Original Article

SDA-CLIP: surgical visual domain adaptation using video and text labels

Yuchong Li[1,2], Shuangfu Jia[3], Guangbi Song[4], Ping Wang[5], Fucang Jia[1,2,6]

[1]Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China; [2]Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China; [3]Department of Operating Room, Hejian People's Hospital, Hejian, China; [4]Medical Imaging Center, Luoping County People's Hospital, Qujing, China; [5]Department of Hepatobiliary Surgery, The First Affiliated Hospital of Guangzhou Medical University, Guangzhou, China; [6]Pazhou Lab, Guangzhou, China

Contributions: (I) Conception and design: Y Li, F Jia; (II) Administrative support: G Song, P Wang, F Jia; (III) Provision of study materials or patients: S Jia, P Wang, F Jia; (IV) Collection and assembly of data: Y Li, S Jia, G Song, P Wang, F Jia; (V) Data analysis and interpretation: Y Li, F Jia; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Correspondence to: Fucang Jia, PhD. Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, 1068 Xueyan Avenue, Shenzhen 518055, China; Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, Shenzhen, China; Pazhou Lab, Guangzhou, China. Email: fc.jia@siat.ac.cn.

Background: Surgical action recognition is an essential technology in context-aware-based autonomous surgery, whereas the accuracy is limited by clinical dataset scale. Leveraging surgical videos from virtual reality (VR) simulations to research algorithms for the clinical domain application, also known as domain adaptation, can effectively reduce the cost of data acquisition and annotation, and protect patient privacy.

Methods: We introduced a surgical domain adaptation method based on the contrastive language-image pretraining model (SDA-CLIP) to recognize cross-domain surgical action. Specifically, we utilized the Vision Transformer (ViT) and Transformer to extract video and text embeddings, respectively. Text embedding was developed as a bridge between VR and clinical domains. Inter- and intra-modality loss functions were employed to enhance the consistency of embeddings of the same class. Further, we evaluated our method on the MICCAI 2020 EndoVis Challenge SurgVisDom dataset.

Results: Our SDA-CLIP achieved a weighted F1-score of 65.9% (+18.9%) on the hard domain adaptation task (trained only with VR data) and 84.4% (+4.4%) on the soft domain adaptation task (trained with VR and clinical-like data), which outperformed the first place team of the challenge by a significant margin.

Conclusions: The proposed SDA-CLIP model can effectively extract video scene information and textual semantic information, which greatly improves the performance of cross-domain surgical action recognition. The code is available at https://github.com/Lycus99/SDA-CLIP.

Keywords: Surgical domain adaptation; cross-domain surgical action recognition; video-text learning

Submitted Mar 23, 2023. Accepted for publication Aug 03, 2023. Published online Sep 21, 2023.
doi: 10.21037/qims-23-376
View this article at: https://dx.doi.org/10.21037/qims-23-376

Introduction

Since the emergence of the information technology era, the integration of big data into clinical surgical practice has become a prevalent trend (1). The data-driven context-aware-based autonomous surgical assistance system is the future development direction of the operating room (2), which will reduce the operating pressure of the surgeon and shorten the learning curve for newcomers. The system mentioned serves multiple purposes in the context of surgical procedures. It monitors surgical procedures (3), supports clinical decision-making (4), and provides early warnings during surgeries (2). Additionally, the system generates task-based performance reports, enabling the assessment of surgeons' surgical skills after the completion of surgeries (5,6). Moreover, the analysis results of surgical videos can be synchronized across multiple devices through cloud-based storage, analysis, and retrieval mechanisms (7).

Context-aware tasks such as surgical action recognition can be effectively learned from a multitude of manually annotated surgical videos. However, due to worries about privacy invasions, many people are reluctant to disclose their own health data (8). In addition, clinical surgical video annotations are challenging and time-consuming (9) since they require prior medical knowledge.

A new trend is to use videos from virtual reality (VR) simulations of surgical actions to develop algorithms to recognize tasks in a clinical-like setting (10), a technique known as surgical domain adaption. The utilization of simulated, harmless, and well-labeled surgical action videos is the major advantage of this approach. Domain adaptation tasks can be divided into hard domain adaptation and soft domain adaptation according to whether the target domain data is included in the training set. Specifically, hard domain adaptation refers to using only VR domain data for model training, which is also called unsupervised domain adaptation, and soft domain adaptation employs VR domain and few clinical domain data for model training.

Traditional surgical action recognition methods only leverage data from the clinical domain. However, these models' performance degrades a great deal in other data domains, which is called domain-shift (11). Many efforts have been made to address this issue by aligning the data distribution in the source and target domains through non-linear transformations (*Figure 1*). A representation of the typical distribution of original data in the source and target domains is provided in *Figure 1A*, previous surgical domain transforms in novel spaces are displayed in *Figure 1B*, wherein the category distributions of the source and target domains are consistent and contiguous. Image-to-image (I2I) (12) adopts a 2-step strategy, first using cycle-generative adversarial network (CycleGAN) to transfer simulated images into clinical-style images, and then harnessing these transformed images for training. The models presented in (10) aim to extract instrument edges or segment masks in surgical images. However, when the manifold structures of data distribution are complex, such transformations in pursuit of domain consistency can lead to a loss of semantic information and a reduction in model accuracy (13). Consequently, we explored the introduction of novel components to establish the relationship between the 2 data domains without transforming the original space.

We aimed to clarify whether there any components that are innately consistent across both domains. We proposed to connect the source and target domains with the help of text labels corresponding to surgical action videos. As shown in *Figure 1C*, the surgical action videos for both domains are linked to text labels by image-text contrastive learning. Compared to single-modality models that are solely trained on images, multi-modality models have the theoretical potential to learn more effective and discriminative latent space representations (14). In addition, the multi-modality features possess the capability to be transferred to other data domains, which is especially relevant in situations where data availability is scarce, such as in clinical surgical videos. With the blooming of contrastive learning, the contrastive language-image pretraining model (CLIP) (15) was proposed to learn multi-modality feature representations from image-text pairs. Additionally, it has been experimentally validated that the model is robust when processing images from various domains. The knowledge learned by the CLIP model can be successfully transferred to downstream tasks (16,17). Unlike mutually orthogonal one-hot labels, text labels encompass abundant semantic information. The similarity between text labels from various categories can be viewed as prior knowledge for assessing their relevance (18). Moreover, in the field of medical image analysis, although visual features of the same anatomical structure are domain-specific, its textual information can remain consistent. Consequently, Liu *et al.* (19) employed a text-based approach to construct uniform and robust models across datasets. Such findings have encouraged us to investigate the application of image-text models to the surgical domain adaptation task.

