CPR-Coach: Recognizing Composite Error Actions based on Single-class Training

Shunli Wang[1,2], Shuaibing Wang[1,2], Dingkang Yang[1,2], Mingcheng Li[1,2], Haopeng Kuang[1,2], Xiao Zhao[1,2], Liuzhen Su[1,2], Peng Zhai[1,2], Lihua Zhang[1,2,3,4*]

[1]Academy for Engineering and Technology, Fudan University [2]Cognition and Intelligent Technology Laboratory

[3]Engineering Research Center of Al and Robotics, Ministry of Education

[4]Al and Unmanned Systems Engineering Research Center of Jilin Province

{slwang19,lihuazhang}@fudan.edu.cn

Abstract

Fine-grained medical action analysis plays a vital role in improving medical skill training efficiency, but it faces the problems of data and algorithm shortage. Cardiopulmonary Resuscitation (CPR) is an essential skill in emergency treatment. Currently, the assessment of CPR skills mainly depends on dummies and trainers, leading to high training costs and low efficiency. For the first time, this paper constructs a vision-based system to complete error action recognition and skill assessment in CPR. Specifically, we define 13 types of single-error actions and 74 types of composite error actions during external cardiac compression and then develop a video dataset named CPR-Coach. By taking the CPR-Coach as a benchmark, this paper investigates and compares the performance of existing action recognition models based on different data modalities. To solve the unavoidable "Single-class Training & Multi-class Testing" problem, we propose a human-cognition-inspired framework named ImagineNet to improve the model's multi-error recognition performance under restricted supervision. Extensive comparison and actual deployment experiments verify the effectiveness of the framework. We hope this work could bring new inspiration to the computer vision and medical skills training communities simultaneously. The dataset and the code are publicly available on https://github.com/Shunli-Wang/CPR-Coach.

1. Introduction

High professionalism and data shortage seriously hinder the development of fine-grained medical action analysis technology [22, 60]. This paper takes Cardiopulmonary Resuscitation (CPR) as the research example, which is a critical life-saving technique for cardiac and respiratory arrest. CPR aims to restore the patient's spontaneous breathing and circulation. According to the American Heart Association (AHA), high-quality and standard CPR is the core of effective treatment, while improper actions will reduce the treatment effectiveness. Traditional CPR skill assessment usually requires the participation of the examiner and the dummy equipped with force sensors. The cost of this hy-

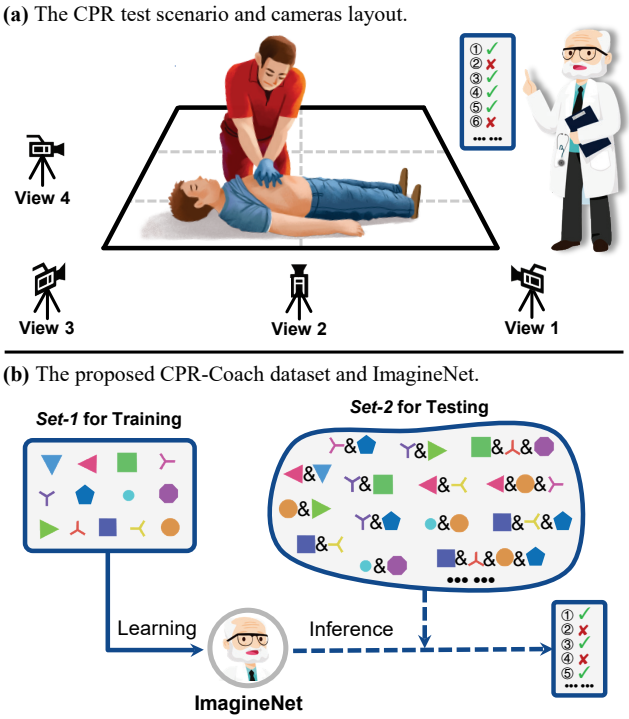


Figure 1. (a) shows the multi-view capture system. (b) illustrates the structure of the CPR-Coach dataset and the function of the ImagineNet. Each colored mark represents an error action class.

*Corresponding author.

