Revisiting Distillation for Continual Learning on Visual Question Localized-Answering in Robotic Surgery

Long Bai[1], Mobarakol Islam[2], and Hongliang Ren[1,3](✉)

[1] Department of Electronic Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong SAR, China
b.long@link.cuhk.edu.hk

[2] Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, London, UK
mobarakol.islam@ucl.ac.uk

[3] Shun Hing Institute of Advanced Engineering, CUHK, Hong Kong SAR, China
hlren@ee.cuhk.edu.hk

Abstract. The visual-question localized-answering (VQLA) system can serve as a knowledgeable assistant in surgical education. Except for providing text-based answers, the VQLA system can highlight the interested region for better surgical scene understanding. However, deep neural networks (DNNs) suffer from catastrophic forgetting when learning new knowledge. Specifically, when DNNs learn on incremental classes or tasks, their performance on old tasks drops dramatically. Furthermore, due to medical data privacy and licensing issues, it is often difficult to access old data when updating continual learning (CL) models. Therefore, we develop a non-exemplar continual surgical VQLA framework, to explore and balance the rigidity-plasticity trade-off of DNNs in a sequential learning paradigm. We revisit the distillation loss in CL tasks, and propose rigidity-plasticity-aware distillation (RP-Dist) and self-calibrated heterogeneous distillation (SH-Dist) to preserve the old knowledge. The weight aligning (WA) technique is also integrated to adjust the weight bias between old and new tasks. We further establish a CL framework on three public surgical datasets in the context of surgical settings that consist of overlapping classes between old and new surgical VQLA tasks. With extensive experiments, we demonstrate that our proposed method excellently reconciles learning and forgetting on the continual surgical VQLA over conventional CL methods. Our code is publicly accessible at github.com/longbai1006/CS-VQLA.

L. Bai and M. Islam—Co-first authors.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_7.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 68–78, 2023.
https://doi.org/10.1007/978-3-031-43996-4_7

1 Introduction

Trustworthy and reliable visual question-answering (VQA) models have proved their potential in the medical domain [17,22]. A deep learning (DL)-based surgical VQA system [22] has been developed as a surgical training and popularization tool for junior surgeons, medical students, and patients. However, one pivotal problem with surgical VQA is the lack of localized answers. VQA can provide the answer to the question, but cannot relate the answers to its localization at an instance level. Surgical scenarios with various similar instruments and actions may further confuse the learners. Answers with localization can further assist learners in dealing with confusion. In this case, a surgical visual-question localized-answering (VQLA) system can thereby be established for effective surgical training and scene understanding [3].

Meanwhile, catastrophic forgetting has become a largely discussed topic in deep neural networks. Deep neural networks (DNNs) shall abruptly and drastically forget old knowledge when learning new [16]. Various continual learning (CL) methods have been proposed to mitigate catastrophic forgetting and study the balance of rigidity and plasticity in deep models [16,20]. Rigidity refers to the ability of the model not to diverge and remember old knowledge, while plasticity represents the acquisition of new knowledge by DNNs [4]. Some pioneering works have attempted to tackle the CL problem in the medical domain [6]. Catastrophic forgetting may occur in various real-world medical scenarios, e.g., data collected (i) over time, and (ii) across devices/institutions. More seriously, due to issues of data privacy, storage, and licensing, old data may not be accessible anymore [13]. Therefore, it is necessary to develop a non-exemplar CL method for surgical VQLA tasks to resist catastrophic forgetting in clinical applications.

Furthermore, most medical decision-making tasks shall include classes overlapping with the old tasks and newly appeared classes, as shown in Fig. 1. We should not distillate the entire previous model when we deal with CL with overlapping classes. Firstly, the model will not emphasize new classes and have a high bias toward overlapping classes rather than new classes. Overlapping classes will dominate the model prediction if we naively follow the distillation from existing CL models. Secondly, catastrophic forgetting will be severe in old non-overlapping classes and the overlapping classes will dominate in the model prediction, and forget the old classes. For this purpose, we revisit distillation methods in CL and design a Continual Surgical VQLA (CS-VQLA) framework for learning incremental classes by balancing the performance of the old overlapping and non-overlapping classes. CS-VQLA has the following attributes: (i) it is a multi-task model including answering and localization, (ii) domain shift and class increment problems both exist, (iii) there may be overlapping classes between old and new tasks. These points shall further complicate the CL tasks.

