





CPR-CLIP: Multimodal Pre-training for Composite Error Recognition in CPR Training

Shunli Wang , Dingkang Yang , Peng Zhai , and Lihua Zhang , *Member, IEEE*

Abstract—The expensive cost of the medical skill training paradigm hinders the development of medical education, which has attracted widespread attention in the intelligent signal processing community. To address the issue of composite error action recognition in Cardiopulmonary Resuscitation (CPR) training, this letter proposes a multimodal pre-training framework named CPR-CLIP based on prompt engineering. Specifically, we design three prompts to fuse multiple errors naturally on the semantic level and then align linguistic and visual features via the contrastive pre-training loss. Extensive experiments verify the effectiveness of the CPR-CLIP. Ultimately, the CPR-CLIP is encapsulated to an electronic assistant, and four doctors are recruited for evaluation. Nearly four times efficiency improvement is observed in comparative experiments, which demonstrates the practicality of the system. We hope this work brings new insights to the intelligent medical skill training and signal processing communities simultaneously. Code is available on <https://github.com/Shunli-Wang/CPR-CLIP>.

Index Terms—Human action analysis, cross-modal interaction, action quality assessment, CPR skill training.

I. INTRODUCTION

INTELLIGENT medical skill training systems have received continuous attention in signal processing community in recent years. Statistics [1] show that the mean price of a U.S. medical education was about \$300,000, 75% of students took on loans, and their average debt at graduation was \$200,000. Medical institutions and training centers are highly concerned about improving training efficiency and saving costs, as this affects the quality of medical services nationally.

Fortunately, with the flourishing development of signal processing technologies [2], [3], [4], [5], [6], [7], some algorithms [8], [9], [10], [12], [13], [14], [24] have been introduced into medical skill analysis research. Studies in medical skill analysis mainly focus on surgical skill evaluation [8], [9], [11], [15] and operating skills identification on surgical robot systems [13], [14], [25], [26]. Despite the progress in datasets and algorithms, these methods still face two challenges: on the one hand, most of these models only support skill classification from *Novice/Medium/Expert*, which lacks interpretability in application. On the other hand, these algorithms usually have poor interactivity. The outputs are the scores of each category, making the model unable to become effective assistants. Overall, there is a considerable gap between current medical skill training systems and practical application.

This work was supported in part by the National Key R&D Program of China under Grants 2021ZD0113502, in part by Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0103.

The authors are with the Academy for Engineering and Technology, Fudan University, Shanghai 200433, China (e-mail: slwang19@fudan.edu.cn; dkyang20@fudan.edu.cn; pzhai@fudan.edu.cn; lihuazhang@fudan.edu.cn).

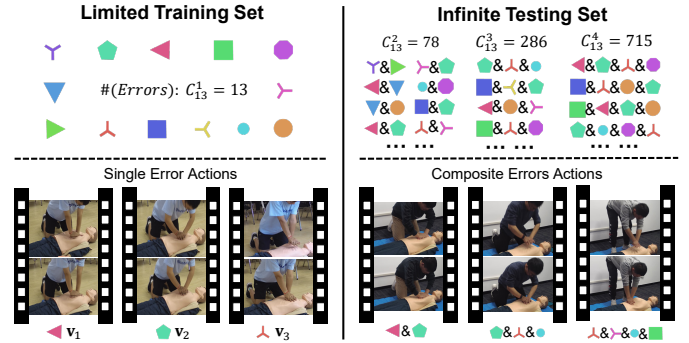


Fig. 1. Different colored marks represent different error actions. There is a significant challenge in recognizing composite error actions with limited single-class samples. The vast divergence in data distribution poses higher requirements of generalization for algorithms.