This work is supported by the National Key R&D Program of China (2021ZD0113502).

brid evaluation method is too high to conduct large-scale training system deployment [72, 74]. In this paper, we build an intelligent system that automatically identifies wrong actions in CPR during skill training, thus significantly reducing the assessment cost and improving training efficiency.

As far as we know, there is no clear definition of specific error types of CPR actions, and no research has been done to explore the vision-based CPR skill assessment. To fill this gap, we first identify 13 types of common error actions (shown in Figure 2(a)) under the guidance of the latest version of *AHA Guidelines for CPR & ECC* [5] and doctors. A visual system is constructed to capture videos of the rescue process, as shown in Figure 1(a). Subsequently, we create a dataset named CPR-Coach, which consists of two parts: *Set-1* that contains single-class actions, and *Set-2* that contains composite error actions. Figure 1(b) graphically depicts the dataset's structure through colored marks.

Existing action recognition frameworks [12, 21, 31, 58] have been able to handle the single-class action recognition task. We can directly migrate these models to CPR-Coach *Set-1* to evaluate the fine-grained errors recognition performance. However, these models cannot meet the actual application in the CPR test. In actual CPR skill assessment, rescuers are likely to make multiple mistakes simultaneously, and a qualified coach is supposed to point out all mistakes exactly. If the number of single errors is 13, the total number of composite errors can reach a frightening 8,191 ($\sum_{n=1}^{13} C_{13}^{n} = 2^{13} - 1$). It is impossible to conduct exhaustive collection to cover all these error combinations.

To solve this dilemma, let us re-think how a real coach works. This coach must not have seen all the wrong action combinations, but he can still give the correct judgment according to the single-error action knowledge. This is because human beings have extremely strong knowledge reasoning and generalization abilities [41]. Inspired by this, this paper proposes a concise framework named ImagineNet to handle the intractable *Single-class Training & Multi-class Testing* problem properly. The function of the ImagineNet is shown in Figure 1(b). The essence of the ImagineNet is a human-inspired feature combination training strategy. As its name implies, it can *Imagine* composite error features based on restricted single-class error actions and achieves high performance in the unseen composite error recognition task. By regarding *Set-1* as the training set and *Set-2* as the testing set, we can examine the ImagineNet, which plays the role of *Coach*. Sufficient experimental results confirm the effectiveness of the framework.

The main contributions of this paper are as follows:

- To the best knowledge, we propose the first dataset named CPR-Coach in the visual CPR assessment task, which supports fine-grained action recognition and composite error recognition tasks.
- Taking the CPR-Coach dataset as a benchmark, we explore and compare the existing action recognition models based on different modality information.
- We propose a human-cognition-inspired framework ImagineNet, which significantly improves the composite error recognition performance with restricted supervision.

2. Related Work

Human Action Recognition. Video-based Human Action Recognition (HAR) is one of the representative tasks of video understanding. There exist some HAR benchmarks [1, 2, 24, 30, 45] and frameworks [9, 12, 19, 21, 31, 47, 51, 52, 58, 63, 65]. Benefiting from the availability of sports videos, some fine-grained HAR datasets [46, 54, 62] are proposed in sports. Fine-grained action recognition in medical field mainly focuses on surgical workflow recognition, such as laparoscopic cholecystectomy [26, 37, 53, 56], cataract surgery [4, 44], and *Da Vinci* surgical system operation [6, 22, 36, 43]. Although these benchmarks delineate surgical workflows, they are usually limited in scale and focus only on the interaction of surgical instruments with tissues, without recording and identifying incorrect operations by subjects. To fill the gap of fine-grained action recognition in CPR training, this paper proposes the first dataset named CPR-Coach, which contains indistinguishable errors and complex composite error classes, putting forward higher requirements for action recognition models.
Action Quality Assessment. Action Quality Assessment (AQA) aims to identify the score or rank specific skilled actions. Wang *et al.* [60] found that publicly available AQA datasets and algorithms in sports [8, 38–40, 42, 59, 61] are more than those in the medical field [3, 17, 22, 27, 44, 49, 50, 56, 71–73, 76], which is mainly caused by the high professionalism of medical data acquisition. Existing studies on medical AQA can be divided into three categories: surgical skill evaluation [44, 49, 50, 75, 76] under the OSATS system [34], operating skills identification based on *Da Vinci* surgical systems [3, 17, 22, 33, 67], and skill assessment in laparoscopic surgery [13, 27, 28, 56, 71–73]. These research only rated medical actions and did not conduct detailed analysis. As the CPR testing focuses more on specific errors and is not suitable for judging through scores, this paper extends the concept of traditional AQA to CPR.
Multi-Label Learning Algorithms. Different from traditional classification tasks, multi-label learning faces the challenge of exponential growth in the number of class label spaces [32, 70]. Existing solutions are mainly divided into two categories: Convert the multi-label problem into multiple independent binary classification problems [11, 16, 69], or improve the algorithm to adapt to multi-label data [25, 29, 57]. In addition, Dmitriev *et al.* [15] explored the setting of samples with one positive label. Cole *et al.* [18] explored the same topic in multi-class image segmentation tasks. Although the composite error recogni-

