# Surgical-VQA: Visual Question Answering in Surgical Scenes Using Transformer

Lalithkumar Seenivasan[1], Mobarakol Islam[2], Adithya K Krishna[3], and Hongliang Ren[1,4,5(✉)]

[1] Department of Biomedical Engineering, National University of Singapore, Singapore, Singapore
lalithkumar_s@u.nus.edu, ren@nus.edu.sg

[2] Biomedical Image Analysis Group, Imperial College London, London, UK
m.islam20@imperial.ac.uk

[3] Department of ECE, National Institute of Technology, Tiruchirappalli, India
108118004@nitt.edu

[4] Department of Electronic Engineering, Chinese University of Hong Kong, Hong Kong, Hong Kong
hlren@ee.cuhk.edu.hk

[5] Shun Hing Institute of Advanced Engineering, Chinese University of Hong Kong, Hong Kong, Hong Kong

**Abstract.** Visual question answering (VQA) in surgery is largely unexplored. Expert surgeons are scarce and are often overloaded with clinical and academic workloads. This overload often limits their time answering questionnaires from patients, medical students or junior residents related to surgical procedures. At times, students and junior residents also refrain from asking too many questions during classes to reduce disruption. While computer-aided simulators and recording of past surgical procedures have been made available for them to observe and improve their skills, they still hugely rely on medical experts to answer their questions. Having a Surgical-VQA system as a reliable 'second opinion' could act as a backup and ease the load on the medical experts in answering these questions. The lack of annotated medical data and the presence of domain-specific terms has limited the exploration of VQA for surgical procedures. In this work, we design a Surgical-VQA task that answers questionnaires on surgical procedures based on the surgical scene. Extending the MICCAI endoscopic vision challenge 2018 dataset and workflow recognition dataset further, we introduce two Surgical-VQA datasets with classification and sentence-based answers. To perform Surgical-VQA, we employ vision-text transformers models. We further introduce a residual MLP-based VisualBert encoder model that enforces interaction between visual and text tokens, improving performance in classification-based answering. Furthermore, we study the influence of the number of input image patches and temporal visual features on the model performance in both classification and sentence-based answering.

## 1   Introduction

Lack of medical domain-specific knowledge has left many patients, medical students and junior residents with questions lingering in their minds about medical diagnosis and surgical procedures. Many of these questions are left unanswered either because they assume these questions to be thoughtless, or students and junior residents refrain from raising too many questions to limit disruptions in lectures. The chances for them finding a medical expert to clarify their doubts are also slim due to the scarce number of medical experts who are often overloaded with clinical and academic works [6]. To assist students in sharpening their skills in surgical procedures, many computer-assisted techniques [2,17] and simulators [14,18] have been proposed. Although the systems assist in improving their skills and help reduce the workloads on academic professionals, the systems don't attempt to answer the student's doubts. While students have also been known to learn by watching recorded surgical procedures, the tasks of answering their questions still fall upon the medical experts. In such cases, a computer-assisted system that can process both the medical data and the questionnaires and provide a reliable answer would greatly benefit the students and reduce the medical expert's workload [20]. Surgical scenes are enriched with information that the system can exploit to answer questionnaires related to the defective tissue, surgical tool interaction and surgical procedures.

With the potential to extract diverse information from a single visual feature just by varying the question, the computer vision domain has seen a recent influx of vision and natural language processing models for visual question answering (VQA) tasks [15,23,27]. These models are either built based on the long short-term memory (LSTM) [5,21] or attention modules [20,22,29]. In comparison to the computer vision domain, which is often complemented with massive annotated datasets, the medical domain suffers from the lack of annotated data, limiting the exploration of medical VQA. The presence of domain-specific medical terms also limits the use of transfer learning techniques to adapt pre-trained computer-vision VQA models for medical applications. While limited works have been recently reported on medical-VQA [20] for medical diagnosis, VQA for surgical scenes remains largely unexplored.