In this work, **(1)** We establish a non-exemplar CS-VQLA framework. While being applied to surgical education and scene understanding, the framework can learn data in a streaming manner and effectively resist catastrophic forgetting. **(2)** We revisit the distillation method for CL, and propose rigidity-plasticity-aware distillation (RP-Dist) and self-calibrated heterogeneous distilla-

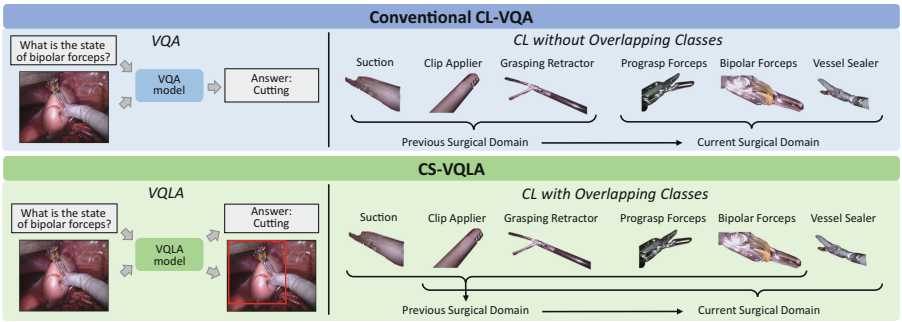

Fig. 1. Comparison between conventional CL-VQA and our CS-VQLA. Besides providing localized answers, our CS-VQLA framework also pays attention to the issue of overlapping and non-overlapping classes in sequential surgical domains.

tion (SH-Dist) for the output logits and intermediate feature maps, respectively. The weight aligning (WA) technique is further integrated to adjust model bias between old and new data. **(3)** Extensive comparison and ablation studies prove the outstanding performance of our method in mitigating catastrophic forgetting, demonstrating its potential in real-world applications.

2 Methodology

2.1 Preliminaries

Problem Definition. We define the continual learning sequence with $\mathcal{TP}$ time periods, and $t \in \{1, ..., \mathcal{TP}\}$ means the current time period. $\mathcal{D}_t$ denotes the training dataset at time period t, with x representing a sample of the input question and image pair in $\mathcal{D}_t$. $\mathcal{C}_{old}$ denotes the classes appearing in previous time period $\{1, ..., t-1\}$, and $\mathcal{C}_{new}$ represents the classes appearing in current time period t. Furthermore, we define the classes existing in both $\mathcal{C}_{old}$ and $\mathcal{C}_{new}$ as *overlapping classes* $\mathcal{C}_{op}$, and define unique classes in $\mathcal{C}_{old}$ as *old non-overlapping classes* $\mathcal{C}_{no}$. F stands for the output feature map from the network backbone.

Knowledge Distillation (KD) [9,26] on output logits [16] or intermediate feature map [19] is a widely used approach to retain knowledge on old tasks. With z^o and z^{cl} denote the output logits from the old and CL model, respectively, we can formulate the logits distillation loss [16] as:

$$\mathcal{L}_{LKD} = \sum_{c=0}^{\mathcal{C}_{old}} -p_T^o(x) \log\left(p_T^{cl}(x)\right) \tag{1}$$

in which $p_T^o(x) = SM(z^o/T)$ and $p_T^{cl}(x) = SM(z^{cl}/T)$ represent the probabilities. SM means Softmax. T is temperature normalization for all old classes.