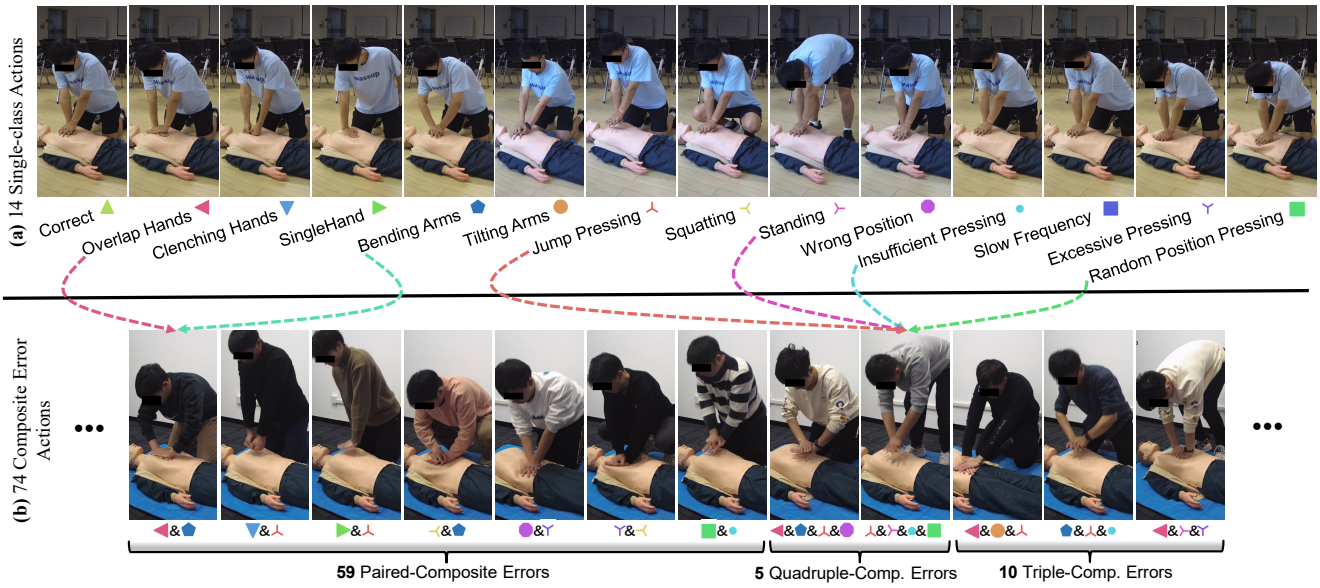


Figure 2. Structure of the CPR-Coach. (a) *Set-1* consists of a *Correct* class and 13 types of single-error actions. (b) *Set-2* consists of 74 composite error actions (59 paired-, 10 triple-, and 5 quadruple-composite errors). For clarity, different marks with different colors are adopted to represent 14 single classes. Due to space limitations, this figure only shows the generation process of one paired- and one quadruple-composite error actions. All 74 composite error actions are enumerated and annotated in detail in the supplementary material.