In this paper, we propose a novel framework for surgical domain adaptation (SDA; SDA-CLIP) to recognize cross-

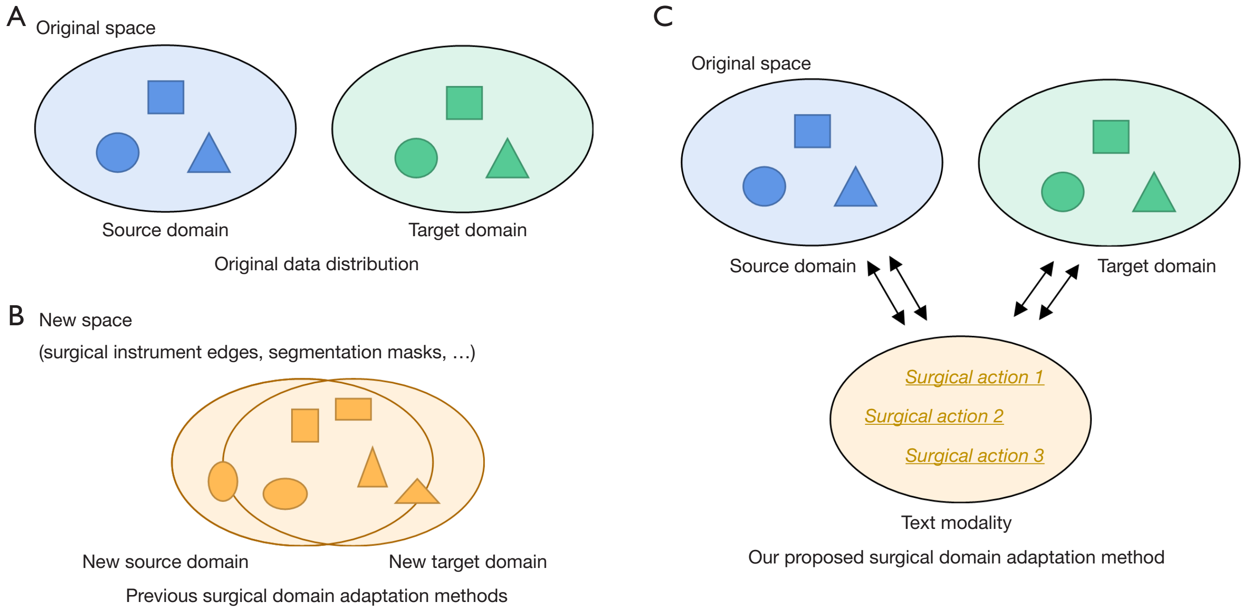


Figure 1 Different surgical domain adaptation methods. (A) The distribution of the original data in the source and target domains. The same shapes represent the same action categories. (B) The source and target domains are aligned to a new space by non-linear transformations to achieve domain-aligned data distributions. (C) The source and target domains are bridged by text labels. The original data distribution and semantic features are preserved.

domain surgical actions. Specifically, we first extended the original one-hot labels to text prompt labels by templates. After that, we input the training videos into the video encoder and input the text prompts into the text encoder to obtain the visual embeddings and text embeddings, respectively, and then the visual embeddings were fused through the fusion model. Finally, the surgical action recognition task was implemented by calculating the similarity between visual embeddings and text embeddings. When evaluating the model, test videos from another domain can be directly fed into the model without changing the model parameters or framework.

Our key contributions are as follows:

(I) We propose a novel framework that applies a video-text pair-based network to recognize cross-domain surgical actions, demonstrating the feasibility of using text information for surgical domain adaptation and video analysis.

(II) We introduce intra- and inter-modality loss functions to further constrain the learning of the network. The efficacy of different components is further analyzed.

(III) Our method achieves state-of-the-art performance on 2 tasks on the MICCAI 2020 EndoVis Challenge SurgVisDom dataset, greatly surpassing the winner's method in the challenge.

Methods

In this section, we illustrate different modules in the network, including video encoder, text encoder, loss functions, and other components. The overview of our proposed SDA-CLIP is shown in *Figure 2*.

Model architecture

For a mini-batch of size n, we use $V=[v_1,v_2,\cdots,v_n]$ to represent each video clip in it, and the corresponding labels are denoted by $C=[c_1,c_2,\cdots,c_n]$. The video clip v_i consists of L_i frames. In the video pre-process, we divide one video clip into s segments, and then randomly select a frame from each segment to constitute an image sequence representing the video clip, denoted as $\hat{v}_i=\left[x_1^i,x_2^i,\cdots,x_s^i\right]\in R^{s\times C\times W\times H}$.

Video encoder

In order to extract the discriminative embedding of the pictures in the video, we exploit the CLIP pre-trained

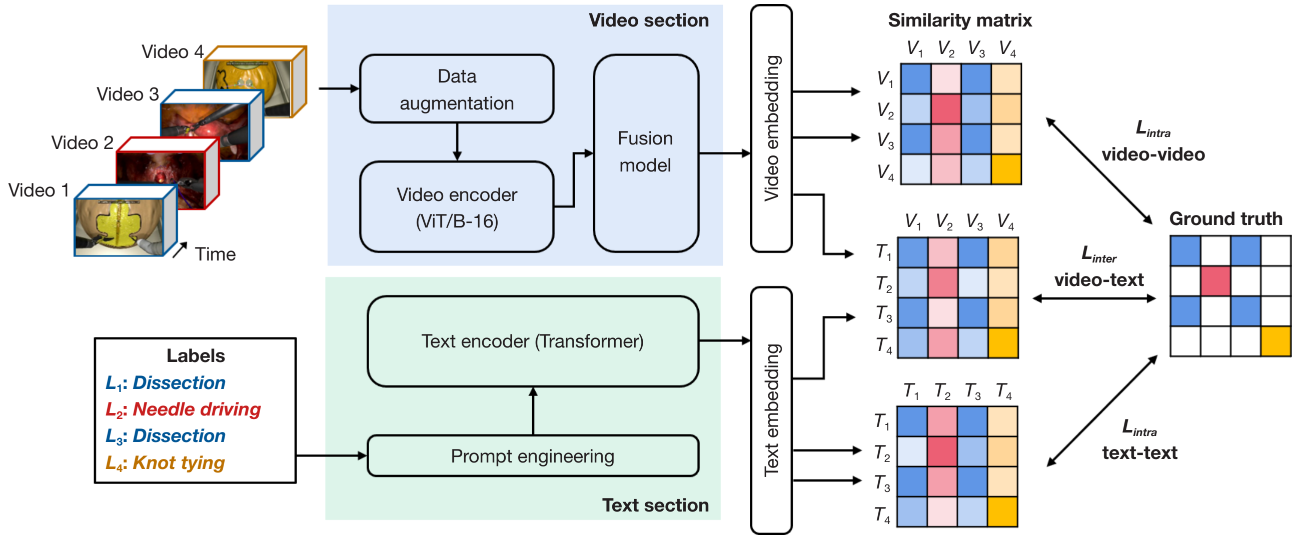


Figure 2 Overview of our proposed SDA-CLIP. Labels or videos of the same color belong to the same class, so the corresponding similarity coefficient in the ground truth matrix is 1. The white blocks indicate that the coefficient is 0. $V_{1\text{-}4}$, $T_{1\text{-}4}$ denotes video embedding of videos 1–4 and text embedding of labels 1-4, respectively. SDA-CLIP, surgical domain adaptation method based on the contrastive language-image pretraining model; ViT, Vision Transformer.