To address the above issues, this study explores composite error recognition and actual deployment in Cardiopulmonary Resuscitation (CPR) training. CPR is a core life-saving technique for cardiac and respiratory arrest. Inappropriate CPR actions will cause unsatisfactory effects and secondary damage [11]. Subjects usually make multiple mistakes during testing, while the training set cannot contain all combinations of errors, which leads to the composite error recognition task. As shown in Fig. 1, the form of the composite error recognition task is: given a set that only contains single-error samples, the algorithm is required to accurately identify complicated error combinations in application. Restricted supervision conditions exceed the capabilities of traditional action recognition algorithms [16], [17], [18], [19], [20]. Inspired by the Contrastive Language-Image Pre-Training (CLIP) paradigm [27], [29], [28], this letter introduces the prompt engineering [30], [31] into a composite error recognition task and proposes a multi-modal pre-training framework named CPR-CLIP. The motivation is neat and straightforward: aligning language embeddings with augmented visual features through minimizing a multimodal contrastive loss function. This approach fully utilizes the advantage of natural language, which can integrate and describe composite errors smoothly, thus improving the generalization of the model. Extensive experimental results verify the effectiveness brought by the contrastive pre-training.

Unlike previous studies [9], [11], [14] that stopped after verifying the performance of the proposed methods, we leverage the merits of the multi-modal framework and transform it into a real assistant for doctors. Controlled experiments on time-consuming and precision are conducted to verify the effectiveness of the electronic assistant in actual deployment.

In summary, our contributions are as follows:

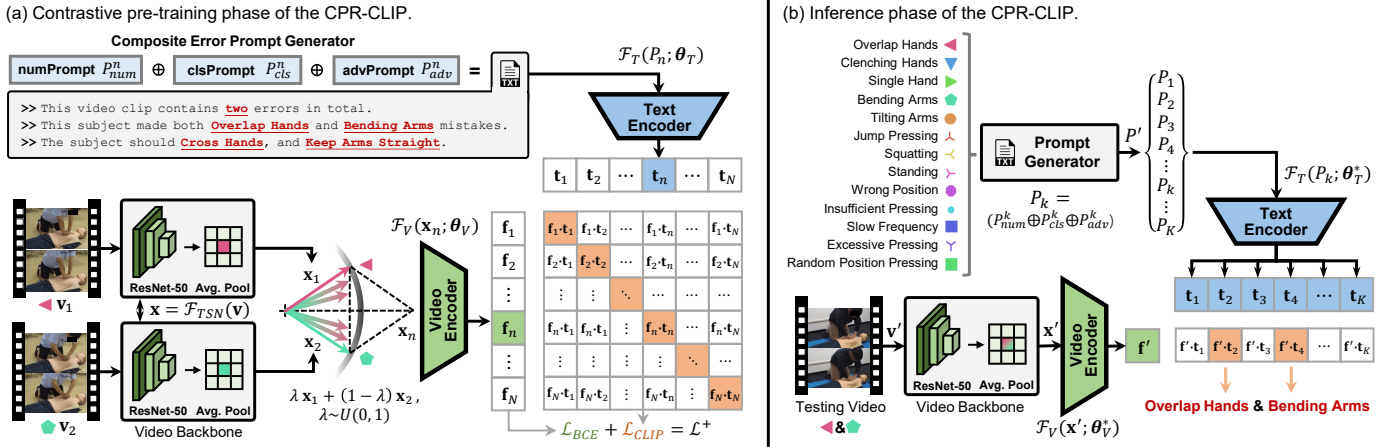


Fig. 2. Architecture of the CPR-CLIP. (a) shows the multi-modal pre-training phase of the CPR-CLIP. The **visual pathway**, **language pathway**, and the **loss part** are colored in different themes for clarity. This figure only depicts the situation of two inputs, and more inputs follow the same pattern. (b) shows the inference phase of the CPR-CLIP. Testing videos are sampled from *Set-2* of CPR-Coach, which contains abundant error combinations.

- We propose the first prompt-based pre-training framework named CPR-CLIP for composite error action recognition in CPR training;
- Extensive experimental results demonstrate that the CPR-CLIP have achieved promising performances on composite error recognition task;
- We deploy the CPR-CLIP model in practice to verify its effectiveness. Results show that the system can effectively improve the doctors' efficiency by nearly 4 times.

II. PROPOSED METHOD

The architecture of the CPR-CLIP is shown in Fig. 2(a). It consists of three stages: video feature extraction, prompt generation and embedding, and contrastive loss computing.