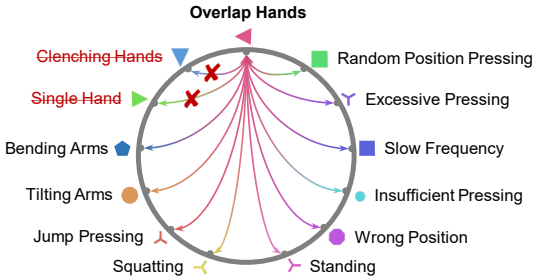


Figure 3. The selection strategy of the composite error actions. In this case, *Overlap Hands* is selected as the primary class, and two impossible co-occurrence combinations are deleted.

Dataset	#Actions	Modality	#Videos	#Views	Available
FLS-ASU [71]	1	RGB	28	2	✘
Sharma *et al.* [48]	2	RGB	33	1	✘
Bettadapura *et al.* [10]	3	RGB	64	2	✘
Zia *et al.* [74]	2	RGB	104	1	✘
Zhang *et al.* [72]	1	RGB	546	1	✘
Chen *et al.* [14]	3	RGB	720	2	✘
Cataract-101 [44]	2	RGB	101	1	✔
Hei-Chole [56]	7	RGB	33	1	✔
MISTIC-SL [17]	4	RGB+Kinematics	49	1	✘
JIGSAWS [22]	3	RGB+Kinematics	103	1	✔
UI-PRMD [54]	10	RGB+Kinematics	1,000	1	✔
CPR-Coach (Ours)	14+74	RGB+Flow+Pose	5,664	4	✔

Table 1. Comparison with existing fine-grained medical action analysis analysis and assessment datasets. More detailed comparison results are listed in the supplementary material.

tion belongs to multi-label learning, the conditions are more stringent, *i.e.*, the training samples only contain single-class samples. The proposed ImagineNet follows an algorithm transformation strategy and thoroughly improves the recognition performance through feature-combining strategies.

3. CPR-Coach Dataset

As shown in Figure 2, the proposed CPR-Coach dataset is divided into two parts: *Set-1* that contains 1 type of correct action and 13 types of single-error actions, and *Set-2* that contains 74 types of composite error actions.

Considering the exponential growth of the total number of composite error actions (8,191 classes for 13 single-error actions), this paper mainly focuses on paired combinations and several common multi-error combinations. Based on the filtering strategy in Figure 3, we remove 19 impossible combinations from 78 pairs ($C_{13}^2 = 78$) and finally get 59 paired-composite error actions. All deleted combinations have been confirmed by emergency doctors. In addition, 10 triple errors and 5 quadruple errors are selected by these professional doctors based on actual experience. Finally, we built a label space containing 74 combination errors.

Data Collection. We build a video capture system with four high-resolution cameras to record the rescue process, as shown in Figure 1(a). In order to ensure the diversity of the dataset, we recruited 12 volunteers to participate in data collection. Multiple participants enrich the visual feature diversity of the proposed dataset. Three volunteers were assigned to *Set-1*, while nine were assigned to *Set-2*. Single-class actions in *Set-1* are performed for 42 times. In *Set-2*, paired-composite error actions are performed for 12 times, while others are performed for 8 times. All actions are carried out under the guidance of professional doctors to ensure the quality of each external cardiac compression action.

