5	183.8
VideoSwin + AOANet [9]	41.5	29.7	46.5	58.0	288.1
SIG-Former [26]	36.2	30.1	35.8	52.0	181.7
SwinMLP-TranCAP [24]	39.8	28.7	39.2	51.9	195.9
M^2Transformer [5]	43.2	30.9	46.6	57.8	317.8
Ours *w/o* SCL, MC-Align	40.3	29.6	46.4	55.7	279.8
Ours *w/o* MC-Align	44.3	32.4	52.6	61.8	298.9
Ours *w/o* SCL	45.8	32.9	53.2	62.7	325.1
Ours	**48.1**	**35.1**	**56.1**	**64.9**	**368.4**

This dataset reveals robotic nephrectomy procedures acquired by the da Vinci X or Xi system, and is annotated with surgical actions between 9 possible tools and surgical targets [23]. We follow the official split in [24] with 11 sequences for training and 3 sequences for test. In this way, these two datasets can comprehensively evaluate the captioning tasks under both surgical videos and images.

Implementation Details. We implement our SCA-Net and state-of-the-art captioning methods [5,9,21,24,26] in PyTorch [20]. We optimize the SCA-Net and compared captioning methods using Adam with the batch size of 12 for both captioning datasets. All models are trained for 20 and 50 epochs in neurosurgery and EndoVis datasets, respectively. We adopt the step-wise learning rate decay strategy to facilitate training convergence, where the learning rate is initialized as 1×10^{-2} and halved after every 5 epochs. The loss coefficients λ_1 of L_{SCL} and λ_2 of of L_{MCA} are empirically set to 0.1 and 0.01, respectively. All experiments are performed on a single NVIDIA A100 GPU.

Evaluation Metrics. To evaluate the captioning performance, we adopt standard metrics, including BLEU@4 [19], METEOR [3], SPICE [2], ROUGE [12] and CIDEr [22]. Specifically, BLEU@4 [19] evaluates the 4-gram precision of the predicted caption, and CIDEr [22] is based on the n-gram similarity with TF-IDF weights. METEOR [3] considers both precision and recall. ROUGE [12] and SPICE [2] measure the matching between predictions and ground truth. The higher scores of these metrics indicate better performance in surgical captioning.

3.2 Comparison on Neurosurgery Video Captioning

To evaluate the performance of our SCA-Net, we perform a comprehensive comparison with the state-of-the-art captioning methods, including Self-Seq [21], AOANet [9], SIG-Former [26], M^2Transformer [5], and SwinMLP-TranCAP [24]. As illustrated in Table 1, our SCA-Net achieves the best performance, with the overwhelming BLEU@4 of 48.1%, METEOR of 35.1% and CIDEr of 368.4%.

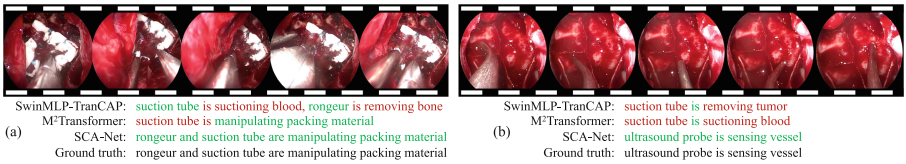


Fig. 4. The qualitative comparison between our SCA-Net and state-of-the-arts. With surgical concept alignment, our SCA-Net generates more accurate surgical descriptions.

Noticeably, our SCA-Net outperforms the surgical captioning work, SwinMLP-TranCAP [24], by a large margin, *e.g.*, 16.9% in SPICE and 13.0% in ROUGE. This advantage confirms that the proposed surgical concept alignment can alleviate the modalities gap in surgical captioning. Moreover, compared with the second-best M^{2}Transformer [5] with meshed attention between the visual encoder and the text decoder, our SCA-Net obtains superior performance with a remarkable increase of 9.5% in SPICE and 7.1% in ROUGE. These experimental results demonstrate the performance advantage of our SCA-Net over state-of-the-arts in the neurosurgery video captioning.

Ablation Study. To further validate the effectiveness of SCL and MC-Align, we perform the detailed ablation study in Table 1. Specifically, we implement three ablative baselines of the proposed SCA-Net, by removing the MC-Align (denoted as *w/o* MC-Align) and the SCL (denoted as *w/o* SCL) individually, as well as removing both (denoted as *w/o* SCL, MC-Align). As illustrated in Table 1, the proposed SCL and MC-Align can bring an individual improvement of 4.0% and 5.5% in BLEU@4, respectively, to the baseline of 40.3%. Furthermore, the SCL and MC-Align can work together to facilitate the captioning, with a BLEU@4 gain of 7.8%. These ablation experiments confirm that the proposed SCL and MC-Align play an important role in solving the modality gap in surgical video captioning, resulting in the performance advantage of our SCA-Net.

Qualitative Analysis. We present qualitative results of our SCA-Net and state-of-the-arts [5,24] on neurosurgery video captioning. In Fig. 4(a), SwinMLP-TranCAP [24] and M^{2}Transformer [5] incorrectly predict the operated targets and ignore important surgical instruments, respectively, and both methods [5,24] cannot recognize the rare instrument *ultrasound probe* as well as the corresponding surgical action in Fig. 4(b). With the help of surgical concept alignment, our SCA-Net can perceive the surgical concepts present in the surgical videos and thus generate correct descriptions in these two complex videos.