A. Video Feature Extraction

Firstly, sample two videos (\mathbf{v}_1, c_1) and (\mathbf{v}_2, c_2) from different classes of single-class error videos from the dataset. Labels c_1 and c_2 indicate the errors in videos \mathbf{v}_1 and \mathbf{v}_2 , respectively. Note that the sampling process meets $c_1 \neq c_2$. \mathbf{v}_1 contains L_1 frames and \mathbf{v}_2 contains L_2 frames, i.e., $\mathbf{v}_1 = \{I_i\}_{i=1}^{L_1}$, $\mathbf{v}_2 = \{I_j\}_{j=1}^{L_2}$. With the help of video backbones such as TSN [16], we can map the original video \mathbf{v} into video feature:

$$\mathbf{x} = \mathcal{F}_{TSN}(\mathbf{v}; \theta_{TSN}), \mathbf{x} \in \mathbb{R}^{T \times D}, \quad (1)$$

where T denotes the number of clips sampled from the original video \mathbf{x} , D denotes the dimension of the video feature, and θ_{TSN} denotes the parameters in the video backbone.

In the feature fusion stage, we adopt the same configuration as [11] for comparability. After enhancing the diversity of feature combinations through introducing randomness $\lambda \sim U(0, 1)$ into the weighted summation, a temporal average pooling operation is added to obtain the video representation:

$$\mathbf{x}_n = \frac{1}{T} \sum_{t=1}^T (\lambda \mathbf{x}_1^t + (1 - \lambda) \mathbf{x}_2^t), \mathbf{x}_n \in \mathbb{R}^D, \quad (2)$$

where \mathbf{x}_1^t denotes the t -th row of the video feature \mathbf{x}_1 .

Finally, the fused features \mathbf{x}_n are transformed into the final video representation through the video feature encoder $\mathcal{F}_V(\cdot)$:

$$\mathbf{f}_n = \mathcal{F}_V(\mathbf{x}_n; \theta_V), \mathbf{f}_n \in \mathbb{R}^D, \quad (3)$$

where θ_V denotes the trainable parameters. The visual encoder $\mathcal{F}_V(\cdot)$ is instantiated through a two-layered MLP network.

B. Prompt Generation and Embedding

Human language can naturally express various composite information fluently [28]. Inspired by this, this letter designs a set of prompt templates for expressing combinations of error actions. Fig. 2(a) illustrates the process of generating prompts for *Overlap Hands* and *Bending Arms* errors in detail. The set of templates comprehensively describes specific error combinations from quantity, classes, and corresponding advice. Detailed definitions of Number Prompt P_{num} , Classes Prompt P_{cls} , and Advice Prompt P_{adv} are as follow:

$$\begin{aligned} P_{num}^n &= \text{“This video clip contains } \{cnt\} \text{ errors in total.”} \\ P_{cls}^n &= \text{“This subject made both } \{c_1\} \text{ and } \{c_2\} \text{ mistakes.”} \\ P_{adv}^n &= \text{“This subject should } \{a_1\}, \text{ and } \{a_2\} \text{.”} \end{aligned}$$

where $\{cnt\}$ represents the number of composite errors, which varies according to the number of composite errors during the training stage. $\{c_1\}$ and $\{c_2\}$ represent specific error categories, while $\{a_1\}$ and $\{a_2\}$ represent corresponding advice, respectively. n represents the n -th sample within a batch.

Three types of prompts P_{num}^n , P_{cls}^n and P_{adv}^n are fused into the final prompt P_n through string concatenation:

$$P_n = P_{num}^n \oplus P_{cls}^n \oplus P_{adv}^n, \quad (4)$$

where \oplus denotes the concatenation operation. Similar to the mapping process in the visual pathway, the prompt P_n is mapped to embeddings through the text encoder:

$$\mathbf{t}_n = \mathcal{F}_T(P_n; \theta_T), \mathbf{t}_n \in \mathbb{R}^D, \quad (5)$$

where θ_T denotes the trainable parameters of the text encoder $\mathcal{F}_T(\cdot)$. The structure of $\mathcal{F}_T(\cdot)$ follows the setting in the original CLIP framework [27]: a 12-layer Transformer with the feature dimension of 512 and 8 attention heads.