ViT (20) model with patch size 16 and 12 layers as the video backbone network. Each image in the sequence is discretized with a non-overlapping size of 16×16 to obtain a series of 1-dimensional (1D) tokens. After that, the [class] token is concatenated at the first position of the sequence and the position embedding is added to each token. Finally, the token sequence is input into encoder blocks to extract video embedding. In order to calculate the similarity coefficient with the text embedding, the video backbone network ends with a linear projection head and the embedding is projected to the dimension d_0 from d_v by matrix $D \in R^{d_v \times d_0}$. Let f_v represent the video encoder, then the output segment-level features of the segment $\hat{v}_l$ can be formulated as follows:

$$f_v\left(x_j^i\right) = \text{transformer-encoder}_{vid}\left(x_j^i\right) D \qquad [1]$$

$$f_{seg}^i = f_v\left(\hat{v}_i\right) = \left[f_v\left(x_1^i\right), f_v\left(x_2^i\right), \cdots, f_v\left(x_s^i\right)\right] \in R^{s \times d_0} \qquad [2]$$

Finally, $F_{seg} = \left[f_{seg}^1, f_{seg}^2, \cdots, f_{seg}^n\right]$ is utilized to represent the segment-level embedding of the mini-batch.

Fusion model

The segment-level features of a clip extracted by the video encoder are integrated into the video-level feature through the fusion model. A naïve thought is to average the segment-level features, which is also known as mean pooling:

$$f_{vid}^i = f_{mean}\left(f_{seg}^i\right) = \text{mean-pooling}\left(\left[f_v\left(x_1^i\right), f_v\left(x_2^i\right), \cdots, f_v\left(x_s^i\right)\right]\right) [3]$$

In addition, the Transformer proposed for modeling language sequence can be easily expanded to features fusion for image sequence, which can perform global modeling and fit more complex action changes.

$$f_{vid}^i = f_{trans}\left(f_{seg}^i\right) = \text{mean-pooling}\left(\text{transformer-encoder}_{fus}\left(f_{seg}^i\right)\right) [4]$$

From another perspective, the naïve mean-pooling strategy is equal to a 0-layer transformer. Our experiments demonstrate that the Transformer fusion model performs better than mean pooling. Let $F_{vid} = \left[f_{vid}^1, f_{vid}^2, \cdots, f_{vid}^n\right]$ represent the video-level embedding of a mini-batch, and the dimension is $R^{n \times d_0}$.

Text encoder

Prompt engineering is the process of expanding text labels for class names into prompts by templates in order to provide task information and prevent polysemy. We first expanded all category names into phrases with the same meaning, such as *Needle-Driving → driving the needle tip*. The expanded category names are easily embedded into the prompt templates. The design of the prompt templates has a significant impact on how well multi-modality models work (21). The prompt types we used are listed below:

Table 1 Different kinds of prompt templates. The curly brackets are used to contain the category name phrase

Template type	Template description
Caption template	{}
General template	a photo of {}, a picture of action {}, playing action of {}, doing a kind of action, {} playing a kind of action, {} can you recognize the action of {}, video classification of {}, a video of {}
Customized template	a photo of {}, a type of surgical action; surgical action of {}; {}, a surgical action; {}, this is a surgical action; {}, a video of surgical action; look, the surgeon is {}; the doctor is performing {}; the surgeon is performing {}

- Caption prompt: only contains the category name (e.g., {driving the needle tip});
- General prompt: does not contains task information (e.g., a photo of {driving the needle tip});
- Customized prompt: contains task information (e.g., a photo of {driving the needle tip}, a type of surgical action).

where the content in the curly brackets represents the category name phrase, and the other parts are the prompt templates.

In (15), it is mentioned that templates containing task information perform better than general templates. As a result, we utilized task-customized templates that included the words "surgeon" or "surgical action", which can offer task hints to the model. To attenuate the effect of template wording, we used 8 templates in both general prompt and customized prompt. The templates were programmatically randomly combined with the category names in each training epoch. Finally, 17 prompt templates were all used in our model (1 caption template, 8 general templates, and 8 customized templates). In the ablation study section, we compare the experimental results of different types of templates. A full list of prompt templates can be found in *Table 1*.

Let $\hat{C} = [y_1, y_2, \cdots, y_n]$ denote the prompt label expanded from the origin label C. After obtaining the prompt labels, the CLIP pre-trained Transformer (22) model with 12 layers is leveraged to extract text embedding. The prompt labels are first encoded by the token embedding layer. Position embedding and attention blocks are also used in the text encoder. The disparity from the video encoder is that the text encoder outputs the [end of term] token instead of the [class] token. We set text encoder g_t and keep consistent embedding dimension with video embedding. The text embedding is calculated as follows:

$$g_t(y_i) = \text{transformer-encoder}_{text}(y_i) \quad [5]$$

$$F_{text} = g_t(\hat{C}) = [g_t(y_1), g_t(y_2), \cdots, g_t(y_n)] \in R^{n \times d_0} \quad [6]$$

Inter- and intra-modality loss functions

To ameliorate training efficiency, the CLIP model exploits the method of predicting image-text pair matching as the pretext task. Since the correct match between images and caption labels is unique, cross-entropy can be harnessed as the loss function. However, in our tasks, different videos of the same category are contained in the mini-batch, which results in the match between videos and texts being no longer one-to-one. Therefore, Kullback-Leibler (KL) divergence is employed to measure the similarity between the prediction and the ground truth. The cosine similarity calculation formula of f_1, f_2 is presented as Eq.7. Cosine similarity can be used to calculate inter-modal loss L_{inter} and intra-modal loss L_{intra}.

$$\text{sim}(f_1, f_2) = \text{cosine-similarity}(f_1, f_2) = \frac{f_1 \cdot f_2^T}{\|f_1\|\|f_2\|} \quad [7]$$