Weight Aligning (WA) [27] is a simple technique to align the weight bias in the classifier layer. We use $\mathbf{W}_{new}$ to represent the weights for newly appeared classes in the classifier, and $\mathbf{W}_{old}$ to denote those of old classes, then we have:

$$\hat{\mathbf{W}}_{new} = \frac{Mean[Norm(\mathbf{W}_{old})]}{Mean[Norm(\mathbf{W}_{new})]} \cdot \mathbf{W}_{new} \tag{2}$$

where *norm* means normalizing all the elements in the vector. In class-incremental learning, WA can effectively avoid the model bias towards new classes.

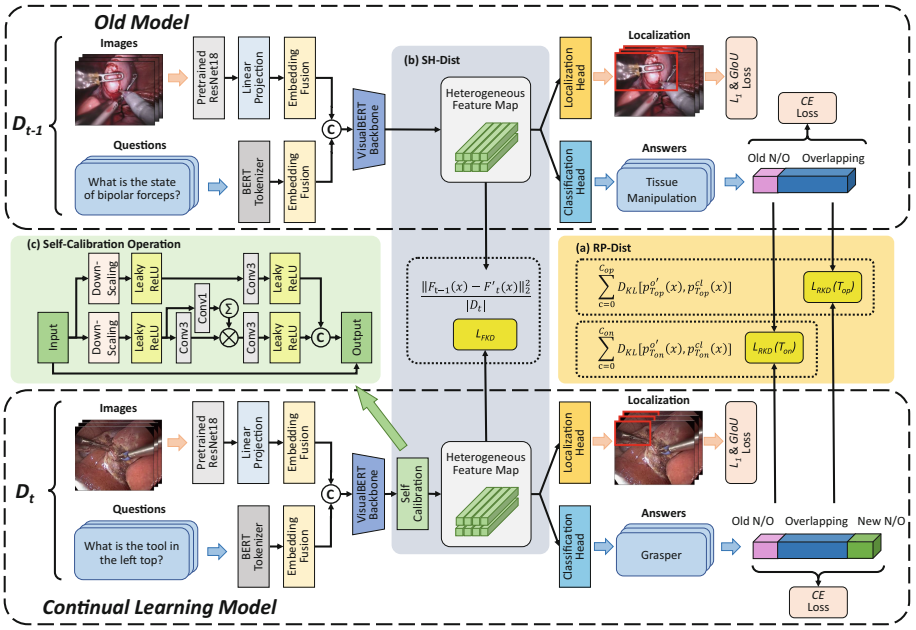

Fig. 2. Overview of our CS-VQLA network. The VQLA model is used to process bimodal input (image and text) and provide predictions for two tasks (answering and localization). The proposed RP-Dist and SH-Dist are designed to help the CL model retain old knowledge from the old model and trade-off model rigidity-plasticity. 'N/O' means non-overlapping classes.

2.2 Continual Surgical VQLA (CS-VQLA)

Visual-Question Localized Answering. We define our VQLA framework following [3], by building a parallel detector on top of the VQA-based classification model. Therefore, the VLQA model includes the following components: a ResNet18 [8] pre-trained on ImageNet [5] as a prior image feature extractor, a BERT tokenizer [7], the VisualBERT [15] as the backbone (it can also

be called as the deep feature extractor), a fully-connected layer as the classifier, and a 3-layer MLP as the detector. The classification task is optimized via the cross-entropy loss $\mathcal{L}_{CE}$, and the bounding box regression is optimized by the sum of $\mathcal{L}_1$ and $GIoU$ loss [21]. Thus, the VQLA loss can be formulated as: $\mathcal{L}_{VQLA} = \mu \cdot \mathcal{L}_{CE} + (\mathcal{L}_1 + \mathcal{L}_{GIoU})$, where μ is set as 100 to balance the optimization progress of the two tasks.

Rigidity-Plasticity-Aware Distillation (RP-Dist). The current rigidity-plasticity trade-off is towards the entire model. However, we shall make the rigidity-plasticity aware in overlapping and non-overlapping classes.