C. Contrastive Pre-Training Loss

The CPR-CLIP aims to align visual and linguistic features into the same semantic representation space through a self-supervised contrastive pre-training mechanism. We take advantage of the CLIP loss to improve the performance and usability of the network. In a batch that contains N samples, the visual feature $\mathbf{F} = \{\mathbf{f}_n\}_{n=1}^N$ and text feature $\mathbf{T} = \{\mathbf{t}_n\}_{n=1}^N$ are obtained through visual and language pathways, respectively. The cosine similarity between \mathbf{f}_n and \mathbf{t}_n is defined as:

$$\text{sim}(\mathbf{f}_n, \mathbf{t}_n) = \frac{\mathbf{f}_n \cdot \mathbf{t}_n}{\|\mathbf{f}_n\| \|\mathbf{t}_n\|}. \quad (6)$$

Therefore, the similarity matrix in a batch is represented as:

$$S(\mathbf{F}, \mathbf{T}) = \begin{bmatrix} \text{sim}(\mathbf{f}_1, \mathbf{t}_1) & \cdots & \text{sim}(\mathbf{f}_1, \mathbf{t}_N) \\ \vdots & \ddots & \vdots \\ \text{sim}(\mathbf{f}_N, \mathbf{t}_1) & \cdots & \text{sim}(\mathbf{f}_N, \mathbf{t}_N) \end{bmatrix}. \quad (7)$$

By adopting the softmax normalization function along rows and columns of the $S(\mathbf{F}, \mathbf{T})$, we can obtain the text-wise similarity matrix $S_T(\mathbf{F}, \mathbf{T})$ and video-wise similarity matrix $S_V(\mathbf{F}, \mathbf{T})$, respectively. Afterward, a Ground-Truth (GT) matrix $M_{GT} \in \mathbb{1}^{N \times N}$, $\mathbb{1} = \{0, 1\}$ is created according to the consistency of visual and language labels within a batch. In M_{GT} , positions where video features and linguistic features match are padded with 1, while others with 0.

The Kullback–Leibler (KL) divergence is adopted as the metric between similarity matrices and M_{GT} . The multi-modal contrastive pre-training loss of the CPR-CLIP is expressed as:

$$\mathcal{L}_{CLIP} = \frac{1}{2} (KL[S_F(\mathbf{F}, \mathbf{T}) || M_{GT}] + KL[S_V(\mathbf{F}, \mathbf{T}) || M_{GT}]). \quad (8)$$

The goal of the optimizer is to find the optimal parameters (θ_V, θ_T) of CPR-CLIP to minimize the loss:

$$(\theta_V^*, \theta_T^*) = \arg \min_{(\theta_V, \theta_T)} \mathcal{L}_{CLIP}. \quad (9)$$

Although the \mathcal{L}_{CLIP} provides supervision during training, it is insufficient because it belongs to the self-supervised paradigm. Therefore, we follow the design of the network head in ImagineNet [11] and add the Binary Cross Entropy (BCE) loss to the CPR-CLIP. This variant is named by CPR-CLIP+ for discrimination. Loss of the CPR-CLIP+ is expressed as:

$$\mathcal{L}^+ = \mathcal{L}_{CLIP} + \mathcal{L}_{BCE}. \quad (10)$$

D. Inference of the CPR-CLIP

Details of the inference stage are shown in Fig. 2(b). In the language pathway, assuming there are K types of independent errors, we can obtain a prompt set $P' = \{P_k\}_{k=1}^K$ through the prompt templates mentioned above. For the k -th class, the input prompt is obtained through:

$$P_k = P_{num}^k \oplus P_{cls}^k \oplus P_{adv}^k. \quad (11)$$

After that, an embedding set $\mathbf{T}' = \{\mathbf{t}_k\}_{k=1}^K$ corresponding to the testing prompts set P' is obtained through the pre-trained text encoder: $\mathbf{t}_k = \mathcal{F}_T(P_k; \theta_T^*)$, $\mathbf{t}_k \in \mathbb{R}^D$. In the visual pathway, after the target video \mathbf{v}' is mapped to \mathbf{x}' through the

video backbone, the visual feature \mathbf{f}' is generated by the video encoder: $\mathbf{f}' = \mathcal{F}_V(\mathbf{x}'; \theta_V^*)$, $\mathbf{f}' \in \mathbb{R}^D$.