Inter-modality loss function

Following the form of the symmetric loss in CLIP, we symmetrically calculated the similarity coefficient between video embedding F_{vid} and text embedding F_{text}, and softmax was leveraged to convert it into the form of a probability distribution. Video-to-text and text-to-video similarity matrix can be denoted as S_{VT}, S_{TV}, respectively. Taking video-to-text as an example, let $S_{vt}^{i,j}$ represent the similarity coefficient between i-th video embedding F_{vid}^{i} and j-th text embedding F_{text}^{j}, the score can be formulated as follows:

$$L_{inter} = \frac{1}{2} E_{(V,T) \sim D_S}\left[KL(S_{VT}, Gt) + KL(S_{TV}, Gt)\right] \quad [8]$$

Intra-modality loss function

In previous work in the field of contrastive learning, features of the same class were considered to be distributed in similar positions in the embedding space (23). Inspired by this thought, we additionally calculated the intra-modality loss to enhance the representation ability of the single-modality embedding. We used S_{VV} and S_{TT} to denote the video similarity matrix and text similarity matrix, respectively. Taking video intra-modality loss as an example, video embeddings of the same category are encouraged to be more similar, which helps the encoder to mine potential class-level features rather than sample-level features. KL divergence is maintained for the loss computation.

$$L_{intra} = \frac{1}{2} E_{(V,T) \sim D_S} \left[KL\left(S_{VV}, Gt\right) + KL\left(S_{TT}, Gt\right) \right] \qquad [9]$$

It is worth noting that (24) and (25) use consistency loss to train the network. We hope to strengthen the models' robustness through the loss function. The consistency loss is for the target domain image, and the model is expected to keep the output of the perturbed images consistent with the original images. Differently, for videos of the same class in the source domain, the intra-modality loss hopes that the embeddings of the same modality are consistent, so as to determine the class-level features.

In summary, the total loss function is composed of L_{inter} and L_{intra}, which can be formulated as follows:

$$L = L_{inter} + \lambda \times L_{intra} \qquad [10]$$

where λ is hyperparameter.

Results

Dataset

We evaluated our proposed model on the MICCAI 2020 EndoVis Challenge SurgVisDom dataset (10). The training set of the SurgVisDom dataset consists of 450 VR domain and 26 clinical-like domain surgical video clips from da Vinci simulator and da Vinci system (Xi or Si), respectively. VR domain clips are captured at 60 fps and 1280×720 resolution, and clinical domain clips are captured at 20 fps and 960×540 resolution. Both VR and clinical clips contain 3 categories: dissection (DS), knot-tying (KT), and needle-driving (ND), and each clip includes only 1 category. The test set of the SurgVisDom dataset consists of 16 long videos of the clinical domain, which means that each video includes at least 1 action. Some inactive frames that do not match any action are not used for evaluation.

Evaluation metrics

We utilized the metrics and evaluation methods proposed in the Challenge to comprehensively compare different methods. The evaluation metrics are mean weighted F1-score, mean unweighted F1-score, mean global F1-score, and mean balanced accuracy score. For prediction p and ground truth y, the following metrics were utilized to evaluate different models.

Balanced accuracy

Let C, c_i represent the total number of classes in the dataset and the number of samples in i-th class, respectively. The balanced accuracy score can be formulated as follows:

$$BalancedAccuracy\left(p, y\right) = \frac{1}{C} \sum_{i=1}^{C} \sum_{j=1}^{c_i} 1_{p=y} \qquad [11]$$

F1 score

Firstly, precision (PR) and recall (RE) are computed by $PR = \frac{|y \cap p|}{|p|}$, $RE = \frac{|y \cap p|}{|y|}$, respectively. Then, F1 score is calculated as follows:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{PR} \cdot \mathrm{RE}}{\mathrm{PR} + \mathrm{RE}} \qquad [12]$$

The unweighted F1-score and global F1-score correspond to macro-F1 and micro-F1, respectively.

Among the 4 metrics, the mean weighted F1-score and the mean balanced accuracy score are harnessed to handle the data imbalance problem. The evaluation method consists of 2 steps. First, the 4 metrics between the label and the prediction of each video in the test set are calculated, and then the average value of each metric is taken as the final result. The weighted F1-score is leveraged as the ranking foundation.

Implementation details

Considering the amount of data, the video clips of the training set and test set were sampled according to their respective frame rates. The black border and the simulation operation interface are cut off due to them not being related to the surgery. The model was implemented on PyTorch and trained on Titan RTX GPU.

Our model used the parameters of the CLIP model pre-trained on the WIT-400M dataset as the initial weights.

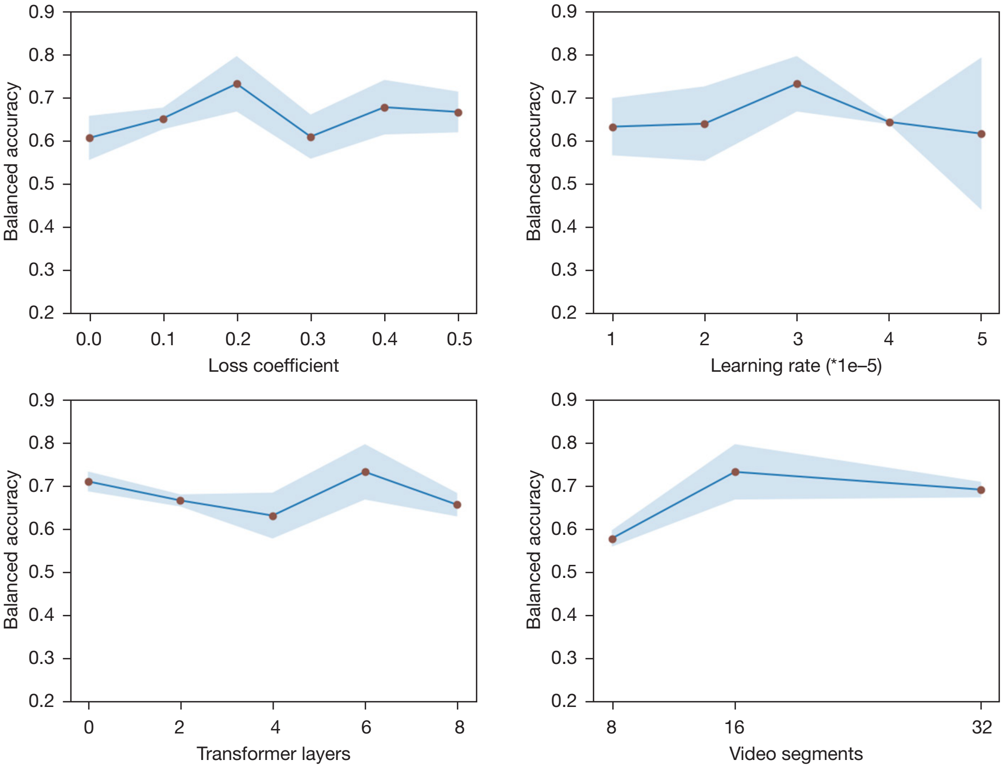


Figure 3 Using 5-fold cross-validation to select model parameters. The blue area represents the SD, and the dots represent the mean value. SD, standard deviation.