There is no overlap between $\mathcal{C}_{old}$ and $\mathcal{C}_{new}$ in an ideal class-incremental learning setup, so the temperature T in Equ. 1 is set to 2 by [16]. However, in a real-world application setup, T should not smooth the logits equally for old non-overlapping $\mathcal{C}_{on}$ and overlapping classes $\mathcal{C}_{op}$. Specifically, through adjusting for T, we shall endow the model greater plasticity on C_{op}, and keep the rigidity on C_{on}. We first establish a regularized distillation loss. Originally, the old model shall serve as the 'teacher' model in CL-based distillation. Instead of directly distilling the old model output logits, we construct a perfect pseudo teacher for distillation. To begin with, a pseudo answering label set a' can be built from the old model classification probabilities $p^o(x)$ via $a' = Max[p(x)]$. Based on the idea of label smoothing, we can manually setup a pseudo old model to have a high probability of predicting a correct class, and its probability distribution shall be:

$$p^{o\prime}(x) = \begin{cases} \lambda & \text{if } x = a' \\ (1-\lambda)/(\mathcal{C}_{old} - 1) & \text{if } x \neq a' \end{cases} \tag{3}$$

When λ is set to a very high number (e.g., $\lambda \geq 0.9$), we will have a high probability of getting a correct class, allowing the teacher model to have perfect performance. The probability output of the CL model can be optimized with this pseudo-teacher based on the Kullback-Leibler divergence D_{KL}:

$$\mathcal{L}_{RKD} = \sum_{c=0}^{\mathcal{C}_{old}} D_{KL}\left[p_T^{o'}(x), p_T^{cl}(x)\right] \tag{4}$$

T is the KD temperature used to generate soft probabilities for the pseudo old model. As discussed above, this naive setting of T is not suitable for general CL scenarios. Therefore, we treat T_{op} and T_{on} differently to strengthen the plasticity on $\mathcal{C}_{op}$ and the rigidity on $\mathcal{C}_{on}$ respectively. $\mathcal{L}_{RKD}$ can thereby be rewritten as:

$$\mathcal{L}_{RKD} = \sum_{c=0}^{\mathcal{C}_{op}} D_{KL}\left[p_{T_{op}}^{o'}(x),\, p_{T_{op}}^{cl}(x)\right] + \sum_{c=0}^{\mathcal{C}_{on}} D_{KL}\left[p_{T_{on}}^{o'}(x),\, p_{T_{on}}^{cl}(x)\right] \tag{5}$$

We keep $T_{op} > T_{on}$ to balance the rigidity and plasticity trade-off in the CL model, and set $T_{op} = 25$, $T_{on} = 20$ empirically in our implementation.

Self-Calibrated Heterogeneous Distillation (SH-Dist). Works have discussed the use of self-calibration to improve model performance [18,32]. However, assuming we obtain an old model and we would like to conduct CL training on it, we can hardly modify the old model itself directly. Therefore, we perform a self-calibration operation on the heterogeneous output features F_t from the VisualBERT backbone to get self-calibrated feature F_t'. The details can be referred to at the bottom of Fig. 2. Without engaging more learnable parameters, we endow the heterogeneous features with adaptively modeled long-range context information. Therefore, we can construct our feature distillation using the self-calibrated feature map F_t' and the old model feature map F_{t-1}. $\mathcal{L}_2$ loss is used to minimize the distance between F_t' and F_{t-1} empirically by following [19]:

$$\mathcal{L}_{FKD} = \frac{\|F_{t-1}(x) - F_t'(x)\|_2^2}{|\mathcal{D}_t|} \tag{6}$$

Subsequently, the self-calibrated feature map F_t' shall be propagated through the parallel classifier and detector for the multi-task prediction.