During inference, the similarity matrix in Eq. (7) degenerates into a K -dimensional vector for the video feature \mathbf{f}' , which indicates the similarity scores between \mathbf{f}' and \mathbf{T}' :

$$S_V(\mathbf{f}', \mathbf{T}') = [\text{sim}(\mathbf{f}', \mathbf{t}_1), \cdots, \text{sim}(\mathbf{f}', \mathbf{t}_K)]. \quad (12)$$

CPR-CLIP also supports video retrieval patterns. Given a specific query prompt embedding \mathbf{t}' and the entire video features set \mathbf{F}' , the CPR-CLIP generates a video-wise similarity vector:

$$S_T(\mathbf{F}', \mathbf{t}') = [\text{sim}(\mathbf{f}_1, \mathbf{t}'), \cdots, \text{sim}(\mathbf{f}_M, \mathbf{t}')]^T. \quad (13)$$

This advantage of CPR-CLIP can be used for fast retrieval function with natural language among large-scale video sets.

III. EXPERIMENTS

A. Dataset and Evaluation Metrics

The CPR-Coach dataset [11] provides 14 single-class actions and 74 composite error actions in four different perspectives, containing 4,544 videos. All experiments in this letter maintained the same settings with [11] for comparability: single-class error videos in *Set-1* are used for training, while composite error videos in *Set-2* are used for testing. The performance of composite error action recognition is measured through *mAP* and *mmit mAP*. *mAP* averages the precision of all K classes, $mAP = (\sum_{k=1}^K AP_k)/K$. *mmit mAP* averages the precision over all M videos, $mmit mAP = (\sum_{m=1}^M AP_m)/M$. Two metrics correspond to the *macro mAP* and *micro mAP* in [23], respectively.

B. Implementation Details

All experiments are implemented on a system with an AMD EPYC 7742@2.25GHz CPU and an NVIDIA Tesla A800 GPU. The input resolution of the video backbone is set to 224×224 . The training epoch is set to 60, corresponding to 32k pre-training iterations under batch size $N = 32$. The SGD optimizer is adopted with the base learning rate of 0.001 and attenuated by 0.1 at 20 and 40-th epochs. The visual encoder is loaded from the pre-trained model and frozen during training.

C. Performance Evaluation

Table I lists the performance of direct migration methods, CPR-CLIP, and the variant model CPR-CLIP+. Taking the vanilla migrations as baselines, Table I annotates the performance gains brought by the proposed mechanisms. Results show that the pre-training process brings significant performance improvements. For example, 9.89% *mAP* and 8.66% *mmit mAP* gains are observed on CPR-CLIP w/ Video Swin Transformer [20]. Fig. 3 visualizes the loss of CPR-CLIP and performance metrics of each epoch during training. Continuous decrease of the loss indicates that the language embeddings and visual features are gradually and successfully aligned into the same space. Decreasing loss and increasing performance confirm the effectiveness of the CPR-CLIP. Table I also demonstrates the indispensability of BCE loss, which brings additional performance gains to the CPR-CLIP.

TABLE I
PERFORMANCE COMPARISON AMONG DIRECT MIGRATION, CPR-CLIP,
AND THE VARIANT MODEL CPR-CLIP+.

Model	mAP	Δ	mmit mAP	Δ
TSN [16]	0.5598	—	0.6143	—
CPR-CLIP	0.6034	\uparrow 4.36%	0.6727	\uparrow 5.84%
CPR-CLIP+	0.6417	\uparrow 8.19%	0.7030	\uparrow 8.87%
TSM [17]	0.5662	—	0.6618	—
CPR-CLIP	0.6401	\uparrow 7.39%	0.7074	\uparrow 4.56%
CPR-CLIP+	0.7076	\uparrow 14.14%	0.7602	\uparrow 9.84%
ST-GCN [18]	0.5776	—	0.6692	—
CPR-CLIP	0.6028	\uparrow 2.52%	0.6831	\uparrow 1.39%
CPR-CLIP+	0.6358	\uparrow 5.82%	0.7127	\uparrow 4.35%
ViViT [19]	0.5582	—	0.6651	—
CPR-CLIP	0.6503	\uparrow 9.21%	0.7494	\uparrow 8.43%
CPR-CLIP+	0.7251	\uparrow 16.69%	0.7754	\uparrow 11.03%
Video Swin [20]	0.5696	—	0.6701	—
CPR-CLIP	0.6685	\uparrow 9.89%	0.7567	\uparrow 8.66%
CPR-CLIP+	0.7439	\uparrow 17.43%	0.7924	\uparrow 12.23%

