International Journal of Computer Assisted Radiology and Surgery
https://doi.org/10.1007/s11548-025-03406-0

ORIGINAL ARTICLE

Evaluating the generalizability of video-based assessment of intraoperative surgical skill in capsulorhexis

Zhiwei Gong[1] · Bohua Wan[2] · Jay N. Paranjape[3] · Shameema Sikder[1,4] · Vishal M. Patel[3] · S. Swaroop Vedula[1]

Received: 4 November 2024 / Accepted: 15 April 2025

Abstract

Purpose Assessment of intraoperative surgical skill is necessary to train surgeons and certify them for practice. The generalizability of deep learning models for video-based assessment (VBA) of surgical skill has not yet been evaluated. In this work, we evaluated one unsupervised domain adaptation (UDA) and three semi-supervised (SSDA) methods for generalizability of models for VBA of surgical skill in capsulorhexis by training on one dataset and testing on another.
Methods We used two datasets, D99 and Cataract-101 (publicly available), and two state-of-the-art models for capsulorhexis. The models include a convolutional neural network (CNN) to extract features from video images, followed by a long short-term memory (LSTM) network or a transformer. We augmented the CNN and the LSTM with attention modules. We estimated accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC).
Results Maximum mean discrepancy (MMD) did not improve generalizability of CNN-LSTM but slightly improved CNN transformer. Among the SSDA methods, Group Distributionally Robust Supervised Learning improved generalizability in most cases.
Conclusion Model performance improved with the domain adaptation methods we evaluated, but it fell short of within-dataset performance. Our results provide benchmarks on a public dataset for others to compare their methods.

Keywords Surgical skill assessment · Domain adaptation · Transformer · Cataract surgery

Introduction

Patients have poor outcomes when their surgeons have poor skill [1, 2]. Assessment of surgical skill is critical to support surgeons' learning during their training and independent practice [3]. Valid and rapid surgical skill assessment is crucial to support surgeons' learning needs globally.

Traditionally, surgical skill assessment was based upon human experts and direct observation, which is subjective and inefficient. On the other hand, surgical data science methods have enabled data-driven assessment of surgical skill using various sources of data. Among the various data sources, videos of the surgical field potentially provide the most information on surgical skill. Consequently, video-based assessment (VBA) is being evaluated for routine use in training and practicing surgeons [3, 4].

Generalizability of deep learning models for VBA of surgical skill across datasets is a critical yet nontrivial problem.

Zhiwei Gong and Bohua Wan contributed equally to this work.

✉ S. Swaroop Vedula
swaroop@jhu.edu

Zhiwei Gong
zhiweigong75@gmail.com

Bohua Wan
bwan2@jhu.edu

Jay N. Paranjape
jparanj1@jhu.edu

Shameema Sikder
ssikder1@jhmi.edu

Vishal M. Patel
vpatel36@jhu.edu

[1] Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD 21218, USA

[2] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA

[3] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA

[4] Wilmer Eye Institution, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA

Published online: 22 May 2025

Using intraoperative videos of surgery in patients, few studies have reported within-dataset performance of deep learning models for VBA of surgical skill [5–7]. None of the studies evaluated or reported methods for generalizability of the deep learning models in new datasets. Dataset shifts are the norm and they can result from many factors such as different surgeons, sensors, data collection protocols, and annotation schema. Deep learning models for VBA of surgical skill for a given step or procedure are useful when they are generalizable.

Related work: The primary objective of domain adaptation methods is to improve the target domain performance of models that are developed on source domain. Both semi-supervised and unsupervised methods for domain adaptation have been proposed in the context of various tasks. Semi-supervised domain adaptation (SSDA) methods utilize only a small fraction of labeled target domain data and unsupervised domain adaptation (UDA) methods require no labeled data from target domain. They both stand out in the context of VBA of surgical skill due to limited access to target domain data. In [8], Paranjape et al. developed a UDA method using maximum mean discrepancy (MMD) [9] for instrument classification in surgical video images. Their work demonstrated that MMD could effectively reduce domain shifts in surgical video data by minimizing the divergence between latent representations, providing a foundation for adapting UDA techniques to temporal tasks.

For VBA of intraoperative surgical skill, [5] proposed a model that includes ResNet50, followed by a LSTM, in which both networks were augmented by attention mechanisms. This attention-augmented architecture achieved promising within-dataset performance for skill assessment in capsulorhexis and established a baseline for subsequent methods. However, it did not evaluate generalizability to external datasets, highlighting the need to establish external validity. Thus, our study directly addresses this gap by adopting the ResNet-LSTM model as a backbone to rigorously evaluate cross-dataset generalizability using domain adaptation methods. A 3D-convolutional neural network was used by a team participating in a challenge reported in [7]. Another study used ResNet101 to extract features that were then analyzed with a multilayer perceptron to predict skill [6]. Both works demonstrate the ResNet's efficiency in capturing both spatial and temporal information of surgical videos.