In this work, **(i)** we design a Surgical-VQA task to generate answers for questions related to surgical tools, their interaction with tissue and surgical procedures (Fig. 1). **(ii)** We exploit the surgical scene segmentation dataset from
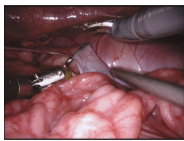


| Questions | Answers |
|---|---|
| **Q1:** what organ is being operated? | **A1:** kidney |
| **Q2:** what is the state of bipolar forceps? | **A2:** retraction |
| **Q3:** what is the state of suction? | **A3:** tissue manipulation |
| **Q4:** where is bipolar forceps located? | **A4:** bipolar forceps is located at left-top |
| **Q5:** what tools are operating the organ? | **A5:** the tools operating are bipolar forceps, monopolar curved scissors, and suction |

**Fig. 1.** Surgical-VQA: Given a surgical scene, the model predicts answer related to surgical tools, their interactions and surgical procedures based on the questionnaires.

the MICCAI endoscopic vision challenge 2018 (EndoVis-18) [3] and workflow recognition challenge dataset (Cholec80) [25], and extend it further to introduce two novel datasets for Surgical-VQA tasks. **(iii)** We employ two vision-text attention-based transformer models to perform classification-based and sentence-based answering for the Surgical-VQA. **(iv)** We also introduce a residual MLP (ResMLP) based VisualBERT ResMLP encoder model that outperforms VisualBERT [15] in classification-based VQA. Inspired by ResMLP [24], cross-token and cross-channel sub-modules are introduced into the VisualBERT ResMLP model to enforce interaction among input visual and text tokens. **(v)** Finally, the effects on the model's performance due to the varied number of input image patches and inclusion of temporal visual features are also studied.

## 2   Proposed Method

### 2.1   Preliminaries

**VisualBERT** [15]**:** A multi-layer transformer encoder model that integrates BERT [9] transformer model with object proposal models to perform vision-and-language tasks. BERT [9] model primarily processes an input sentence as a series of tokens (subwords) for natural language processing. By mapping to a set of embeddings ($E$), each word token is embedded ($e \in E$) based on token embedding $e_t$, segment embedding $e_s$ and position embedding $e_p$. Along with these input word tokens, VisualBERT [15] model processes visual inputs as unordered visual tokens that are generated using the visual features extracted from the object proposals. In addition to text embedding from BERT [9], it performs visual embedding ($F$), where, each visual token is embedded ($f \in F$) based on visual features $f_o$, segment embedding $f_s$ and position embedding $f_p$. Both text and visual embedding are then propagated through multiple encoder layers in the VisualBERT [15] to allow rich interactions between both the text and visual tokens and establish joint representation. Each encoder layer consists of an (i) self-attention module that establishes relations between tokens, (ii) intermediate module and (iii) output module consisting of hidden linear layers to reason across channels. Finally, the encoder layer is followed by a pooler module.

### 2.2   VisualBERT ResMLP

In our proposed VisualBERT ResMLP encoder model, we aim to further boost the interaction between the input tokens for vision-and-language tasks. The VisualBERT [15] model relies primarily on its self-attention module in the encoder layers to establish dependency relationships and allow interactions among tokens. Inspired by residual MLP (ResMLP) [24], the intermediate and output modules of the VisualBERT [15] model are replaced by cross-token and cross-channel modules to further enforce interaction among tokens. In the cross-token module, the inputs word and visual tokens are transposed and propagated forward, allowing information exchange between tokens. The resultant is then transposed

back to allow per-token forward propagation in the cross-channel module. Both cross-token and cross-channel modules are followed by element-wise summation with a skip-connection (residual-connection), which are layer-normalized. The cross-token output $(X_{CT})$ and cross-channel output $(X_{CC})$ is theorised as:

$$X_{CT} = Norm(X_{SA} + (A((X_{SA})^T))^T) \tag{1}$$

$$X_{CC} = Norm(X_{CT} + C(GeLU(B(X_{CT})))) \tag{2}$$

where, $X_{SA}$ is the self-attention module output, $A$, $B$ and $C$ are the learnable linear layers, $GeLU$ is the GeLU activation function [12] and NORM is the layer-normalization.