Overall Framework. Figure 2 shows the overview of our CS-VQLA framework. The given image and question input are respectively processed as feature embedding by pre-trained ResNet18 and BERT tokenier, and fed to the VisualBERT backbone after embedding fusion. Then the output heterogeneous feature map is used to train the parallel predictors. The loss functions establish the essential components of our CS-VQLA framework. In the initial time period $t = 0$, the model is only trained on the VQLA loss. When $t > 0$, we combine the VQLA loss for training on the current dataset D_t, with the RP-Dist & SH-Dist loss to retain the old knowledge. We can summarize our final loss function as follows:

$$\mathcal{L} = \begin{cases} \mathcal{L}_{VQLA} & t = 0 \\ \alpha \cdot \mathcal{L}_{VQLA} + \beta \cdot \mathcal{L}_{RKD} + \gamma \cdot \mathcal{L}_{FKD} & t > 0 \end{cases} \tag{7}$$

We set $\alpha = \beta = 1$ and $\gamma = 5$ in our implementation. Furthermore, WA is deployed after training on each time period, to balance the weight bias of new classes on the classification layer. Through the combination of multiple distillation paradigms and model weight adjustment, we successfully realize the general continual learning framework in the VQLA scenario.

3 Experiments

3.1 Dataset and Setup

Dataset. We construct our continual procedure as follows: when $t = 0$, we train on EndoVis18 Dataset, $t = 1$ on EndoVis17 Dataset, and $t = 2$ on M2CAI Dataset. Therefore, we can establish our CS-VQLA framework with a large initial step, and several smaller sequential steps. When splitting the dataset, we isolate the training and test sets in different sequences to avoid information leakage.

EndoVis18 Dataset is a public dataset with 14 videos on robotic surgery [1]. The question-answer (QA) pairs are accessible in [22], and the bounding box annotations are from [10]. The answers are in single-word form with three categories (organ, interaction, and locations). We further extend the QA pairs and include cases when the answer is a surgical tool. Besides, if the answer is regarding the organ-tool interaction, the bounding box shall contain both the organ and the tool. Statistically, the training set contains 1560 frames with 12741 QA pairs, and the test set contains 447 frames with 3930 QA pairs.

EndoVis17 Dataset is a public dataset with 10 videos on robotic surgery [2]. We randomly select frames and manually annotate the QA pairs and bounding boxes. The training set contains 167 frames with 1034 QA pairs, and the test set contains 40 frames with 201 QA pairs.

M2CAI Dataset is also a public robotic surgery dataset [24,25], and the location bounding box is publicly accessible in [11]. Similarly, we randomly select 167

Table 1. Comparison experiments from the time period $t = 0$ to $t = 1$. **Bold** and underlined represent best and second best, respectively. 'W/O' denotes 'without', and 'W/I' denotes 'within'. 'N/O' means non-overlapping. 'Old N/O' represents the classes that exist in $t = 0$ but do not exist in $t = 1$, and 'New N/O' represents the opposite. 'Overlapping' denotes the classes that exist in both $t = 0, 1$.

$t = 1$	Methods	Old N/O		Overlapping		New N/O		EndoVis18		EndoVis17		Average	
		Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU
W/O CL	Base ($t = 0$)	6.11	62.12	64.60	75.68	✗	✗	62.65	75.23	✗	✗	✗	✗
	FT	0.00	62.55	38.07	71.88	86.96	80.41	34.86	71.29	81.59	78.30	58.23	74.79
W/I CL	LwF [16]	0.00	**63.40**	54.36	69.94	73.91	77.47	53.20	69.53	43.79	74.64	48.50	72.08
	WA [27]	0.76	60.70	55.11	71.61	52.17	78.10	52.85	71.02	63.68	76.71	58.26	73.86
	iCaRL [20]	0.76	62.23	55.87	72.36	43.48	**79.51**	53.85	71.76	58.21	**78.41**	56.03	75.08
	IL2A [29]	0.00	57.74	53.00	69.88	56.52	78.20	51.48	69.23	48.76	75.75	50.12	72.49
	PASS [30]	0.00	56.49	54.01	70.08	69.57	77.60	51.70	69.46	65.67	74.56	58.69	72.01
	SSRE [31]	0.00	60.29	54.04	70.34	65.22	76.07	51.76	69.78	64.68	75.44	58.22	72.61
	CLVQA [14]	0.00	59.87	51.83	72.98	65.22	78.36	49.14	72.40	72.14	76.40	60.64	74.40
	CLiMB [23]	0.00	60.16	52.88	72.99	69.57	77.37	50.13	72.44	74.13	75.87	62.13	74.16
	Ours	**1.53**	61.08	**56.98**	**74.57**	**78.26**	78.59	**54.33**	**74.02**	**75.12**	77.02	**64.73**	**75.52**

Table 2. Comparison experiments from the time period $t = 0, 1$ to $t = 2$. 'Old N/O' represents the classes that exist in $t = 0, 1$ but do not exist in $t = 2$, and 'New N/O' represents the opposite. 'Overlapping' denotes the classes that exist in $t = 0, 1, 2$.