Dataset Statistics. Table 1 compares the proposed CPR-

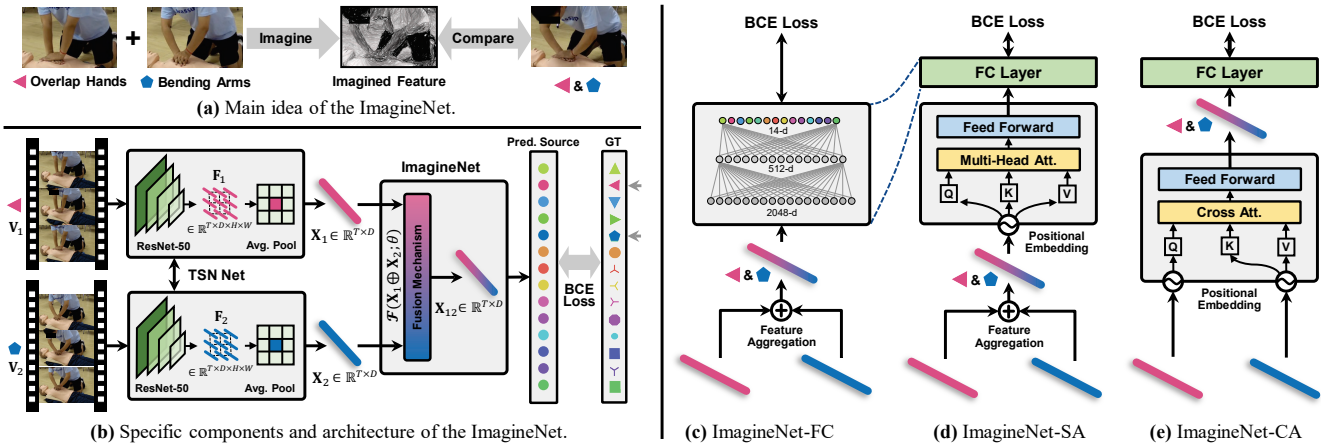


Figure 4. (a) and (b) demonstrate the main idea and specific network architecture of the proposed ImagineNet, respectively. Two error actions *Overleap Hands* and *Bending Arms* are selected for visualization. The ImagineNet simulates the thinking and judgment process of a real experienced coach concisely. The knowledge base only includes single-class actions, while real applications will encounter unseen composite errors. (c), (d) and (e) show three feature fusion mechanisms. Note that only two inputs are displayed for clarity.

Item	Data
Perspectives	4
FPS	25
Video Resolution	4096×2160 (4K)
Number of Participants	12
Classes of Single-class Actions	1+13=14
Classes of Composite Error Actions	59+10+5=74
Frames (RGB)	2,217,756
Frames (RGB+Flow)	6,644,596
Videos	5,664
Avg. Len. of Videos	19.52s
Storage Size	450GB

Table 2. Summary of statistics of the CPR-Coach dataset.

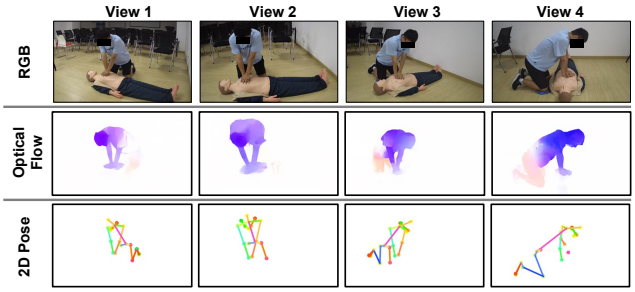


Figure 5. Three types of modality information on four views provided by CPR-Coach.

Coach dataset with existing fine-grained medical action analysis datasets. The CPR-Coach dataset has surpassed existing datasets in terms of data scale, action granularity, and modal complexity. More comparisons with other datasets are listed in the supplementary materials. Table 2 summarizes the statistics of the CPR-Coach dataset. It contains 4.6K videos and 2.2M frames in total. The storage size of the entire dataset is 450GB. The CPR-Coach also provides optical flow images generated by the TV-L1 algorithm [66] and 2D skeletons of the rescuer obtained by Alphapose [20]. Figure 5 shows three modality information from four perspectives: RGB frames, optical flow, and 2D poses.

Supported Tasks. As the first multi-perspective dataset to explore fine-grained composite actions in medical scenarios, the CPR-Coach can support multiple studies. Firstly, we can evaluate existing HAR models on fine-grained error recognition tasks on *Set-1*. Secondly, by taking *Set-1* as the training set and *Set-2* as the testing set, we can explore the composite error action recognition task under constrained supervision. Thirdly, the influence of combining different perspectives and modes on the algorithm can be explored. The following experiments follow these ideas.

4. ImagineNet

Figure 4(a) shows the main idea of the proposed human-cognition-inspired framework ImagineNet. With restricted supervision training data, the *Imagine* process can freely combine features to improve the multi-label recognition performance. Taking the classic Temporal Segment Network (TSN) [58] as the basic network, the detailed architecture of ImagineNet is shown in Figure 4(b).