TABLE II
PERFORMANCE COMPARISON BETWEEN CLIP-CPR+ AND SOTAS.

Model	Backbone	mAP	mmit mAP
CBP [21]	TSM [17]	0.6864	0.7487
Block [22]		0.6651	0.7222
ImagineNet-FC [11]		0.7053	0.7566
CPR-CLIP+		0.7076	0.7602
CBP [21]	Video Swin [20]	0.6951	0.7524
Block [22]		0.6801	0.7322
ImagineNet-FC [11]		0.7082	0.7638
CPR-CLIP+		0.7439	0.7924

Table II compares the performance of CLIP-CPR+ with the existing state-of-the-art (SOTA) methods. Two feature aggregation methods, Compact Bilinear Pooling (CBP) [21] and Block [22], are implemented and measured for comprehensive comparison. Results show that the combination of contrastive pre-training loss \mathcal{L}_{CLIP} and BCE loss \mathcal{L}_{BCE} in CPR-CLIP+ effectively improves the generalization performance in composite error recognition tasks. The results confirm the complementation of contrastive self-supervision and full supervision. In other words, the CPR-CLIP+ adds contrastive pre-training loss \mathcal{L}_{CLIP} based on the ImagineNet-FC. Therefore, the performance comparison belongs to ablation results, which also confirms the effectiveness of the proposed framework.

D. Ablation Studies

Ablation studies are conducted on three types of prompts P_{cnt} , P_{cls} , and P_{adv} of the CPR-CLIP. Table III lists the performance under various ablation settings. Macroscopically, all three prompts contribute to the final performance. Results show that P_{cls} has the highest weight, which is consistent with our intuition because P_{cls} explicitly describes error information. The advice prompt P_{adv} also provides a positive impact as it introduces richer semantic information during pre-training. Removing the number prompt P_{num} brings 0.73% *mmit mAP* improvement of the CPR-CLIP w/ TSM. This is mainly caused by the default setting that $\{cnt\}$ is *one* during inference, thus resulting in the misalignment issue. Although enriching the types of prompts improves the effectiveness of the multimodal contrastive pre-training process, the improvement is still lower than the performance gain brought by the BCE loss. The entire framework requires the combination and complementarity between \mathcal{L}_{CLIP} and \mathcal{L}_{BCE} .

TABLE III
ABLATION STUDIES OF THREE TYPES OF PROMPTS ON CPR-CLIP.

Backbone	Variants	P_{num}	P_{cls}	P_{adv}	mAP	mmit mAP
TSM [17]	CPR-CLIP	✓	✓	✓	0.6401	0.7074
		✗	✓	✓	0.6298	0.7147
		✓	✗	✓	0.4498	0.5480
		✓	✓	✗	0.5651	0.6870
	CPR-CLIP+	✓	✓	✓	0.7076	0.7602
Video Swin [20]	CPR-CLIP	✓	✓	✓	0.6685	0.7567
		✗	✓	✓	0.6351	0.7328
		✓	✗	✓	0.4525	0.5910
		✓	✓	✗	0.5591	0.7470
	CPR-CLIP+	✓	✓	✓	0.7439	0.7924

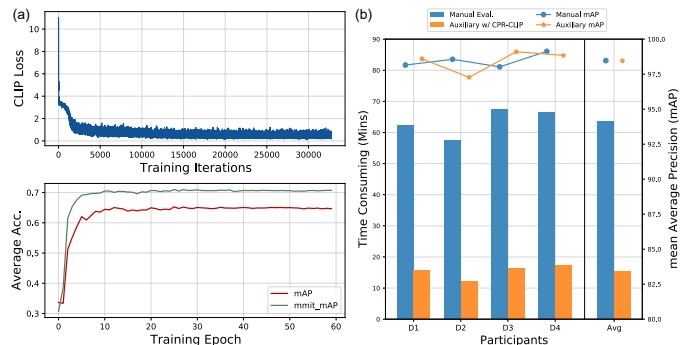


Fig. 3. (a) shows the contrastive pre-training loss of the CPR-CLIP model and the testing performance after each training epoch. (b) demonstrates the time consumption and evaluation performance of four doctors.