Contributions: In this paper, we evaluate SSDA and UDA methods for generalizability of models for VBA of surgical skill in capsulorhexis with the following contributions: (1) We establish the state-of-the-art performance of SSDA and UDA methods for generalizability of algorithms for VBA of intraoperative skill in a critical step in cataract surgery. Our literature search did not retrieve any studies evaluating generalizability of networks for VBA of surgical skill for any procedure. (2) We evaluate the utility of UDA with MMD for prediction tasks that use temporal models. (3) We motivate new research on generalizability of networks for VBA of intraoperative surgical skill because we use a publicly accessible target dataset.

Methods

Preliminaries: In the skill assessment task, N_s videos from the source domain $\mathcal{D}_s$ and N_t videos from the target domain $\mathcal{D}_t$ are given. Each source data consists of (X_s, y_s), where $X_s = \{x_{s1}, x_{s2}, ..., x_{sn} | x_{sn} \in \mathbb{R}^{3\times H\times W}\}$ denotes a source video with n frames, in which 3 indicates the channel dimension of the frame. The label $y_s \in \{0, 1\}$ denotes the corresponding skill assessment label. Target video and label pair (X_t, y_t) is similarly defined, where $X_t = \{x_{t1}, x_{t2}, ..., x_{tm} | x_{tm} \in \mathbb{R}^{3\times H\times W}\}$ denotes a target video with m frames and $y_t \in \{0, 1\}$ denotes the skill label. SSDA and UDA methods are adopted to improve deep learning model's ability in predicting skill labels given videos from the target domain.

Model architecture

Our work builds upon the findings of [5, 10], we use the model architecture shown in Figure 1, incorporating a feature extractor and a LSTM [11] with spatial and temporal attention mechanisms [12, 13]. We use a second network architecture in which we replace the LSTM cell and the temporal attention module in Figure 1 with a transformer-based architecture [14].

Unsupervised domain adaptation

Let $\mathcal{D}_s = \{(X_s, y_s)_i | i \in [1, N_s]\}$ and $\tilde{\mathcal{D}}_t = \{(\tilde{X}_t, \tilde{y}_t)_i | i \in [1, \tilde{N}_t]\}$ be the source domain with labels and target domain without labels. The goal of performing unsupervised domain adaptation approach is to train a robust skill assessment model using samples from $\mathcal{D}_s$ and $\tilde{\mathcal{D}}_t$ and show good performance when we test on $\tilde{\mathcal{D}}_t$.

We employ an additional MMD loss within the UDA framework. Following the residual blocks of ResNet50, an adaptation layer with MMD is introduced. We hypothesize that this layer has the potential to reduce distributional disparities in temporal features between the source and target domains. The MMD loss is calculated as follows:

$$L_{\text{MMD}}(X_s, X_t) = \mathbb{E}[\kappa(x_s, x_s')] + \mathbb{E}[\kappa(x_t, x_t')] - 2\mathbb{E}[\kappa(x_s, x_t)], \qquad (1)$$

where $\kappa(\cdot)$ is the Gaussian kernel function. The total loss is $L_{\text{total}} = L_{\text{BCE}} + \lambda L_{\text{MMD}}$.