The model was trained for 50 epochs by AdamW optimizer, and the first 10% epochs were set to warm-up phase. We use multi-scale crop, random horizontal flip, random color jitter, and random gray scale for data augmentation. The Rand Augment strategy (26) with N=4, M=9 was also employed to alleviate the overfitting. Our ablation experiments found that the trainable text encoder could perform better, and therefore we did not freeze it.

On the basis of the ActionCLIP (16), the hyperparameters were chosen by 5-fold cross-validation on the training set including VR and clinical-like clips. We chose the balanced accuracy score as the metric for hyperparameters selection and the comparison of different configurations is shown in *Figure 3*. According to the experiments, the video clip was divided into *s*=16 segments. A 6 layers transformer fusion model and λ=0.2 were adopted. The learning rates of the visual and text encoders were set to 3×10^{-5} and 5×10^{-5} for soft and hard domain adaptation tasks, respectively. The learning rate of the fusion model is 10 times that of the encoder. During the training stage, a frame was randomly selected from each divided segment during each epoch. During the inference stage, we first cut the test video into clips with 128 frames, and then used the same method as in the training process to infer the category of each clip, which represents the category of all the included frames. Our ablation experiments found that the trainable text encoder could perform better. Therefore, we did not freeze it and set the learning rate the same as the visual encoder.

Comparison with state-of-the-art methods

For the hard domain adaptation task, a total of 2 teams participated. The results are presented in *Table 2*. Team Parakeet utilized the 2DVGG16 (27) network to extract 8 frames of image features as input, then a 3-dimensional (3D) convolution kernel was harnessed to further fuse spatio-temporal information. Finally, the fully connected layer was exploited to obtain action recognition results. Team SK employed 2 pre-processing methods to extract instrument segmentation masks and contours. After that, the masks, contours, and the original RGB images were input to the SlowFast (28) network pre-trained on the Kinetics400

Table 2 Action recognition results of different models on task 1: hard domain adaptation

Method	Weighted F1-score	Unweighted F1-score	Global F1-score	Balanced accuracy
Rand	0.450	0.207	0.327	0.327
SK	0.460	0.225	0.370	0.369
Parakeet	0.470	0.266	0.475	0.559
SDA-CLIP	0.659 (+18.9%)	0.546 (+28.0%)	0.647 (+17.2%)	0.609 (+5.0%)

SDA-CLIP, surgical domain adaptation method based on the contrastive language-image pretraining model.

Table 3 Action recognition results of different models on task 2: soft domain adaptation

Method	Weighted F1-score	Unweighted F1-score	Global F1-score	Balanced accuracy
Rand	0.450	0.207	0.327	0.327
SK	0.600	0.414	0.599	0.644
ECBA	0.790	0.488	0.742	0.776
Parakeet	0.800	0.604	0.774	0.778
SDA-CLIP	0.844 (+4.4%)	0.667 (+6.3%)	0.858 (+8.4%)	0.811 (+3.3%)

SDA-CLIP, surgical domain adaptation method based on the contrastive language-image pretraining model.

dataset to obtain the prediction results.

It was evident that our SDA-CLIP substantially improved all metrics, surpassing the first-place team of the challenge by 18.9%, +28.0%, +17.2%, and +5.0% in each metric, respectively. The weighted F1-scores of other models were similar to random guessing, meaning that the model only learns limited knowledge. Although the balanced accuracy of the Parakeet was acceptable, the poverty in weighted F1-score illustrated that the model tended to predict positive samples as negative.

For the soft domain adaptation task, a total of 3 teams participated. The results are presented in *Table 3*. Team Parakeet and Team SK utilized the same models as the hard task. Team ECBA tapped into an image segmentation network. They proposed RASNet to extract surgical instrument segmentation in VR and clinical domains, then 3D ResNet-18 (29) was used to extract spatio-temporal fusion features. Finally, the stride convolution was leveraged to obtain the action prediction of the test video. Our SDA-CLIP still achieved the best performance on all metrics, outperforming the first-place team of the challenge by 4.4%, 6.3%, 1.9%, and 3.3%, respectively. Although team SK used multi-modal image information (segmentation mask, edge counter) and reached a fairly high 80% weighted F1-score, our model was still able to improve the accuracy on multiple metrics, illustrating the effectiveness of using video text pairs.

In summary, our SDA-CLIP not only achieved superior performance on the SurgVisDom dataset but also outperformed the state-of-the-art methods by a significant margin. Furthermore, our model does not design additional unit modules for the domain adaptation task, which indicates that the method can be extended to other tasks, such as surgical action segmentation, phase recognition, and so on.

Ablation study

In order to illustrate the efficacy of our contributions, ablation models were constructed for evaluation.

(I) Pure ViT: the model only leverages the video encoder and ends with a linear projection head to classify videos. The cross-entropy loss is selected as the loss function between labels and predictions.

(II) ViT + Text-f: auxiliary text modality is utilized in the model, and the video-text matching loss based on KL divergence is employed as the loss function. However, the text encoder's parameters are frozen, similar to many nature-scene CLIP-driven models.

(III) ViT + Text-t: given the domain gap between the surgical and general scenes, we made the parameters of the text encoder trainable. This

Table 4 Ablation settings on key components for surgical domain adaptation

Method	Component		Metrics (hard/soft)			
	Text	L_{intra}	Weighted F1	Unweighted F1	Global F1	Balanced accuracy
Pure ViT	×	×	0.531**/0.630**	0.390*/0.489**	0.536*/0.684**	0.585/0.756
ViT + Text-f	Freeze	×	0.569*/0.791**	0.438*/0.659**	0.621/0.796**	0.696/0.822
ViT + Text-t	Train	×	0.579*/0.798	0.428*/0.636	0.610/0.797*	0.677/0.769
SDA-CLIP-f	Freeze	√	0.655/0.829	0.520/0.653	0.630/0.857	0.603/0.784
SDA-CLIP	Train	√	0.659/0.844	0.546/0.667	0.647/0.858	0.609/0.811

* and ** denote that the P values of the paired *t*-test are less than 0.05 and 0.01, respectively. Text, text encoder; L_{intra}, intra-modality loss function; hard, hard domain adaptation task; soft, soft domain adaptation task; ViT, Vision Transformer; SDA-CLIP, surgical domain adaptation method based on the contrastive language-image pretraining model.