3.3 Comparison on EndoVis Image Captioning

To further confirm the effectiveness of surgical captioning, we perform the comparison on the public EndoVis image captioning dataset. As shown in Table 2, the end-to-end captioning methods [24,26] outperform the detector-based works using instrument bounding box as auxiliary annotations [9,21], by optimizing

the visual encoder to meet the requirement of the captioning task. In particular, our SCA-Net with Swin Transformer [14] as visual encoder achieves the best performance of four metrics (*e.g.*, 47.6% in BLEU@4 and 58.4% in SPICE), and outperforms the surgical state-of-the-art [24] with the advantage of 7.3% in BLEU@4 and 5.1% in METEOR. These comparisons confirm that our SCA-Net with surgical concept alignment can produce more accurate surgical captions.

Table 2. Comparison on EndoVis-2018 image captioning dataset. Best and second best results are **highlighted** and underlined.

Method	Aux. Anno	BLEU@4	METEOR	SPICE	CIDEr
FasterRCNN + Self-seq [21]	✓	29.5	28.3	49.6	180.1
FasterRCNN + AOANet [9]	✓	37.7	32.4	<u>58.0</u>	181.1
SIG-Former [26]	✗	42.6	<u>33.5</u>	52.4	<u>282.6</u>
SwinMLP-TranCAP [24]	✗	40.3	31.3	54.7	250.4
M^2Transformer [5]	✗	<u>43.0</u>	32.5	55.3	245.2
Ours	✗	**47.6**	**36.4**	**58.4**	**300.8**

4 Conclusion

To achieve accurate surgical video captioning, we propose the SCA-Net to mitigate the semantic gap of visual and text modalities with surgical concepts. Specifically, we devise the SCL to enable the SCA-Net with the perception of surgical concepts in visual and text modalities, respectively. Moreover, we propose the MC-Align to mutually coordinate visual and text representations with surgical concept representations of the other modality for multi-modal decoding, thereby generating more accurate captions with aligned multi-modal knowledge. Extensive experiments on neurosurgery and nephrectomy datasets confirm the advantage of our SCA-Net over state-of-the-arts on the surgical captioning.

Acknowledgments. This work is supported by National Key R&D Program of China under Grant No. 2021YFE0205700, National Natural Science Foundation of China (No. 62276260, 62076235, 62176254, 61976210, 62002356, 62006230), sponsored by Zhejiang Lab (No. 2021KH0AB07) and the InnoHK program.

References

1. Allan, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Anderson, P., Fernando, B., Johnson, M., Gould, S.: SPICE: semantic propositional image caption evaluation. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9909, pp. 382–398. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46454-1_24

3. Banerjee, S., Lavie, A.: METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop, pp. 65–72 (2005)
4. Bieck, R., et al.: Generation of surgical reports using keyword-augmented next sequence prediction. Curr. Direct. Biomed. Eng. **7**(2), 387–390 (2021)
5. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: CVPR, pp. 10578–10587 (2020)
6. Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N.: OperA: attention-regularized transformers for surgical phase recognition. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 604–614. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_58
7. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. In: ICLR (2021)
8. Elnikety, S., Badr, E., Abdelaal, A.: Surgical training fit for the future: the need for a change. Postgrad. Med. J. **98**(1165), 820–823 (2022)
9. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: ICCV, pp. 4634–4643 (2019)
10. Khosla, P., et al.: Supervised contrastive learning. In: NeurIPS, vol. 33, pp. 18661–18673 (2020)
11. Lin, C., Zheng, S., Liu, Z., Li, Y., Zhu, Z., Zhao, Y.: SGT: scene graph-guided transformer for surgical report generation. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 507–518. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_48
12. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
13. Lin, K., et al.: SwinBERT: end-to-end transformers with sparse attention for video captioning. In: CVPR, pp. 17949–17958 (2022)
14. Liu, Z., et al.: Swin transformer: hierarchical vision transformer using shifted windows. In: ICCV, pp. 10012–10022 (2021)
15. Liu, Z., et al.: Video swin transformer. In: CVPR, pp. 3202–3211 (2022)
16. Loper, E., Bird, S.: NLTK: the natural language toolkit. arXiv preprint cs/0205028 (2002)
17. Madani, A., et al.: Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. Ann. Surg. (2020)
18. Nwoye, C.I., et al.: CholecTriplet 2021: a benchmark challenge for surgical action triplet recognition. Med. Image Anal. **86**, 102803 (2023)
19. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: ACL, pp. 311–318 (2002)
20. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 (2019)
21. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR, pp. 7008–7024 (2017)
22. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: CIDEr: consensus-based image description evaluation. In: CVPR, pp. 4566–4575 (2015)
23. Xu, M., Islam, M., Lim, C.M., Ren, H.: Class-incremental domain adaptation with smoothing and calibration for surgical report generation. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 269–278. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_26

24. Xu, M., Islam, M., Ren, H.: Rethinking surgical captioning: end-to-end window-based MLP transformer using patches. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) MICCAI 2022. LNCS, vol. 13437, pp. 376–386. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_36
25. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: CoCa: contrastive captioners are image-text foundation models. Trans. Mach. Learn. Res. (2022)
26. Zhang, J., Nie, Y., Chang, J., Zhang, J.J.: Surgical instruction generation with transformers. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 290–299. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_28