### 2.3  VisualBert ResMLP for Classification

**Word and Visual Tokens:** Each question is converted to a series of word tokens generated using BERT tokenizer [9]. Here, the BERT tokenizer is custom trained on the dataset to include surgical domain-specific words. The visual tokens are generated using the final convolution layer features of the feature extractor. A ResNet18 [11] model pre-trained on ImageNet [8] is used as the feature extractor. While the VisualBERT [15] model uses the visual features extracted from object proposals, we bypass the need for object proposal networks by extracting the features from the entire image. By employing adaptive average pooling to the final convolution layer, the output shape (s) is resize to s = [batch size x $n$ x $n$ x 256], thereby, restricting the number of visual tokens
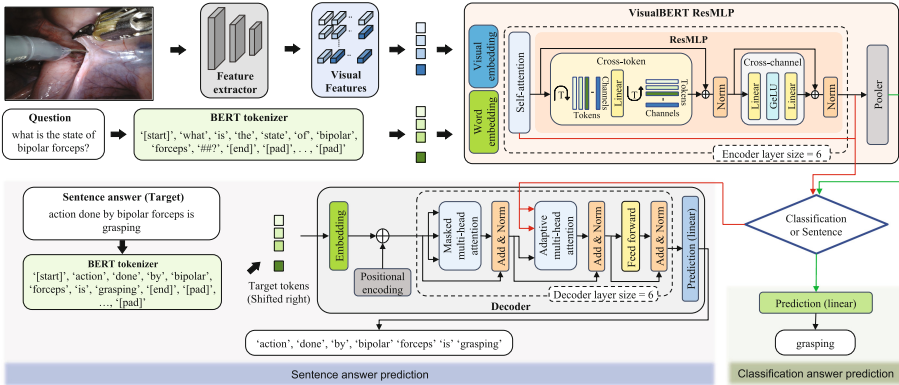


**Fig. 2.** Architecture: Given an input surgical scene and questions, its text and visual features are propagated through the vision-text encoder (VisualBERT ResMLP). **(i) classification-based answer:** The encoder output is propagated through a prediction layer for answer classification. **(ii) Sentence-based answer:** The encoder is combined with a transformer decoder to predict the answer sentence word-by-word (regressively).

(patches) to $n^2$. **VisualBERT ResMLP Module:** The word and visual tokens are propagated through text and visual embedding layers, respectively, in the VisualBert ResMLP model. The embedded tokens are then propagated through 6 layers of encoders (comprising self-attention and ResMLP modules) and finally through the pooling module. Embedding size = 300 and hidden layer feature size = 2048 are set as VisualBert ResMLP model parameters. **Classification-based answer:** The output of the pooling module is propagated through a prediction (linear) layer to predict the answer label.

### 2.4   VisualBert ResMLP for Sentence

**Word and Visual Tokens:** In addition to tokenizing the words in the questions and visual features in the image as stated in Sect. 2.3, the target sentence-based answer is also converted to a series of word tokens using the BERT tokenizer. **VisualBERT ResMLP encoder:** The VisualBERT ResMLP model follows the same parameters and configuration as stated in Sect. 2.3. **Sentence-based answer:** To generate the answer in a sentence structure, we propose to combine the vision-text encoder model (VisualBERT [15] or VisualBERT ResMLP) with a multi-head attention-based transformer decoder (TD) model. Considering the sentence-based answer generation task to be similar to an image-caption generation task [7], it is positioned as a next-word prediction task. Given the question features, visual features and the previous word (in answer sentence) token, the decoder model predicts the next word in the answer sentence. During the training stage, as shown in Fig. 2, the target word tokens are shifted rights and, together with vision-text encoder output, are propagated through the transformer decoder layers to predict the corresponding next word. During the evaluation stage, the beam search [28] technique is employed to predict the sentence-based answer without using target word tokens. Taking the '[start]' token as the first word, the visual-text encoder model and decoder is regressed on every predicted word to predict the subsequent word, until a '[end]' token is predicted.