$t = 2$	Methods	Old N/O		Overlapping		New N/O		EndoVis18		EndoVis17		M2CAI16		Average	
		Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU
W/O CL	FT	4.00	60.90	19.08	58.69	55.56	70.83	15.57	58.75	41.79	60.18	51.06	69.48	36.14	62.80
W/I CL	LwF [16]	4.20	62.91	**42.75**	63.34	27.78	72.68	38.04	63.25	41.29	**62.71**	31.91	69.65	37.08	65.20
	WA [27]	6.80	59.32	40.62	61.55	**55.56**	73.06	36.67	61.20	36.32	60.86	40.43	69.80	37.81	63.96
	iCaRL [20]	2.00	58.55	38.06	58.59	41.67	73.42	33.46	58.30	38.81	61.12	38.30	**71.01**	36.85	63.48
	IL2A [29]	11.00	57.25	33.80	58.50	27.78	72.71	30.61	58.28	37.31	58.29	36.17	67.10	34.70	61.22
	PASS [30]	22.60	56.57	24.80	58.39	30.56	72.07	23.52	58.07	37.81	58.91	**41.49**	66.68	34.27	61.22
	SSRE [31]	13.80	58.12	19.49	57.02	47.22	73.31	18.12	57.00	27.36	58.09	40.43	67.47	28.64	60.85
	CLVQA [14]	21.80	58.01	36.54	62.92	25.00	74.09	34.40	62.35	39.80	61.05	36.17	68.89	36.79	64.10
	CLiMB [23]	23.00	57.03	38.90	64.30	33.33	73.45	36.62	63.50	42.29	61.35	40.43	69.20	39.78	64.68
	Ours	**28.20**	**68.14**	41.04	**65.74**	44.44	**74.41**	**39.13**	**66.21**	**46.77**	62.17	**41.49**	70.04	**42.46**	**66.14**

frames and annotate 449 QA pairs for the training set, and 40 frames with 94 QA pairs in different videos for the test set.

Implementation Details. We compare our solution against the fine-tuning (FT) baseline and state-of-the-art (SOTA) methods, including LwF [16], WA [27], iCaRL [20], IL2A [29], PASS [30], SSRE [31], CLVQA [14], and CLiMB [23]. All the methods are implemented using [28][1], with PyTorch and on NVIDIA RTX 3090 GPU. We removed the exemplars in all methods for a non-exemplar comparison. All methods are firstly trained on EndoVis18 ($t = 0$) for 60 epochs with a learning rate of 1×10^{-5}, and then trained on EndoVis17 ($t = 2$) and M2CAI ($t = 2$) for 30 epochs with a learning rate of 5×10^{-5}. We use Adam optimizer [12] and a batch size of 64. Answering and localization performance are evaluated by Accuracy (Acc) and mean intersection over union (mIoU), respectively.

Table 3. Ablation experiments from the time period $t = 0$ to $t = 1$, and from $t = 1$ to $t = 2$. To observe the contribution of each component, we degenerate the proposed RP-Dist and SH-Dist to the normal distillation paradigm, and remove the WA module.