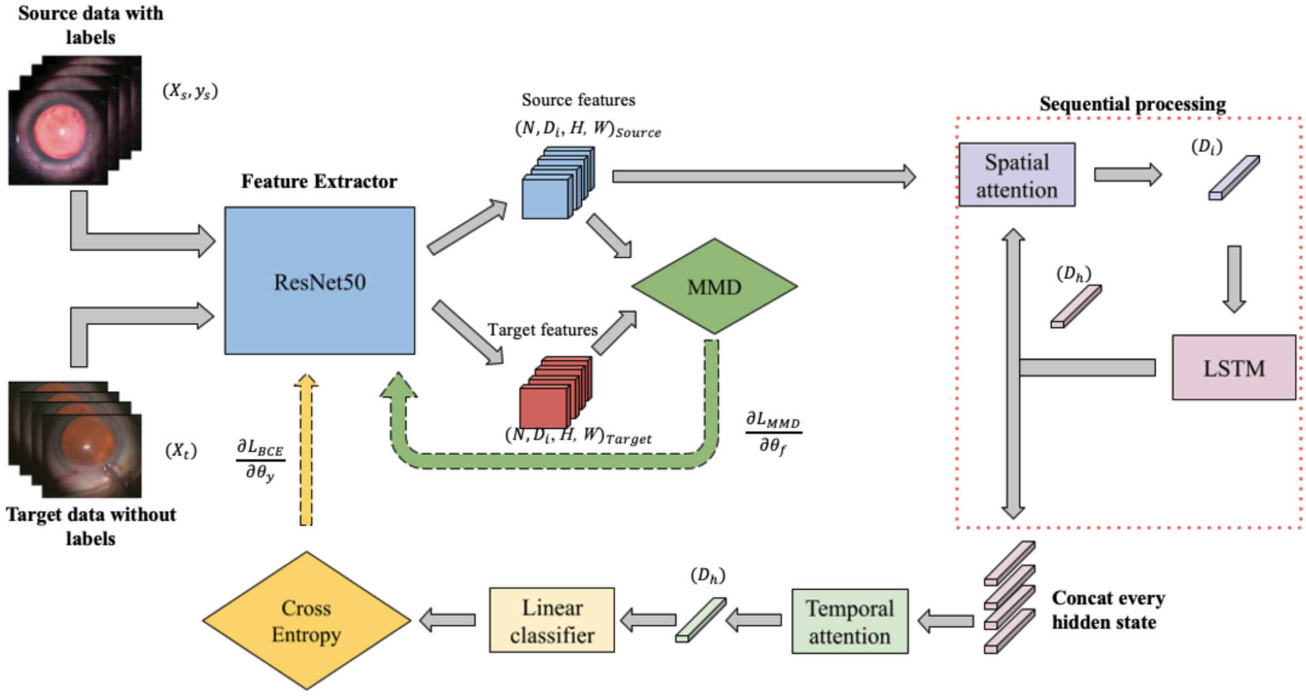

Fig. 1 Architecture of our models based on the work of [5]. Given all frames of the RGB video in the source and the target datasets as input, a CNN backbone extracts per-frame source and target image features, which are used to compute the MMD loss. The source features are then passed into the spatial attention module to produce per-frame features. These features are concatenated together as input to the temporal attention module, the final hidden state of the temporal attention module is passed into a linear classifier to predict the skill label

Semi-supervised domain adaptation

Given a small subset of the target domain with labeled samples denoted as $\mathcal{D}_t = \{(X_t, y_t)_i | i \in [1, N_t]\}$, where $N_t << \tilde{N}_t$. SSDA aims to train a robust model using samples from $\mathcal{D}_s$ and $\mathcal{D}_t$ and show good performance when we employ it on $\tilde{\mathcal{D}}_t$.

We evaluate three SSDA methods. The first method, denoted as Vanilla SSDA, incorporates a small number of samples from the target domain into the source domain. The second strategy is predicated upon the Group Distributionally Robust Supervised Learning (Group-DRSL) as delineated in [15]. The third method, termed Weighted-Group-DRSL, extends upon the second approach by integrating class weighting.

Vanilla SSDA

This method simply combines all samples from the source domain and a small number of target samples with labels as the training set to train the model. The combined training set aims to harness the inherent patterns of the source domain while also capturing the nuances of the target domain through its labeled samples. Such an approach seeks to enhance the model's adaptability and generalization capabilities, ensuring a more comprehensive understanding of both domains during training.

Group-DRSL

We augment the Vanilla SSDA methodology with the Group-DRSL framework [15]. This framework extends conventional DRSL based on the latent prior probability change assumption [16]. We introduce $k \in \{1, 2, ..., K\}$, a latent variable, which serves the function of categorizing the dataset into K distinct groups. Let P_s and P_t be the probability distribution of the source domain samples and the target domain samples, respectively. The latent probability change assumption [16] necessitates the equality of $P(X, y|k)$ across domains. In our analysis, this assumption is validated when the target data samples are drawn independently and identically from the target distribution. Consequently, we can infer that $P_s(X, y|k = \text{target}) = P_t(X, y|k = \text{target})$. To enhance the robustness of this approach, we leverage an adversarial training paradigm by introducing a learnable weight, denoted as $w(k)$, which serves to re-weight the loss for samples, ensuring a balanced representation between both the source and target domains. This method aims to minimize:

$$\min_{\theta} \sup_{w \in W} \frac{1}{N} \sum_{k=\text{source}}^{\{\text{target},\text{source}\}} n_k w(k) L(k;\theta), \tag{2}$$

$$W = \{w \in \mathbb{R}^2 | \frac{1}{N} \sum_{k=\text{source}}^{\{\text{target},\text{source}\}} n_k w(k) = 1, w \leq 0\}, \tag{3}$$

where $\sup_{w \in W}$ means the smallest number w that is greater-than-or-equal to every number in the set that makes the largest value of $\frac{1}{N} \sum_{k=\text{source}}^{\{\text{target},\text{source}\}} n_k w(k) L(k;\theta)$, and $L(k;\theta)$ represents the averaged loss within group k. In our experiment, $w(k = \text{target})$ can be simulated by one learnable parameter, and $w(k = \text{source}) = \frac{N - n_t w(k=\text{target})}{n_s}$.