E. Practical Deployment of the CPR-CLIP

As described in Eq. (13), the CPR-CLIP supports video retrieving with natural language, which can be used as an electronic assistant. Under this setting, doctors no longer need to suffer from judging each video individually but instead use natural language to query all videos through CPR-CLIP and then review the assistant results. To verify the actual effectiveness of the system, we recruited four doctors and conducted comparative experiments on the time-consuming and evaluation quality. All testing videos in CPR-Coach *Set-2* are randomly divided into two groups: G_1 and G_2 . Firstly, these doctors are asked to identify all errors on each video in G_1 . Afterwards, they are asked to use CPR-CLIP as an auxiliary retrieval tool to identify all videos in G_2 . Time-consuming and precision under two settings will be recorded in the end. Fig. 3 summarizes the average time-consuming and mean average precision of all four doctors. As expected, the CPR-CLIP significantly improves efficiency without compromising the evaluation quality. The assistant system has helped doctors improve their efficiency by nearly 4 times on average.

IV. CONCLUSION

In this letter, we propose a contrastive pre-training framework named CPR-CLIP to address the composite error recognition issue in CPR training. Extensive experiments and practical deployment demonstrate the effectiveness of the CPR-CLIP under the *Single-class Training & Multi-class Testing* setting. This study only focuses on the external cardiac compression action analysis, not the entire process of CPR. We will continue to explore the application of CPR-CLIP in complex temporal medical action analysis and retrieval in the future.