Methods			$t = 0$ to $t = 1$								$t = 1$ to $t = 2$							
			Old N/O		Overlapping		New N/O		Average		Old N/O		Overlapping		New N/O		Average	
RP	SH	WA	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU	Acc	mIoU
✓	✗	✗	0.00	60.58	55.30	72.94	73.91	77.62	62.66	74.37	11.40	58.29	40.96	64.72	30.56	73.31	40.80	64.85
✗	✓	✗	0.00	59.94	56.23	73.66	**82.61**	78.04	64.33	74.90	9.80	57.17	38.14	65.30	38.89	74.15	40.26	64.35
✗	✗	✓	0.00	60.33	53.08	72.45	52.17	76.18	61.94	74.30	12.20	59.49	39.75	64.81	41.67	72.12	39.07	64.81
✗	✓	✓	0.00	60.97	54.59	74.12	73.91	78.52	61.83	74.89	11.00	65.16	40.69	64.45	41.67	74.18	39.63	65.45
✓	✗	✓	0.00	61.04	56.08	74.11	60.87	78.91	61.60	74.88	11.00	59.10	39.25	62.60	22.22	71.66	39.16	64.55
✓	✓	✗	0.00	60.07	54.56	74.15	73.91	**79.79**	61.81	75.00	10.40	63.32	39.14	64.40	27.78	73.89	40.21	65.65
✓	✓	✓	**1.53**	**61.08**	**56.98**	**74.57**	78.26	78.59	**64.73**	**75.52**	**28.20**	**68.14**	41.04	**65.74**	**44.44**	**74.41**	**42.46**	**66.14**

3.2 Results

Except for testing on three datasets separately, we set three specific categories in our continual learning setup: *old non-overlapping* (old N/O) classes, *overlapping* classes, and *new non-overlapping* (new N/O) classes. By measuring the performance in these three categories, we can easily observe the catastrophic forgetting phenomenon and the performance of mitigating catastrophic forgetting.

As shown in Table 1 & 2, firstly, catastrophic forgetting can be apparently observed in the performance of FT. Then, among all baselines, iCaRL achieves the best performance when the model learns from $t = 0$ to $t = 1$, and gets to forget when there are more time periods. On the contrary, LwF exhibits a strong retention of old knowledge, but a lack of ability to learn new. Our proposed methods demonstrate superior performance in almost all metrics and classes. In classification tasks, the overall average of our methods outperforms the second best with 2.60% accuracy improvement at $t = 1$ and 2.68% at $t = 2$. In localization tasks, our method is 0.44 mIoU higher than the second best at $t = 1$

[1] github.com/G-U-N/PyCIL.

and 0.94 mIoU higher at $t = 2$. The results prove the remarkable ability of our method to balance the rigidity-plasticity trade-off. Furthermore, an ablation study is conducted to demonstrate the effectiveness of each component in our proposed method. We (i) degenerate the RP-Dist to original logits distillation [16], (ii) degenerate the SH-Dist to normal feature distillation [19], and (iii) remove the WA module, as shown in Table 3. Experimental results show that each component we propose or integrate plays an essential role in the final rigidity-plasticity trade-off. Therefore, we demonstrate that each of our components is indispensable. More evaluation and ablation studies can be found in the supplementary materials.

4 Conclusion

This paper introduces CS-VQLA, a general continual learning framework on surgical VQLA tasks. This is a significant attempt to continue learning under complicated clinical tasks. Specifically, we propose the RP-Dist on output logits, and the SH-Dist on the intermediate feature space, respectively. The WA technique is further integrated for model weight bias adjustment. Superior performance on VQLA tasks demonstrates that our method has an excellent ability to deal with CL-based surgical scenarios. Except for giving localized answers for better surgical scene understanding, our solution can conduct continual learning in any questions in surgical applications to solve the problem of class increment, domain shift, and overlapping/non-overlapping classes. Our framework can also be applied when adapting a vision-language foundation model in the surgical domain. Therefore, our solution holds promise for deploying auxiliary surgical education tools across time/institutions. Potential future works also include combining various surgical training systems (e.g., mixed reality-based training, surgical skill assessment) to develop an effective and comprehensive virtual teaching system.

Acknowledgements. This work was funded by Hong Kong RGC CRF C4063-18G, CRF C4026-21GF, RIF R4020-22, GRF 14203323, GRF 14216022, GRF 14211420, NSFC/RGC JRS N_CUHK420/22; Shenzhen-Hong Kong-Macau Technology Research Programme (Type C 202108233000303); Guangdong GBABF #2021B1515120035. M. Islam was funded by EPSRC grant [EP/W00805X/1].