Table 5 Recognition results with 3 kinds of prompt templates for surgical domain adaptation

Template type	Metrics (hard/soft)			
	Weighted F1	Unweighted F1	Global F1	Balanced accuracy
Caption template	0.565**/0.776	0.446/0.541*	0.600*/0.788	0.654*/0.734
General template	0.596/0.782	0.367*/0.531	0.602**/0.782	0.628**/0.750
Customized template	0.631/0.796	0.474/0.583	0.677/0.796	0.702/0.754

We conducted the paired *t*-test for caption template *vs.* customized template and general template *vs.* customized template. * and ** denote the P values less than 0.05 and 0.01, respectively. hard, hard domain adaptation task; soft, soft domain adaptation task.

setting is commonly employed in CLIP-driven medical image analysis methods (30-32).

(IV) SDA-CLIP-f: the text encoder's parameters are frozen, and the intro-modality loss function is employed.

(V) SDA-CLIP: we employed trainable text encoder and intra-modality loss to constrain the learning of the network (i.e., our complete proposed model).

The results of different models are presented in *Table 4*. We can see that the Pure ViT model obtained averaged recognition results on the soft domain adaptation task, and the capability of the transformer backbone is verified on the hard domain adaptation task. On top of that, adding the text modality steadily boosted the performance on both tasks. As shown in *Table 4*, the improvement is statistically significant. Comparing the results of ViT + Text and SDA-CLIP, we observed that the L_{intra} can improve the model property in almost all evaluation metrics. Although the balanced accuracy score of the SDA-CLIP model had decreased, the increase in the weighted F1-score indicated that the ability to recognize false negative samples had improved, which is meaningful to clinical application. Additionally, we found that the trainable text encoder can slightly improve the weighted F1-score in both tasks. This is due to the domain gap between surgical and general text descriptions. In summary, both elements contribute to the eventual progress.

In prompt engineering, the category names are expanded by the caption, general, and customized templates. Here, we compare the recognition results of various kinds of prompt templates. None of the models in the following utilize intra-modality losses for either visual or text modalities.

(I) Caption template: the text label for each sample in the training batch is generated using the unique caption prompt template provided in *Table 1*.

(II) General template: the prompt template for each sample in every training batch is randomly selected from the 8 candidate general templates in *Table 1*.

(III) Customized template: the prompt template is randomly selected from the 8 candidate customized templates in *Table 1*.

The hyperparameters are consistent with previous experiments. From *Table 5*, we can see that the task-customized templates performed better than other template

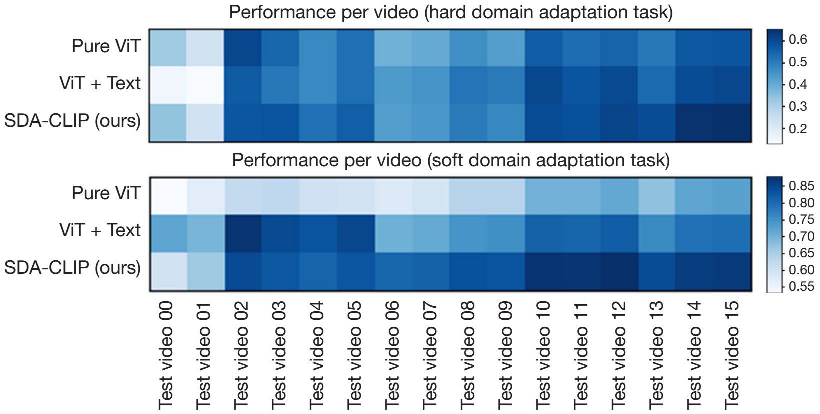


Figure 4 Weighted F1-score for each ablation model for each test video in hard and soft domain adaptation tasks. ViT, Vision Transformer; SDA-CLIP, surgical domain adaptation method based on the contrastive language-image pretraining model.

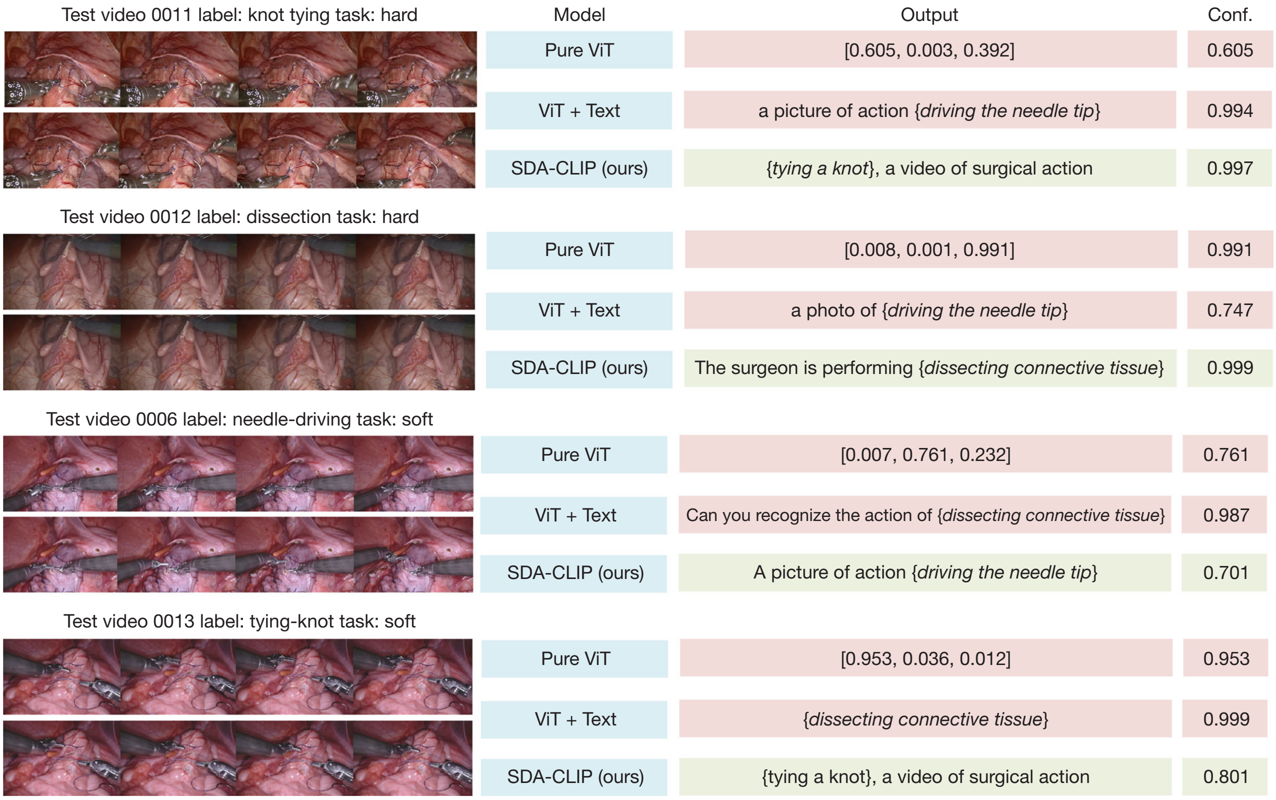
types. Compared to the caption template, the customized templates improved the weighted F1 score by 6.6% and 2.0% on the 2 tasks respectively. This conclusion is identical with that of another study (10). Additionally, we calculated the P-values using the *t*-test to conduct statistical analysis. When comparing customized templates and other options, it can be observed that we obtained $P<0.05$ in practically all metrics in the hard domain adaption task. There was no statistical significant difference between general templates and caption templates in the paired *t*-test. This indicates that using the task description terms "surgical action" and "surgeon" in prompt engineering can provide useful information and improve the model's performance. The benefit provided by the customized templates in the soft domain adaptation task is less pronounced than in the earlier task. In our opinion, this is due to knowledge about the target domain being leaked from some of the clinical videos that were already included in the training set. The information offered by task-customized templates becomes somewhat redundant as a result of this leakage.