REFERENCES

- [1] D. Asch, J. Grischkan, S. Nicholson, "The cost, price, and debt of medical education," *New England J. of Med.*, vol. 383, pp. 6-9, 2020.
- [2] F. Murtaza, M. H. Yousaf, S. A. Velastin and Y. Qian, "End-to-end temporal action detection using bag of discriminant snippets," *IEEE Signal Process. Lett.*, vol. 26, pp. 272-276, 2019.
- [3] J. Kong, Y. Bian and M. Jiang, "MTT: Multi-scale temporal transformer for skeleton-based action recognition," *IEEE Signal Process. Lett.*, vol. 29, pp. 528-532, 2022.
- [4] J. P. Ebenezer, Z. Shang, Y. Wu, H. Wei, S. Sethuraman and A. C. Bovik, "Making video quality assessment models robust to bit depth," *IEEE Signal Process. Lett.*, vol. 30, pp. 488-492, 2023.
- [5] Y. Teng, C. Song and B. Wu, "Recognizing social relationships in long videos via multimodal character interaction," *IEEE Signal Process. Lett.*, vol. 30, pp. 573-577, 2023.
- [6] H. Wu, M. Li, Y. Liu, H. Liu, C. Xu and X. Li, "Transtl: spatial-temporal localization transformer for multi-label video classification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2022, pp. 1965-1969.
- [7] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, and J. Yuan, "Action-stage emphasized spatiotemporal VLAD for video action recognition," *IEEE Trans. on Image Process.*, vol. 28, pp. 2799-2812, 2019.
- [8] Q. Zhang and B. Li, "Relative hidden markov models for video-based evaluation of motion skills in surgical training," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, pp. 1206-1218, 2015.
- [9] A. Zia, Y. Sharma, V. Bettadapura, E. Sarin, and I. Essa, "Video and accelerometer-based motion analysis for automated surgical skills assessment," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, pp. 443-455, 2018.
- [10] S. Wang, D. Yang, P. Zhai, C. Chen, and L. Zhang, "TSA-Net: Tube self-attention network for action quality assessment," in *Proc. ACM Int. Conf. on Multimedia*, pp. 4902-4910, 2021.
- [11] S. Wang, Q. Yu, S. Wang, D. Yang, L. Su, X. Zhao, H. Kuang, P. Zhang, P. Zhai, L. Zhang, "CPR-Coach: Recognizing composite error actions based on single-class training," in *arXiv:2309.11718*, pp. 1-35, 2023.
- [12] A. Vakanski, H. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data (Basel)*, pp. 1-15, 2018
- [13] L. Zappella, B. Bejar, G. Hager, and R. Vidal, "Surgical gesture classification from video and kinematic data," *Med. Image Anal.*, vol. 17, pp. 732-745, 2013.
- [14] N. Ahmidi, P. Poddar, J. Jones, S. Vedula, L. Ishii, G. Hager, and M. Ishii, "Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 10, pp. 981-991, 2015.
- [15] Y. Sharma, T. Plotz, N. Hammerld, S. Mellor, R. McNaney, P. Olivier, S. Deshmukh, A. McCaskie, and I. Essa., "Automated surgical osats prediction from videos," in *IEEE Int. Symp. on Biomed. Ima.*, pp. 461-464, 2014
- [16] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. ValGool, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Europ. Conf. Comput. Vision*, 2016, pp. 20-36.
- [17] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE Int. Conf. on Comput. Vision*, 2019, pp. 7082-7092.
- [18] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. AAAI Conf. on Artif. Intell.*, 2018, pp. 7444-7452.
- [19] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić and C. Schmid, "ViViT: A Video Vision Transformer," in *Proc. IEEE Int. Conf. on Comput. Vision*, 2021, pp. 6816-6826.
- [20] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin and H. Hu, "Video Swin Transformer," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recog.*, 2022, pp. 3192-3201.
- [21] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell, "Compact bilinear pooling," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recog.*, 2016, pp. 317-326
- [22] H. Ben-Younes, R. Cadene, N. Thome, and M. Cord, "Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection," in *Proc. AAAI Conf. on Artif. Intell.*, 2019, pp. 8102-8109.
- [23] M. Monfort, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, D. Gutfrued, C. Vondrick, et al., "Moments in time dataset: one million videos for event understanding", *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019, pp. 502-508.
- [24] S. Wang, D. Yang, P. Zhai, Q. Yu, T. Suo, Z. Sun, K. Li, and L. Zhang, "A survey of video-based action quality assessment," in *Proc. Int. Conf. on Networking Sys. of AI*, pp. 1-9, 2021.
- [25] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. Vedula, G. Lee, M. Lee, and G. Hager. "Recognizing surgical activities with recurrent neural networks," In *Proc. Int. Conf. Med. Image Comput. and Comput.-Assisted Intervention*, pp. 551-558, 2016.
- [26] Y. Gao, S. Vedula, C. Reiley, N. Ahmidi, B. Varadarajan, H. Lin, L. Tao, L. Zappella, B. Bejar, D. Yuh, C. Chen, R. Vidal, S. Khudanpur and G. Hager, "JHU-ISI gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *Proc. Int. Conf. Med. Image Comput. and Comput.-Assisted Intervention*, pp. 1-10, 2014.
- [27] A. Radford, J. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, pp. 8748-8763, 2021
- [28] M. Wang and J. Xing and Y. Liu, "ActionCLIP: A new paradigm for video action recognition," in *arXiv:2109.08472*, pp. 1-11, 2021
- [29] J. Huang, W. Chen, S. Yang, D. Xie, S. Pu, and Y. Zhuang, "Transductive clip with class-conditional contrastive learning," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 3858-3862, 2022
- [30] M. Li, L. Chen, Y. Duarr, Z. Hu, J. Feng, J. Zhou, and J. Lu, "Bridge-Prompt: Towards ordinal action understanding in instructional videos," in *Proc. IEEE Conf. on Comput. Vis. Pattern Recog.*, pp. 19848-19857, 2022.
- [31] J. Zhao, R. Li, Q. Jin, X. Wang and H. Li, "Memobert: Pre-training model with prompt-based learning for multimodal emotion recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, pp. 4703-4707, 2022