References

1. Allan, M., et al.: 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190 (2020)
2. Allan, M., et al.: 2017 robotic instrument segmentation challenge. arXiv preprint arXiv:1902.06426 (2019)
3. Bai, L., Islam, M., Seenivasan, L., Ren, H.: Surgical-VQLA: transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. arXiv preprint arXiv:2305.11692 (2023)

4. De Lange, M., et al.: A continual learning survey: defying forgetting in classification tasks. IEEE Trans. Pattern Anal. Mach. Intell. **44**(7), 3366–3385 (2021)
5. 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)
6. Derakhshani, M.M., et al.: Lifelonger: a benchmark for continual disease classification. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. LNCS, vol. 13432, pp. 314–324. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16434-7_31
7. 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)
8. 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)
9. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
10. 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
11. Jin, A., et al.: Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. In: IEEE Winter Conference on Applications of Computer Vision (2018)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
13. Lee, C.S., Lee, A.Y.: Clinical applications of continual learning machine learning. Lancet Digit. Health **2**(6), e279–e281 (2020)
14. Lei, S.W., et al.: Symbolic replay: scene graph as prompt for continual learning on VQA task. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 37, pp. 1250–1259 (2023)
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. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Trans. Pattern Anal. Mach. Intell. **40**(12), 2935–2947 (2017)
17. Lin, Z., et al.: Medical visual question answering: a survey. arXiv preprint arXiv:2111.10056 (2021)
18. Liu, J.J., Hou, Q., Cheng, M.M., Wang, C., Feng, J.: Improving convolutional networks with self-calibrated convolutions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10096–10105 (2020)
19. Michieli, U., Zanuttigh, P.: Incremental learning techniques for semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
20. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: iCaRL: incremental classifier and representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2001–2010 (2017)
21. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: a metric and a loss for bounding box regression.

In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 658–666 (2019)
22. Seenivasan, L., Islam, M., Krishna, A., Ren, H.: Surgical-VQA: visual question answering in surgical scenes using transformer. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. LNCS, vol. 13437, pp. 33–43. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-16449-1_4
23. Srinivasan, T., Chang, T.Y., Pinto Alva, L., Chochlakis, G., Rostami, M., Thomason, J.: Climb: a continual learning benchmark for vision-and-language tasks. Adv. Neural Inf. Process. Syst. **35**, 29440–29453 (2022)
24. Stauder, R., Ostler, D., Kranzfelder, M., Koller, S., Feußner, H., Navab, N.: The tum lapchole dataset for the m2cai 2016 workflow challenge. arXiv preprint arXiv:1610.09278 (2016)
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. Yuan, L., Tay, F.E., Li, G., Wang, T., Feng, J.: Revisiting knowledge distillation via label smoothing regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3903–3911 (2020)
27. Zhao, B., Xiao, X., Gan, G., Zhang, B., Xia, S.T.: Maintaining discrimination and fairness in class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13208–13217 (2020)
28. Zhou, D.W., Wang, F.Y., Ye, H.J., Zhan, D.C.: PyCIL: a python toolbox for class-incremental learning (2021)
29. Zhu, F., Cheng, Z., Zhang, X.Y., Liu, C.L.: Class-incremental learning via dual augmentation. Adv. Neural Inf. Process. Syst. **34**, 14306–14318 (2021)
30. Zhu, F., Zhang, X.Y., Wang, C., Yin, F., Liu, C.L.: Prototype augmentation and self-supervision for incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5871–5880 (2021)
31. Zhu, K., Zhai, W., Cao, Y., Luo, J., Zha, Z.J.: Self-sustaining representation expansion for non-exemplar class-incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9296–9305 (2022)
32. Zou, W., Ye, T., Zheng, W., Zhang, Y., Chen, L., Wu, Y.: Self-calibrated efficient transformer for lightweight super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 930–939 (2022)