The performance of the ablation models on each test video is visualized in *Figure 4*, where the metric chosen is the weighted F1-score. Darker colors indicate higher weighted F1-scores, signifying better model performance. Our model's superior performance on the majority of test videos visually demonstrates the effectiveness of the proposed method. *Figure 5* illustrates the results of recognizing various video clips by the 3 ablation models in 2 domain adaptation tasks. In the Pure ViT model, where the last layer is linear, the output consists of probability distributions. For the models utilizing a text encoder, the final output is the text that exhibits the highest similarity to the visual features. The pure ViT model demonstrates an inability to correctly identify results solely with the visual encoder. Leveraging text features and intra-modal loss functions, our SDA-CLIP model can boost the prediction results with higher probabilities towards ground truth surgical actions.

Discussion

Surgical domain adaptation assumes a pivotal role in establishment of a robust data-driven model, which not only safeguards patient privacy but also alleviates the burden on clinicians in terms of data annotation. By introducing the text modality, we proposed a novel multi-modality surgical domain adaptation model grounded in contrastive learning. Our model significantly outperforms the existing surgical domain adaptation methods, showcasing a substantial margin of improvement.

In the pursuit of improving the generalization capability in new data domains, domain adaptation models usually leverage consistent information across different domains. Previous methods (10) focused on extracting the edge or segment mask of the surgical instruments in the image and thus nonlinearly projected the source image data into a novel space. However, this approach may sacrifice information. Empirically, we found that the performance in some metrics is close to that of random guessing (*Table 2*), indicating that the models did not learn effective discriminative features. To address this, we took advantage of text labels as consistent information across domains, enabling the connection of data domains without changing the original data space. Moreover, the choice of the



	Model	Output	Conf.
Test video 0011 label: knot tying task: hard	Pure ViT	[0.605, 0.003, 0.392]	0.605
	ViT + Text	a picture of action {*driving the needle tip*}	0.994
	SDA-CLIP (ours)	{*tying a knot*}, a video of surgical action	0.997
Test video 0012 label: dissection task: hard	Pure ViT	[0.008, 0.001, 0.991]	0.991
	ViT + Text	a photo of {*driving the needle tip*}	0.747
	SDA-CLIP (ours)	The surgeon is performing {*dissecting connective tissue*}	0.999
Test video 0006 label: needle-driving task: soft	Pure ViT	[0.007, 0.761, 0.232]	0.761
	ViT + Text	Can you recognize the action of {*dissecting connective tissue*}	0.987
	SDA-CLIP (ours)	A picture of action {*driving the needle tip*}	0.701
Test video 0013 label: tying-knot task: soft	Pure ViT	[0.953, 0.036, 0.012]	0.953
	ViT + Text	{*dissecting connective tissue*}	0.999
	SDA-CLIP (ours)	{tying a knot}, a video of surgical action	0.801

Figure 5 Comparison of the predictions made by our method under different ablation settings. ViT, Vision Transformer; SDA-CLIP, surgical domain adaptation method based on the contrastive language-image pretraining model; hard, hard domain adaptation task; soft, soft domain adaptation task; Conf., the confidence level of the output.

backbone network is crucial to the final experimental results. Compared to previous studies that utilized ResNet (29) as the backbone encoder, we opted for the more advanced ViT (20). From the ablation studies displayed in *Table 4*, we found that the performance of the Pure ViT model, which only contains a video encoder, exceeds that of previous models in the challenging domain adaptation task.

In general scenes, numerous studies utilizing CLIP (15) have often chosen to keep the text encoder fixed during model training (18,33). Indeed, text in medical settings may differ from that present in general scenes, and many medical vision-language models set the text encoder as trainable (30-32). Consequently, we have made the text encoder trainable to enhance its adaptability to the surgical domain. As shown in *Table 4*, a trainable text encoder can slightly improve the performance in both tasks. In the future, we plan to utilize the encoder trained on biomedical text (31) to extract features.

However, our model does have a few limitations. We observed a drop in the balanced accuracy of our SDA-CLIP. This can be attributed to the class imbalance in the dataset, which would have caused the model to overfit to the dominant class, and the excessive similarity between categories may have also affected model learning. Future research can explore methods such as data resampling and different loss weights to address the class imbalance issue and improve model performance. Furthermore, ensuring real-time recognition is essential for practical clinical applicability. Although we adopted a larger backbone network to enhance model accuracy, this led to decreased inference speed and increased clinical deployment expenses. To address this, we will focus on optimizing the model's efficiency and computational complexity. Techniques such as weight sharing, pruning, and separable convolutions can be applied to reduce the model's parameters and computational demands, making it more lightweight and suitable for deployment in real-world clinical settings. In conclusion, we believe that our approach can serve as a catalyst for inspiring further investigations into effectively analyzing surgical domain adaptation using vision-language models.

Conclusions

We have proposed a novel framework SDA-CLIP for surgical domain adaptation and action recognition. Since our model more effectively utilizes the rich semantic information in video and text labels, it has good transfer and generalization ability. Our SDA-CLIP outperforms the SOTA model by a large margin on multiple tasks in the SurgVisDom dataset. We hope that future work will focus on surgical video analysis using video-text pairs. We will validate the capabilities of the model on a larger dataset with richer textual descriptions, and extend this visual-textual model to other tasks.