## 3   Experiment

### 3.1   Dataset

**Med-VQA:** A public dataset from the ImageCLEF 2019 Med-VQA Challenge [1]. Three categories (C1: modality, C2: plane and C3: organ) of medical question-answer pairs from the dataset are used in this work. The C1, C2 and C3 pose a classification task (single-word answer) for 3825 images with a question. The C1, C2 and C3 consist of 45, 16 and 10 answer classes, respectively. The train and test set split follows the original implementation [1].

**EndoVis-18-VQA:** A novel dataset was generated from 14 video sequences of robotic nephrectomy procedures from the MICCAI Endoscopic Vision Challenge 2018 [3] dataset. Based on the tool, tissue, bounding box and interaction annotations used for tool-tissue interaction detection tasks [13], we generated two versions of question-answer pairs for each image frame. **(i) Classification (EndoVis-18-VQA (C)):** The answers are in single-word form. It consists of 26 distinct answer classes (one organ, 8 surgical tools, 13 tool interactions and 4 tool locations). **(ii) Sentence (EndoVis-18-VQA (S)):** The answers are in a sentence form. In both versions, 11 sequences with 1560 images and 9014 question-answer pairs are used as a training set and 3 sequences with 447 images and 2769 question-answers pairs are used as a test set. The train and test split follow the tool-tissue interaction detection task [19].

**Cholec80-VQA:** A novel dataset generated from 40 video sequences of the Cholec80 dataset [25]. We sampled the video sequences at 0.25 fps to generate the Cholec80-VQA dataset consisting of 21591 frames. Based on tool-operation and phase annotations provided in the Cholec80 dataset [25], two question-answer pairs are generated for each frame. **(i) Classification (Cholec80-VQA (C)):** 14 distinct single-word answers (8 on surgical phase, two on tool state and 4 on number of Tools). **(ii) Sentence (Cholec80-VQA (S)):** The answers are in a sentence form. In both versions, 34k question-answer pairs for 17k frames are used for the train set and $9K$ question-answer pairs for 4.5k frames are used for the test set. The train and test set split follows the Cholec80 dataset [25].

### 3.2 Implementation Details

Both our classification and the sentence-based answer generation models[1] are trained based on cross-entropy loss and optimized using the Adam optimizer. For Classification tasks, a batch size = 64 and epoch = 80 are used. A learning rate = $5 \times 10^{-6}$, $1 \times 10^{-5}$ and $5 \times 10^{-6}$ are used for Med-VQA [1], EndoVis-18-VQA (C) and Cholec80-VQA (C) dataset, respectively. For the sentence-based answer generation tasks, a batch size = 50 is used. The models are trained for epoch = 50, 100 and 51, with a learning rate = $1 \times 10^{-4}$, $5 \times 10^{-5}$ and $1 \times 10^{-6}$ on Med-VQA [1], EndoVis-18-VQA (S) and Cholec80-VQA (S) dataset, respectively.

## 4   Results

The performance of classification-based answering on the EndoVis-18-VQA (C), Cholec80-VQA (C) and Med-VQA (C1, C2 and C3) datasets are quantified based on the accuracy (Acc), recall and F-score in Table 1. It is observed that our proposed encoder (VisualBERT ResMLP) based model outperformed the current medical VQA state-of-the-art MedFuse [20] model and marginally outperformed the base encoder (VisualBERT [15]) based model in almost all datasets. While

---

[1] github.com/lalithjets/Surgical_VQA.git.