The ImagineNet is divided into three stages: feature extraction, feature fusion, and loss computing. Firstly, two video samples $(\mathbf{V}_1, C_1)$ and $(\mathbf{V}_2, C_2)$ are selected from *Set-1* in the feature extraction phase. Note that two videos $\mathbf{V}_1 = \{I_i\}_{i=1}^{N_1}$ and $\mathbf{V}_2 = \{I_i\}_{i=1}^{N_2}$ come from different classes, *i.e.*, $C_1 \neq C_2, C \in \{1, \cdots, 13\}$. N_1 and N_2 represent the total frames of two videos, respectively. I_i denotes the i-th frame in the video. The TSN model selects T clips from raw videos for feature extraction. After spatial average pooling, video features $\mathbf{X}_1 \in \mathbb{R}^{T\times D}$ and $\mathbf{X}_2 \in \mathbb{R}^{T\times D}$ are obtained, where D denotes the dimension of the feature. Secondly, in the feature fusion stage, two different features will be subsequently fused and generate $\mathbf{X}_{12} \in \mathbb{R}^{T\times D}$.

This process is also expressed as $\mathbf{X}_1 \oplus \mathbf{X}_2$. We regard this feature fusion process as the *Imagine* process. As illustrated in Figure 4(c&d&e), this paper provides three feature fusion schemes to realize the imagination process: Fully-Connected Layer based fusion (FC), Self-Attention based fusion (SA), and Cross-Attention based fusion (CA). Finally, in the loss computing stage, the Binary Cross Entropy (BCE) loss is adopted to measure the divergence between the predicted score and the Ground-Truth (GT) labels. Note that the GT labels are in the form of multi-hot encoding.

4.1. Fusion Mechanisms of the ImagineNet

Subfigures in Figure 4(c&d&e) demonstrate three different feature fusion mechanisms: ImagineNet-FC, ImagineNet-SA, and ImagineNet-CA, respectively. The formula representation is omitted in these figures for clarity. Two thick lines with different colors are adopted to represent two video features.

ImagineNet-FC. As shown in Figure 4(c), the video features $\mathbf{X}_1$ and $\mathbf{X}_2$ are fused through the feature addition mechanism. Then a two-layer fully connected neural network maps the fusion feature $\mathbf{X}_{12}$ into predicted scores of 14 classes. This process is formulated as

$$S_{FC} = \mathcal{F}_{FC}(\mathbf{X}_1 \oplus \mathbf{X}_2; \boldsymbol{\theta}_{FC}), \tag{1}$$

where $\mathcal{F}_{FC}(\cdot)$ denotes the neural network, and the plus sign $\oplus$ represents the feature aggregation strategy, which will be described in detail later. $\boldsymbol{\theta}_{FC}$ represents the trainable parameters of $\mathcal{F}_{FC}(\cdot)$.

The BCE loss function is selected for the network optimization:

$$\boldsymbol{\theta}^*_{FC} = \arg\min_{\boldsymbol{\theta}_{FC}} BCE(S_{FC}, GT), \tag{2}$$

where $GT = onehot(C_1) \cup onehot(C_2)$ denotes the composite label in multi-hot encoding form. All parameters are omitted in the subsequent statements for clarity.

ImagineNet-SA. The ImagineNet-SA adds a self-attention module based on the ImagineNet-FC, as shown in Figure 4(d). The motivation is to equip the ImagineNet with a stronger feature extraction and fusion capability to improve the generalization and reasoning ability. The process is expressed as

$$S_{SA} = \mathcal{F}_{FC}(\mathcal{F}_{SA}(\mathbf{X}_1 \oplus \mathbf{X}_2)), \tag{3}$$

where $\mathcal{F}_{SA}(\cdot)$ includes the self-attention and feed forward stages, $\mathcal{F}_{FC}(\cdot)$ is the same as Equ.1. By substituting $\mathbf{X}_{12}$ for $\mathbf{X}_1 \oplus \mathbf{X}_2$, the self-attention mechanism is expressed as

$$\mathbf{X}^{'}_{SA} = LN\left[\mathbf{X}_{12} + softmax\left(\frac{\mathbf{X}_{12}\mathbf{X}^T_{12}}{\sqrt{D}}\right)\mathbf{X}_{12}\right], \tag{4}$$

and the feed forward layer

$$\mathbf{X}_{SA} = LN[\mathbf{X}^{'}_{SA} + \mathcal{F}_{FFN}(\mathbf{X}^{'}_{SA})]. \tag{5}$$