Acknowledgments

Funding: The present study was supported in part by the National Natural Science Foundation of China (Nos. 62172401 and 12026602), the Guangdong Natural Science Foundation (Nos. 2022A1515010439 and 2022A0505020019), the Shenzhen Key Basic Research Grant (No. JCYJ20220818101802005), the Zhuhai Science and Technology Program (No. ZH22017002210017PWC), and the Shenzhen Key Laboratory Program (No. ZDSYS201707271637577).

Footnote

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://qims.amegroups.com/article/view/10.21037/qims-23-376/coif). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. The datasets used in this paper are from public open data.

References

1. Zhang Z. Big data and clinical research: focusing on the area of critical care medicine in mainland China. Quant Imaging Med Surg 2014;4:426-9.
2. Maier-Hein L, Vedula SS, Speidel S, Navab N, Kikinis R, Park A, et al. Surgical data science for next-generation interventions. Nat Biomed Eng 2017;1:691-6.
3. Maier-Hein L, Eisenmann M, Sarikaya D, März K, Collins T, Malpani A, et al. Surgical data science - from concepts toward clinical translation. Med Image Anal 2022;76:102306.
4. Padoy N. Machine and deep learning for workflow recognition during surgery. Minim Invasive Ther Allied Technol 2019;28:82-90.
5. Brown KC, Bhattacharyya KD, Kulason S, Zia A, Jarc A. How to Bring Surgery to the Next Level: Interpretable Skills Assessment in Robotic-Assisted Surgery. Visc Med 2020;36:463-70.
6. Zia A, Guo L, Zhou L, Essa I, Jarc A. Novel evaluation of surgical activity recognition models using task-based efficiency metrics. Int J Comput Assist Radiol Surg 2019;14:2155-63.
7. Gendia A. Cloud Based AI-Driven Video Analytics (CAVs) in Laparoscopic Surgery: A Step Closer to a Virtual Portfolio. Cureus 2022;14:e29087.
8. Gostin LO, Halabi SF, Wilson K. Health Data and Privacy in the Digital Era. JAMA 2018;320:233-4.
9. Ng D, Lan X, Yao MM, Chan WP, Feng M. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. Quant Imaging Med Surg 2021;11:852-7.
10. Zia A, Bhattacharyya K, Liu X, Wang Z, Kondo S, Colleoni E, van Amsterdam B, Hussain R, Hussain R, Maier-Hein L. Surgical visual domain adaptation: Results from the MICCAI 2020 SurgVisDom challenge. arXiv preprint arXiv:2102.13644, 2021.
11. Vercauteren T, Unberath M, Padoy N, Navab N. CAI4CAI: The Rise of Contextual Artificial Intelligence in Computer Assisted Interventions. Proc IEEE Inst Electr Electron Eng 2020;108:198-214.
12. Pfeiffer M, Funke I, Robu MR, Bodenstedt S, Strenger L, Engelhardt S, Roß T, Clarkson MJ, Gurusamy K, Davidson BR, editors. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In: Medical Image Computing and Computer Assisted Intervention-MICCAI. Springer, Cham; 2019;119-27.

13. Cai R, Li Z, Wei P, Qiao J, Zhang K, Hao Z. Learning Disentangled Semantic Representation for Domain Adaptation. IJCAI (U S) 2019;2019:2060-6.
14. Huang Y, Du C, Xue Z, Chen X, Zhao H, Huang L. What makes multi-modal learning better than single (provably). In: Advances in Neural Information Processing Systems. 2021;34:10944-56.
15. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I. Learning transferable visual models from natural language supervision. International Conference on Machine Learning-ICML. 2021;PLMR 139.
16. Wang M, Xing J, Liu Y. ActionCLIP: A new paradigm for video action recognition. arXiv preprint arXiv:2109.08472, 2021.
17. Li M, Chen L, Duan Y, Hu Z, Feng J, Zhou J, Lu J. Bridge-prompt: Towards ordinal action understanding in instructional videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition-CVPR; 2022;19848-57.
18. Wu W, Sun Z, Ouyang W. Revisiting classifier: Transferring vision-language models for video recognition. In: Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI); 2023;1-15.
19. Liu J, Zhang Y, Chen JN, Xiao J, Lu Y, Landman BA, Yuan Y, Yuille A, Tang Y, Zhou Z. CLIP-driven universal model for organ segmentation and tumor detection. arXiv preprint arXiv:2301.00785, 2023.
20. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929,2020.
21. Zhou K, Yang J, Loy CC, Liu Z. Learning to prompt for vision-language models. Int J Comput Vis 2022;130:2337-48.
22. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention is all you need. In: 31st Annual Conference on Neural Information Processing Systems –NIPS. 2017;30.
23. Ye M, Zhang X, Yuen PC, Chang S-F. Unsupervised embedding learning via invariant and spreading instance feature. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition-CVPR; 2019;6203-12.
24. Sahu M, Strömsdörfer R, Mukhopadhyay A, Zachow S. Endo-Sim2Real: Consistency learning-based domain adaptation for instrument segmentation. Medical Image Computing and Computer Assisted Intervention-MICCAI. Springer, Cham 2020;784-94.
25. Sahu M, Mukhopadhyay A, Zachow S. Simulation-to-real domain adaptation with teacher-student learning for endoscopic instrument segmentation. Int J Comput Assist Radiol Surg 2021;16:849-59.
26. Cubuk ED, Zoph B, Shlens J, Le QV. RandAugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition-CVPR; 2020;3008-17.
27. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.
28. Feichtenhofer C, Fan H, Malik J, He K. Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision-ICCV; 2019;6201-10.
29. 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-CVPR; 2016;770-8.
30. Eslami S, Melo G, Meinel C. Does CLIP benefit visual question answering in the medical domain as much as it does in the general domain? arXiv preprint arXiv:2112.13906, 2021.
31. Zhang S, Xu Y, Usuyama N, Bagga J, Tinn R, Preston S, Rao R, Wei M, Valluri N, Wong C, Lungren MP, Naumann T, Poon H. Large-scale domain-specific pretraining for biomedical vision-language processing. arXiv preprint arXiv:2303.00915, 2023.
32. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163, 2022.
33. Tevet G, Gordon B, Hertz A, Bermano AmitH, Cohen-Or D. MotionCLIP: Exposing human motion generation to CLIP space. arXiv preprint arXiv:2203.08063, 2022.

Cite this article as: Li Y, Jia S, Song G, Wang P, Jia F. SDA-CLIP: surgical visual domain adaptation using video and text labels. Quant Imaging Med Surg 2023;13(10):6989-7001. doi: 10.21037/qims-23-376