**Table 1.** Performance comparison of our VisualBERT ResMLP based model against MedFuse [20] and VisualBERT [15] based model for classification-based answering.

| Dataset | MedFuse [20] | | | VisualBert [15] | | | VisualBert ResMLP | | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc | Recall | Fscore | Acc | Recall | Fscore | Acc | Recall | Fscore |
| Med-VQA (C1) | 0.754 | 0.224 | 0.140 | **0.828** | **0.617** | **0.582** | **0.828** | 0.598 | 0.543 |
| Med-VQA (C2) | 0.730 | 0.305 | 0.303 | **0.760** | 0.363 | 0.367 | 0.758 | **0.399** | **0.398** |
| Med-VQA (C3) | 0.652 | 0.478 | 0.484 | 0.734 | 0.587 | 0.595 | **0.736** | **0.609** | **0.607** |
| EndoVis-18-VQA (C) | 0.609 | 0.261 | 0.222 | 0.619 | **0.412** | 0.334 | **0.632** | 0.396 | **0.336** |
| Cholec80-VQA (C) | 0.861 | 0.349 | 0.309 | 0.897 | **0.629** | 0.633 | **0.898** | 0.627 | **0.634** |

**Table 2.** k-fold performance comparison of our VisualBERT ResMLP based model against VisualBERT [15] based model on EndoVis-18-VQA (C) dataset.

| Model | $1^{st}$ Fold | | $2^{nd}$ Fold | | $3^{rd}$ Fold | |
|---|---|---|---|---|---|---|
| | Acc | Fscore | Acc | Fscore | Acc | Fscore |
| VisualBert [15] | 0.619 | 0.334 | 0.605 | 0.313 | 0.578 | 0.337 |
| VisualBert ResMLP | **0.632** | **0.336** | **0.649** | **0.347** | **0.585** | **0.373** |

the improvement in performance against the base model is marginal, a k-fold study (Table 2) on EndoVis-18-VQA (C) dataset proves that the improvement is consistent. Furthermore, our model (159.0M) requires 13.64% lesser parameters compared to the base model (184.2M).

For the sentence-based answering, BLEU score [16], CIDEr [26] and METEOR [4] are used for quantitative analysis. Compared to the LSTM-based MedFuse [20] model, the two variants of our proposed Transformer-based model

**Table 3.** Comparison of transformer-based models ((i) VisualBERT [15] + TD and (ii) VisualBERT ResMLP + TD) against MedFuse [20] for sentence-based answering.

| Model | EndoVis-18-VQA (S) | | | | Cholec80-VQA (S) | | | |
|---|---|---|---|---|---|---|---|---|
| | BLEU-3 | BLEU-4 | CIDEr | METEOR | BLEU-3 | BLEU-4 | CIDEr | METEOR |
| MedFuse [20] | 0.212 | 0.165 | 0.752 | 0.148 | 0.378 | 0.333 | 1.250 | 0.222 |
| VisualBert [15] + **TD** | **0.727** | **0.694** | 5.153 | **0.544** | **0.963** | **0.956** | **8.802** | **0.719** |
| VisualBert ResMLP + TD | 0.722 | 0.691 | **5.262** | 0.543 | 0.960 | 0.952 | 8.759 | 0.711 |
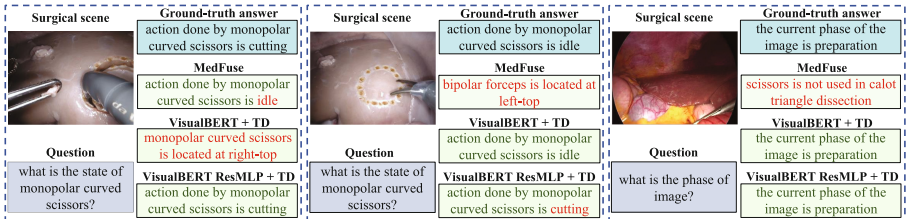


**Fig. 3.** Comparison of sentence-based answer generation by MedFuse [20], Visual-BERT [15] + TD and VisualBERT ResMLP + TD models.