Note that D represents the dimension of video features and $D = 2048$ in TSN [58]. $LN[\cdot]$ denotes the LayerNorm operation. For clarity, the LayerNorm operation and residual links are omitted in Figure 4(d&e) .

ImagineNet-CA. The structure of ImagineNet-CA is shown in Figure 4(e). The main difference between ImagineNet-SA and ImagineNet-CA lies in the feature fusion strategy. Consistent with the above, the computing process is expressed as

$$S_{CA} = \mathcal{F}_{FC}(\mathcal{F}_{CA}(\mathbf{X}_1, \mathbf{X}_2)), \tag{6}$$

where $\mathcal{F}_{CA}(\cdot,\cdot)$ includes a cross-attention module and a feed forward layer. The cross-attention mechanism integrates two video features from different classes:

$$\mathbf{X}^{'}_{CA} = LN\left[\mathbf{X}_1 + softmax\left(\frac{\mathbf{X}_1\mathbf{X}^T_2}{\sqrt{D}}\right)\mathbf{X}_2\right], \tag{7}$$

and the feed forward layer

$$\mathbf{X}_{CA} = LN[\mathbf{X}^{'}_{CA} + \mathcal{F}_{FFN}(\mathbf{X}^{'}_{CA})]. \tag{8}$$

After defining three fusion mechanisms, we can instantiate three ImagineNets and compare their performance.

4.2. Feature Aggregation Strategy

Three feature fusion mechanisms mentioned above are frameworks for implementing ImagineNet, while feature aggregation is a local operation denoted as $\oplus$. Effective feature aggregation methods can make full use of limited samples in *Set-1*, thus improving the generalization performance under the setting of *Single-class Training & Multi-class Testing*. The simplest way to instantiate $\oplus$ in ImagineNet-FC and -SA models is taking the summation of two features. To increase the diversity of the aggregation process, we adopt a random weighted summation mechanism similar to MixUp [68]. The aggregated feature is expressed as follows with two inputs.

$$\mathbf{X}_{12} = \lambda\mathbf{X}_1 + (1-\lambda)\mathbf{X}_2, \lambda \sim U(0,1), \tag{9}$$

where λ is a weight sampled from a uniform distribution $U(0,1)$. The effectiveness of this concise technique is verified in ablation studies. As representatives of feature aggregation methods, CBP [23] and BLOCK [7] are selected for comparison. Weighted summation, CBP, and BLOCK are denoted as Agg-1, Agg-2, and Agg-3, respectively.

4.3. Inference of the ImagineNet

Figure 4(b) only demonstrates the training process of the ImagineNet. It can be found that ImagineNet requires two video features $\mathbf{X}_1$ and $\mathbf{X}_2$ as inputs during training. However, there is only one input video feature of the composite error action during inference. To resolve this mismatch issue, this paper directly adopts the replication method to fill the input. Although the cross-attention in ImagineNet-CA degenerates into the self-attention in ImagineNet-SA during inference, different training process leads to different recognition performance. The two models are still comparable, and the experimental results confirm this analysis.