The learnable parameter aims to maximize the loss term, embodying the adversarial component of the training process. It is periodically frozen per standard adversarial training paradigms. Thus, our approach adheres to domain adaptation principles and improves the model's capacity to generalize from the source to the target domain.

Weighted-group-DRSL

Assigning weights to samples is a conventional strategy to address imbalanced class label distributions. By assigning a higher weight to samples from minority classes, their impact and significance throughout the back-propagation process are enhanced. This is attributed to the assigned weight's capacity to amplify the gradient, ensuring that the less frequent class labels exert a more substantial influence on the model's learning. Consequently, this approach aids in correcting biases toward majority classes. In our case, the total loss can be calculated as follows:

$$L_{\text{total}} = \frac{1}{N} \sum_{(X,y) \in D_s \cup D_t} (w'_{\text{positive}} \hat{y} + w'_{\text{negative}} (1 - \hat{y})) L(X, y), \tag{4}$$

where $\hat{y} = 1$ if y is a positive sample, otherwise, $\hat{y} = 0$. Then, we combine this new total loss with the Group-DRSL framework.

Datasets

We used two video datasets of capsulorhexis-D99 [17] and Cataract-101 [18]. The D99 dataset includes 99 videos captured from the operating microscope, and processed to a resolution of 640 × 480 pixels at 59 frames per second (*fps*). Cataract-101 is a public dataset of 101 cataract surgery videos, which have a resolution of 720 × 540 pixels and a frame rate of 25 *fps*. We used 97 videos in Cataract-101 for which it was possible to evaluate skill. The binary ground truth skill labels for both the datasets were specified by an expert surgeon using the same criteria as in [17].

The research reported in this manuscript is in accordance with protocols approved by the Institutional Review Board at Johns Hopkins Medical Institutions. Informed consent was waived because we used deidentified retrospective data and one of the datasets we used is publicly available.

Variation between the datasets

Figure 2 shows sample images from the datasets and variation in appearance. Between datasets, the videos differ in their length. The mean (standard deviation) of video duration is 8,677 (6,208) frames in D99 and 5,353 (2,914) in Cataract-101. Surgeons in D99 include both faculty and trainees. In Cataract-101, only experienced surgeons were included. The difference in surgical experience introduces covariate shifts that manifest in the speed of movements and video duration. In D99, 50 and 49 videos were labeled as expert and novice, respectively. In Cataract-101, 49 and 48 videos were labeled as expert and novice, respectively. We split each dataset into five folds for cross-validation (Table 1).

Data processing

We sampled frames from videos as in [5] for both datasets. For each video, we extracted 256 frames starting with selecting a frame with uniform probability and then every eighth frame for all models trained on D99, every 32nd frame for LSTM models trained on Cataract-101, and every 128th frame for transformer models trained on Cataract-101 (the interval between frames was empirically chosen). The minimum of an eight-frame interval allowed retention of temporal information relevant for skill. For each time 256 frames are extracted from a video, a distinct clip is retrieved because of the random start. Furthermore, as sampled clips from each video are different in every training epoch, we hypothesize that a comprehensive number of epochs would encompass the full scope of the video through the aggregated sampled clips. During testing, we sampled three clips from each video, and computed the prediction by averaging the results on the three clips. In addition, we resized all images to 256 × 256 size applied random cropping size of 224 × 224. We use the Albumentation 1.01 framework [19] for augmentations. We applied a horizontal flip with a probability of 0.5, the rotation angle is 30 degree. The color jittering was applied with the default values of 0.2 brightness, 0.2 contrast, 0.2 saturation, and 0.2 hue.

Experimental setup

We evaluated the domain adaptation methods on four models: ResNet50 and LSTM with spatial and temporal attention modules (LSTM-ATT), ResNet50 and LSTM without spatial and temporal attention modules (LSTM-no-ATT),

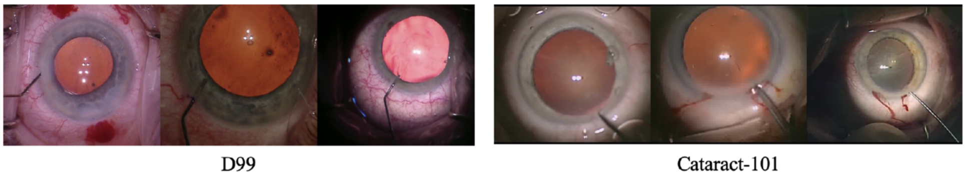

Fig. 2 Sample images from the D99 dataset and the Cataract-101 dataset

Table 1 Statistics of Cataract-101 dataset's cross-validation folds

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Expert	10	10	10	9	10
Novice (Nonexpert)	10	10	10	9	9
Total duration (s)	749.75	750.78	751.57	753.61	752.76

We minimized within-fold duration variances and ensured balanced distribution of expert and novice samples across folds

ResNet50 and transformer with a spatial attention module (Transformer-ATT), and ResNet50 and transformer without a spatial attention module (Transformer-no-ATT). The ResNet50 pretrained on Imagenet [20] was frozen for training. The transformer was not pretrained. An NVIDIA RTX A100 GPU was used for our experiments.