(VisualBERT [15] + TD and VisualBERT ResMLP + TD) performed better on both the EndoVis-18-VQA (S) and Cholec80-VQA (S) dataset (Table 3). However, when compared within the two variants, the variant with the VisualBERT [15] as its vision-text encoder performed marginally better. While the cross-patch sub-module in VisualBERT ResMLP encoder improves performance in classification-based answering, the marginal low performance could be attributed to its influence on the self-attention sub-module. To predict a sentence-based answer, in addition to the encoder's overall output, the adaptive multi-head attention sub-module in the TD utilizes the encoder's self-attention sub-module outputs. By enforcing interaction between tokens, the cross-patch sub-module could have affected the optimal training of the self-attention sub-module, thereby affecting the sentence generation performance. It is worth noting that the VisualBERT ResMLP encoder + TD model (184.7M) requires 11.98% fewer parameters compared to the VisualBERT [15] + TD model (209.8M) while maintaining similar performances. Figure 3 shows the qualitative performance of sentence-based answering.

## 4.1   Ablation Study

Firstly, the performance of VisualBERT encoder and VisualBERT ResMLP encoder-based models for classification-based answering with a varying number of input image patches (patch size = 1, 4, 9, 16 and 25) is trained and studied. From Fig. 4 (a), it is observed that VisualBERT ResMLP encoder-based model generally performs better (in-line with observation in Table 1) than VisualBERT [15] encoder-based model, even with varied number of input patches. Secondly, from Fig. 4 (b), it is also observed that performances generally improved with an increase in the number of patches for classification-based
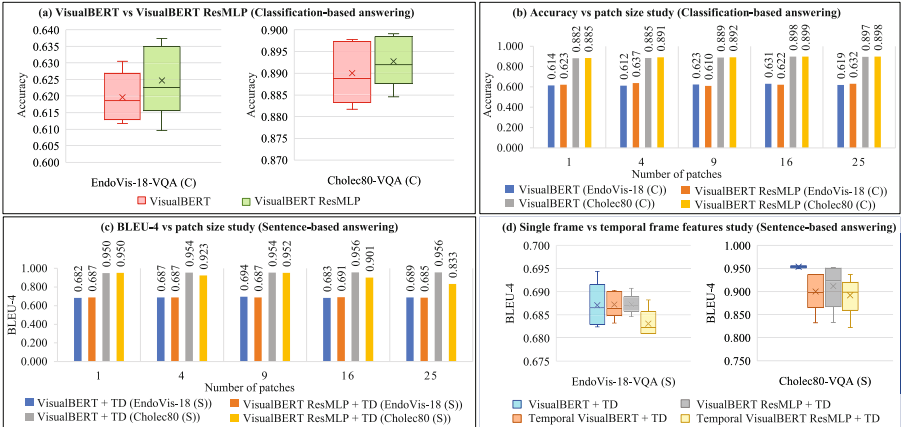


**Fig. 4.** Ablation study: For classification-based answering (a) Performance comparison of VisualBERT [15] vs VisualBERT ResMLP based model; (b) Accuracy vs patch size (number of patches) study. For sentence-based answering: (c) BLEU-4 vs patch size study; (d) Performance of single frame vs temporal visual features.

answering. However, the influence of the number of inputs patches on sentence-based answering remains inconclusive (Fig. 4 (c)). In most cases, the input of a single patch seems to offer the best or near best results.

Finally, the performance of models incorporated with temporal visual features is also studied. Here, a 3D ResNet50 pre-trained on the UCF101 dataset [10] is employed as the feature extractor. With a temporal size $= 3$, the current and the past 2 frames are used as input to the feature extractor. The extracted features are then used as visual tokens for sentence-based answering. Figure 4 (d) shows the model's performance on a varied number of input patches (patch size $= 1, 4, 9, 16$ and $25$) from single frames vs. temporal frames on EndoVis-18-VQA (S) and Cholec80-VQA (S). It is observed that both transformer-based models' performance reduces when the temporal features are used.

## 5  Discussion and Conclusion

We design a Surgical-VQA algorithm to answer questionnaires on surgical tools, their interactions and surgical procedures based on our two novel Surgical-VQA datasets evolved from two public datasets. To perform classification and sentence-based answering, vision-text attention-based transformer models are employed. A VisualBERT ResMLP transformer encoder model with lesser model parameters is also introduced that marginally outperforms the base vision-text attention encoder model by incorporating a cross-token sub-module. The influence of the number of input image patches and the inclusion of temporal visual features on the model's performance is also reported. While our Surgical-VQA task answers to less-complex questions, from the application standpoint, it unfolds the possibility of incorporating open-ended questions where the model could be trained to answer surgery-specific complex questionnaires. From the model standpoint, future work could focus on introducing an asynchronous training regime to incorporate the benefits of the cross-patch sub-module without affecting the self-attention sub-module in sentence-based answer-generation tasks.

## References

1. Abacha, A.B., Hasan, S.A., Datla, V.V., Liu, J., Demner-Fushman, D., Müller, H.: VQA-med: overview of the medical visual question answering task at imageclef 2019. clef2019 working notes. In: CEUR Workshop Proceedings, pp. 9–12. CEUR-WS.org <http://ceur-ws.org>. September

2. Adams, L., et al.: Computer-assisted surgery. IEEE Comput. Graph. Appl. **10**(3), 43–51 (1990)

3. Allan, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)

4. Banerjee, S., Lavie, A.: 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, pp. 65–72 (2005)

5. Barra, S., Bisogni, C., De Marsico, M., Ricciardi, S.: Visual question answering: which investigated applications? Pattern Recogn. Lett. **151**, 325–331 (2021)

6. Bates, D.W., Gawande, A.A.: Error in medicine: what have we learned? (2000)

7. Cornia, M., Stefanini, M., Baraldi, L., Cucchiara, R.: Meshed-memory transformer for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10578–10587 (2020)

8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. IEEE (2009)

9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

10. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3D residual networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, pp. 3154–3160 (2017)

11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

12. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)

13. Islam, M., Seenivasan, L., Ming, L.C., Ren, H.: Learning and reasoning with the graph structure representation in robotic surgery. In: Martel, A.L., et al. (eds.) MICCAI 2020. LNCS, vol. 12263, pp. 627–636. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-59716-0_60

14. Kneebone, R.: Simulation in surgical training: educational issues and practical implications. Med. Educ. **37**(3), 267–277 (2003)

15. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: a simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)

16. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)

17. Rogers, D.A., Yeh, K.A., Howdieshell, T.R.: Computer-assisted learning versus a lecture and feedback seminar for teaching a basic surgical technical skill. Am. J. Surg. **175**(6), 508–510 (1998)

18. Sarker, S., Patel, B.: Simulation and surgical training. Int. J. Clin. Pract. **61**(12), 2120–2125 (2007)

19. Seenivasan, L., Mitheran, S., Islam, M., Ren, H.: Global-reasoned multi-task learning model for surgical scene understanding. IEEE Robot. Autom. Lett. **7**, 3858–3865 (2022)

20. Sharma, D., Purushotham, S., Reddy, C.K.: Medfusenet: an attention-based multimodal deep learning model for visual question answering in the medical domain. Sci. Rep. **11**(1), 1–18 (2021)

21. Sharma, H., Jalal, A.S.: Image captioning improved visual question answering. Multimedia Tools Appl. 1–22 (2021)
22. Sharma, H., Jalal, A.S.: Visual question answering model based on graph neural network and contextual attention. Image Vis. Comput. **110**, 104165 (2021)
23. Sheng, S., et al.: Human-adversarial visual question answering. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
24. Touvron, H., et al.: Resmlp: feedforward networks for image classification with data-efficient training. arXiv preprint arXiv:2105.03404 (2021)
25. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE Trans. Med. Imaging **36**(1), 86–97 (2016)
26. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4566–4575 (2015)
27. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: SIMVLM: simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904 (2021)
28. Wiseman, S., Rush, A.M.: Sequence-to-sequence learning as beam-search optimization. arXiv preprint arXiv:1606.02960 (2016)
29. Zhang, S., Chen, M., Chen, J., Zou, F., Li, Y.F., Lu, P.: Multimodal feature-wise co-attention method for visual question answering. Inf. Fusion **73**, 1–10 (2021)