For all models, a batch size of 2 and 100 training epochs were employed. We used the Adam optimizer with an initial learning rate of 0.001. As validation loss plateaued, the learning rate was reduced by a factor of 10. We set D_i to be 2048 and D_h to be 1024. The dimension of $W_{overall}$, W_{frame}, and W_{hidden} was (1024, 1024), (2048, 1024), and (1024, 1), respectively. The final linear classifier layer was followed by a sigmoid function. We chose λ of MMD loss term to be 0.01. The same hyperparameters were used for all models. Twenty labeled target samples were randomly selected and trained with source samples for Vanilla SSDA, Group-DRSL, and Weighted-Group-DRSL methods. The learnable adversarial parameter $w(k = source)$ was trained every five epochs.

Evaluation metrics

We estimated accuracy, sensitivity (recall), specificity, and the area under the receiver operating characteristic curve (AUC) and computed 95% confidence intervals.

Results

Within-dataset evaluation

Table 2 shows model performance within each dataset, i.e., training and testing on the same dataset (fivefold cross-validation). For D99, our work replicates findings in Hira et al. [5]. Our findings in Cataract-101, which have never been reported before, both LSTM-ATT and Transformer-no-ATT showed similar estimates of sensitivity, specificity, and AUC. However, the estimates were lower than those for the Transformer-no-ATT model on D99. Furthermore, the AUC with a transformer was higher than that with LSTM in D99 but not in Cataract-101. These discrepancies suggest that classifying skill with videos is more difficult in Cataract-101 than in D99.

Between-dataset evaluation

Table 3 shows performance of models trained on one dataset and tested on the other.

Train on D99 and test on Cataract-101

For the LSTM-ATT model using Weighted-Group-DRSL led to the best sensitivity, specificity, and AUC. However, Group-DRSL led to better adaptation for LSTM-no-ATT and Transformer-ATT. For the Transformer-no-ATT model, none of the methods improved model performance. While specificity of the Transformer-no-ATT improved with Vanilla SSDA compared with no adaptation, the sensitivity worsened.

Attention improved model's sensitivity. For both Weighted-Group-DRSL on LSTM and Group-DRSL on transformer models, ablating attention increased false negative.

The best LSTM model (LSTM-ATT with Weighted-Group-DRSL) had higher specificity (fewer false positive predictions) and AUC than the best transformer model (Transformer-ATT with Group-DRSL). Both models had similar sensitivity.

Train on Cataract-101 and test on D99

For the LSTM with and without attention, Group-DRSL and Vanilla SSDA led to the best of sensitivity, specificity, and AUC, respectively. While Group-DRSL led to the best Transformer-no-ATT model, MMD, Vanilla SSDA, and Group-DRSL led to comparable models with Transformer-ATT. Group-DRSL improved specificity and reduced sensi-

Table 2 Within-dataset model performance

Model	D99 → D99				Cataract-101 → Cataract-101			
	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
LSTM-ATT	0.73 (0.57–0.89)	0.76 (0.63–0.86)	0.69 (0.55–0.80)	0.78 (0.69–0.87)	0.73 (0.64–0.81)	0.73 (0.60–0.84)	0.73 (0.60–0.84)	0.76 (0.65–0.86)
LSTM-no-ATT	0.81 (0.72-0.87)	0.78 (0.65–0.88)	0.83 (0.70–0.91)	0.87 (0.80–0.94)	0.68 (0.58–0.76)	0.73 (0.60–0.84)	0.62 (0.48–0.75)	0.70 (0.59–0.80)
Transformer-ATT	0.78 (0.69-0.85)	0.76 (0.63-0.86)	0.79 (0.66–0.88)	0.85 (0.77–0.93)	0.69 (0.59–0.77)	0.63 (0.49–0.75)	0.75 (0.61–0.85)	0.70 (0.60–0.81)
Transformer-no-ATT	0.82 (0.73–0.88)	0.82 (0.70–0.90)	0.81 (0.68–0.90)	0.92 (0.87–0.97)	0.72 (0.63–0.80)	0.73 (0.60–0.84)	0.71 (0.57–0.82)	0.77 (0.67–0.86)

Estimates of performance and 95% confidence intervals for models trained and tested on the same dataset. The models are trained on the dataset shown to the left of the → and tested on the dataset shown to its right

tivity, while MMD and Vanilla SSDA improved sensitivity but not specificity.

The effect of attention was not uniform across models and domain adaptation methods. With the LSTM, ablating attention reduced sensitivity and improved specificity for Group-DRSL but improved both for Vanilla SSDA. With the transformer, ablating attention reduced sensitivity and improved specificity for MMD and Vanilla SSDA while it improved both for Group-DRSL.

The estimates, particularly for sensitivity, were higher for best performing transformer (Transformer-no-ATT with Group-DRSL) than that for the best LSTM (LSTM-no-ATT with Vanilla SSDA).

Discussion

Overall, our models learnt more generalizable information from Cataract-101 than from D99. Sensitivity, specificity, and AUC were higher for the best model generalizing from Cataract-101 to D99 compared with the best model generalizing from D99 to Cataract-101. This is likely because the transformer provides a better temporal representation than the attention module. It may also be because of differences in the frame sampling interval used for training, which was longer for models trained on Cataract-101 than for models trained on D99. Furthermore, the average duration of videos in Cataract-101 was about a third of that in D99. This means the input to our models trained on Cataract-101 had information on a larger portion of the surgery.

Attention had a disparate effect on generalizing from D99 to Cataract-101 and vice versa. When training the best LSTM model (LSTM-ATT with Weighted-Group-DRSL) on D99 and testing on Cataract-101, attention significantly improved sensitivity but only slightly affected specificity. For the same model trained on Cataract-101 and tested on D99, attention reduced estimates of both sensitivity and specificity. When training the best transformer model (Transformer-ATT with Group-DRSL) on Cataract-101 and testing on D99, attention significantly reduced sensitivity and slightly affected specificity. For the same model trained on D99 and tested on Cataract-101, attention significantly improved sensitivity but significantly reduced specificity. From our observations, the effect of attention on generalizability of our models is not clear. Experiments with additional and larger datasets are necessary to explain the effect of attention on generalizability of networks for VBA of surgical skill.

Replacing the temporal attention module and LSTM with a transformer generally improved sensitivity and reduced specificity when adapting from D99 to Cataract-101 (except for Weighted-Group-DRSL). However, when trained on Cataract-101 and tested on D99, the transformer improved sensitivity for models using MMD and Vanilla SSDA but had the opposite effect for models using Group-DRSL and Weighted-Group-DRSL. This increase in the number of false negatives with the transformer may be due to the difference in the interval at which we sampled frames in the videos.

MMD did not prove to be effective to generalize models for VBA of surgical skill in our study, particularly with the LSTM. However, with the transformer, it reduced false negatives (improved sensitivity) and increased false positives (lowered specificity) in most cases. This effect was considerably larger when combined with spatial attention, which is expected because we applied MMD to features extracted from the CNN.

Prior research in VBA of surgical skill using operating room videos is limited to training and testing with the same dataset. Our work extends previous work by training and testing on different datasets. Our findings suggest that some domain adaptation methods, such as Group-DRSL, improve generalizability, but future research should emphasize creating and using large datasets and pretrained vision models.

Table 3 Between-dataset model performance

		D99 → Cataract-101				Cataract-101 → D99			
Model	Domain adaptation	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
LSTM-ATT	None	0.40	0.43	0.37	0.37	0.50	0.00	1.00	0.30
		(0.29–0.53)	(0.27–0.61)	(0.22–0.56)	(0.22–0.52)	(0.38–0.62)	(0.00–0.12)	(0.88–1.00)	(0.16–0.44)
	MMD	0.53	1.00	0.00	0.52	0.60	0.24	0.97	0.78
		(0.40–0.65)	(0.89–1.00)	(0.00–0.12)	(0.36–0.67)	(0.47–0.72)	(0.12–0.42)	(0.83–0.99)	(0.65–0.90)
	Vanilla SSDA	0.53	0.58	0.46	0.56	0.48	0.34	0.62	0.54
		(0.40–0.65)	(0.41–0.74)	(0.29–0.65)	(0.40–0.72)	(0.36–0.61)	(0.20–0.53)	(0.44–0.77)	(0.38–0.69)
	Group-DRSL	0.63	0.55	0.73	0.67	0.64	0.69	0.59	0.61
		(0.50–0.74)	(0.38–0.71)	(0.54–0.86)	(0.53–0.81)	(0.51–0.75)	(0.51–0.83)	(0.41–0.74)	(0.46–0.76)
	Weighted-Group-DRSL	0.70	0.65	0.77	0.74	0.57	0.31	0.83	0.68
		(0.57–0.80)	(0.47–0.79)	(0.58–0.89)	(0.61–0.87)	(0.44–0.69)	(0.17–0.49)	(0.65–0.92)	(0.54–0.83)
LSTM-no-ATT	None	0.37	0.23	0.54	0.33	0.43	0.34	0.52	0.39
		(0.26–0.50)	(0.11–0.40)	(0.35–0.71)	(0.19–0.48)	(0.31–0.56)	(0.20–0.53)	(0.34–0.69)	(0.24–0.55)
	MMD	0.53	1.00	0.00	0.36	0.50	1.00	0.00	0.50
		(0.40–0.65)	(0.89–1.00)	(0.00–0.12)	(0.21–0.50)	(0.38–0.62)	(0.88–1.00)	(0.00–0.12)	(0.38–0.62)
	Vanilla SSDA	0.44	0.29	0.62	0.39	0.72	0.66	0.79	0.75
		(0.32–0.57)	(0.16–0.47)	(0.43–0.78)	(0.24–0.54)	(0.60–0.82)	(0.47–0.80)	(0.62–0.90)	(0.62–0.88)
	Group-DRSL	0.60	0.48	0.73	0.61	0.72	0.48	0.97	0.82
		(0.47–0.71)	(0.32–0.65)	(0.54–0.86)	(0.46–0.76)	(0.60–0.82)	(0.31–0.66)	(0.83–0.99)	(0.71–0.93)
	Weighted-Group-DRSL	0.53	0.35	0.73	0.53	0.71	0.48	0.93	0.84
		(0.40–0.65)	(0.21–0.53)	(0.54–0.86)	(0.37–0.68)	(0.58–0.81)	(0.31–0.66)	(0.78–0.98)	(0.74–0.95)

Table 3 continued

		D99 → Cataract-101				Cataract-101 → D99			
Model	Domain adaptation	Accuracy	Sensitivity	Specificity	AUC	Accuracy	Sensitivity	Specificity	AUC
Transformer-ATT	None	0.40	0.65	0.12	0.44	0.60	0.21	1.00	0.65
		(0.29–0.53)	(0.47–0.79)	(0.04–0.29)	(0.28–0.59)	(0.47–0.72)	(0.10–0.38)	(0.88–1.00)	(0.51–0.80)
	MMD	0.46	0.80	0.07	0.28	0.69	0.79	0.59	0.75
		(0.33–0.58)	(0.63–0.90)	(0.02–0.23)	(0.14–0.41)	(0.56–0.79)	(0.62–0.90)	(0.41–0.74)	(0.62–0.88)
	Vanilla SSDA	0.56	0.67	0.44	0.61	0.69	0.72	0.66	0.75
		(0.43–0.68)	(0.49–0.81)	(0.28–0.63)	(0.46–0.76)	(0.56–0.79)	(0.54–0.85)	(0.47–0.80)	(0.62–0.88)
	Group-DRSL	0.61	0.68	0.54	0.66	0.69	0.59	0.79	0.72
		(0.48–0.73)	(0.50–0.81)	(0.35–0.71)	(0.51–0.80)	(0.56–0.79)	(0.41–0.74)	(0.62–0.90)	(0.59–0.86)
	Weighted-Group-DRSL	0.49	0.13	0.92	0.63	0.50	0.00	1.00	0.73
		(0.37–0.62)	(0.05–0.29)	(0.76–0.98)	(0.48–0.78)	(0.38–0.62)	(0.00–0.12)	(0.88–1.00)	(0.60–0.87)
Transformer-no-ATT	None	0.44	0.48	0.38	0.37	0.48	0.00	0.97	0.34
		(0.32–0.57)	(0.32–0.65)	(0.22–0.57)	(0.22–0.52)	(0.36–0.61)	(0.00–0.12)	(0.83–0.99)	(0.20–0.48)
	MMD	0.47	0.00	1.00	0.47	0.55	0.28	0.83	0.58
		(0.35–0.60)	(0.00–0.11)	(0.88–1.00)	(0.32–0.63)	(0.42–0.67)	(0.15–0.46)	(0.65–0.92)	(0.43–0.73)
	Vanilla SSDA	0.47	0.26	0.73	0.56	0.57	0.45	0.69	0.69
		(0.35–0.60)	(0.14–0.43)	(0.54–0.86)	(0.39–0.72)	(0.44–0.69)	(0.28–0.62)	(0.51–0.83)	(0.55–0.83)
	Group-DRSL	0.44	0.13	0.81	0.51	0.78	0.72	0.83	0.84
		(0.32–0.57)	(0.05–0.29)	(0.62–0.91)	(0.36–0.67)	(0.65–0.86)	(0.54–0.85)	(0.65–0.92)	(0.74–0.95)
	Weighted-Group-DRSL	0.46	0.00	1.00	0.53	0.71	0.52	0.90	0.84
		(0.33—0.58)	(0.00–0.11)	(0.87–1.00)	(0.38–0.69)	(0.58–0.81)	(0.34–0.69)	(0.74–0.96)	(0.73–0.95)

Estimates of performance and 95% confidence intervals for models with different domain adaptation approaches. The models are trained on the dataset shown to the left of the → and tested on the dataset shown to its right. The LSTM-ATT model using Weighted-Group-DRSL had the best sensitivity, specificity, and AUC for D99 → Cataract-10. The Transformer-no-ATT model using Group-DRSL had the best sensitivity, specificity, and AUC for Cataract-101 → D99

Conclusion

In conclusion, our findings provide baseline estimates for generalizability of models for VBA of surgical skill between the D99 and Cataract-101 datasets. Group-DRSL improved model generalizability, but the effect of attention modules on model generalization is unclear.

Funding This research was supported by a grant from the National Institutes of Health, USA; R01EY033065. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declarations

Competing interests The authors declare no competing interests.

References

1. Birkmeyer JD, Finks JF, O'reilly A, Oerline M, Carlin AM, Nunn AR, Dimick J, Banerjee M, Birkmeyer NJ (2013) Surgical skill and complication rates after bariatric surgery. N Engl J Med 369(15):1434–1442
2. Curtis NJ, Foster JD, Miskovic D, Brown CS, Hewett PJ, Abbott S, Hanna GB, Stevenson AR, Francis NK (2020) Association of surgical skill assessment with clinical outcomes in cancer surgery. JAMA Surg 155(7):590–598
3. Pryor AD, Lendvay T, Jones A, Ibáñez B, Pugh C (2023) An American board of surgery pilot of video assessment of surgeon technical performance in surgery. Ann Surg 277(4):591–595
4. Feldman LS, Pryor AD, Gardner AK, Dunkin BJ, Schultz L, Awad MM, Ritter EM (2020) Sages video-based assessment (VBA) program: a vision for life-long learning for surgeons. Surg Endosc 34:3285–3288
5. Hira S, Singh D, Kim TS, Gupta S, Hager G, Sikder S, Vedula SS (2022) Video-based assessment of intraoperative surgical skill. Int J Comput Assist Radiol Surg 17(10):1801–1811
6. Liu D, Jiang T, Wang Y, Miao R, Shan F, Li Z (2019) Surgical skill assessment on in-vivo clinical data via the clearness of operating field. In: Medical Image computing and computer assisted intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22. Springer, pp. 476–484
7. Wagner M, Müller-Stich B-P, Kisilenko A, Tran D, Heger P, Mündermann L, Lubotsky DM, Müller B, Davitashvili T, Capek M (2023) Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. Med Image Anal 86:102770
8. Paranjape JN, Sikder S, Patel VM, Vedula SS (2023) Cross-dataset adaptation for instrument classification in cataract surgery videos. In: International conference on medical image computing and computer-assisted intervention, Springer, pp 739–748
9. Schölkopf B, Platt J, Hofmann T (2007) A kernel method for the two-sample-problem. In: Advances in Neural Information Processing Systems
10. Wan B, Peven M, Hager G, Sikder S, Vedula SS (2024) Spatial-temporal attention for video-based assessment of intraoperative surgical skill. Sci Rep 14(1):26912
11. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
12. Bahdanau D, Cho KH, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd international conference on learning representations, ICLR 2015
13. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short Papers), pp 207–212
14. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N (2020) An image is worth 16 x 16 words: Transformers for image recognition at scale. In: International conference on learning representations
15. Hu W, Niu G, Sato I, Sugiyama M (2018) Does distributionally robust supervised learning give robust classifiers? In: International conference on machine learning, PMLR, pp 2029–2037
16. Sugiyama M, Storkey AJ (2006) Mixture regression for covariate shift. In: Advances in neural information processing systems, vol 19
17. Kim TS, O'Brien M, Zafar S, Hager GD, Sikder S, Vedula SS (2019) Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. Int J Comput Assist Radiol Surg 14:1097–1105
18. Schoeffmann K, Taschwer M, Sarny S, Münzer B, Primus MJ, Putzgruber D (2018) Cataract-101: video dataset of 101 cataract surgeries. In: Proceedings of the 9th ACM multimedia systems conference, pp 421–425
19. Buslaev A, Iglovikov VI, Khvedchenya E, Parinov A, Druzhinin M, Kalinin AA (2020) Albumentations: fast and flexible image augmentations. Information 11(2):125
20. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115:211–252

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.