Model	Modality	Backbone	Config	Epoch	Pre-training	CE Loss		BCE Loss		Multi-Margin Loss	
						Top-1	Top-3	Top-1	Top-3	Top-1	Top-3
TSN [58]	RGB	ResNet-50	1x1x8	50	✗	0.8879	0.9940	0.8829	0.9960	0.8502	0.9901
	RGB	ResNet-50	1x1x8	50	Kinetics-400	0.9067	0.9921	0.8919	0.9940	0.8690	0.9901
	Flow	ResNet-50	1x1x8	50	✗	0.7907	0.9603	0.8304	0.9851	0.7073	0.9355
TSM [31]	RGB	ResNet-50	1x1x8	50	✗	0.9067	0.9901	0.9325	0.9950	0.8433	0.9881
I3D [12]	RGB	ResNet-50	32x2x1	50	✗	0.9692	0.9960	0.9117	0.9940	0.8591	0.9861
TPN [64]	RGB	ResNet-50	8x8x1	50	✗	**0.9802**	0.9960	0.9087	**0.9980**	0.8720	0.9901
C3D [52]	RGB	C3D	16x1x1	50	Sports1M	0.9722	0.9931	0.9702	0.9931	0.8621	0.9802
TIN [47]	RGB	ResNet-50	1x1x8	50	✗	0.8800	0.9901	0.7192	0.9335	0.8393	0.9861
SlowFast [21]	RGB	ResNet-50	4x16x1	256	✗	0.8695	0.9734	0.8719	0.9781	0.8625	0.9688
TimeSFormer [9]	RGB	ViT	8x32x1	50	✗	0.8879	0.9921	0.8998	0.9940	0.8462	0.9762
ST-GCN [63]	Pose	ST-GCN	1x1x300	50	✗	0.9246	**0.9970**	0.9187	0.9881	0.9196	**0.9970**
PoseC3D [19]	Pose	ResNet3D-50	1x1x300	240	✗	0.9208	0.9922	0.9035	0.9715	0.8837	0.9606
Two-Stream [51]	RGB+Flow	TSN+TSN_Flow	Late-Fusion	50	✗	0.9533	0.9891	0.9479	0.9825	0.9296	0.9802
	RGB+Pose	TSN+ST-GCN	Late-Fusion	50	✗	0.9782	0.9962	0.9608	0.9941	**0.9692**	0.9960
MMNet [65]	RGB+Pose+RoI	MS-G3D+Incep.-v3	Late-Fusion	80	✗	0.9756	0.9960	**0.9772**	0.9940	0.9512	0.9876

Table 3. Single-class recognition performance of existing HAR models on CPR-Coach *Set-1*. The first and second accuracy in each column are highlighted in **bold** and underlined, respectively. More results in different settings are summarized in the supplementary materials.

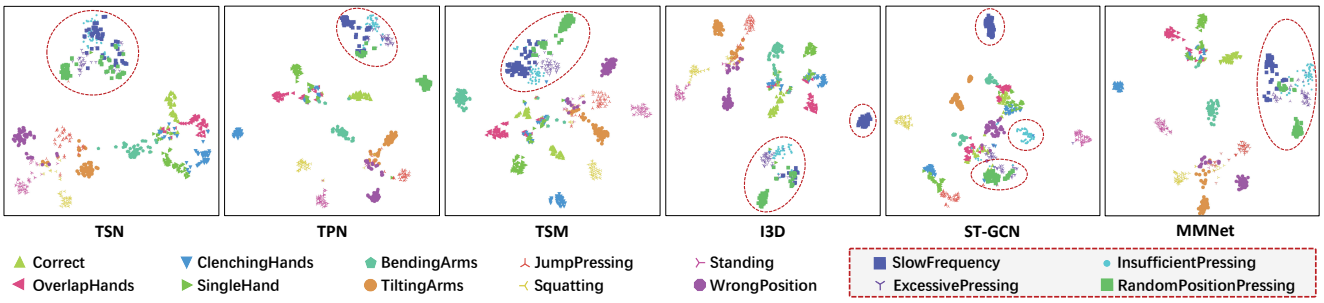


Figure 6. Visualization of the action features through t-SNE. The red box in the legend highlights four confusing classes. We use red circles to highlight these four classes of scatters in figures to compare the performance of these networks more clearly.

5. Experiments

5.1. Action Recognition on CPR-Coach *Set-1*

Compared with traditional HAR datasets, the CPR-Coach focuses on distinguishing subtle errors in CPR. In Figure 2, it is difficult to find the nuances of these actions. CPR-Coach puts forward higher requirements for the action recognition models. Therefore, we take *Set-1* of the CPR-Coach as a benchmark and conduct single-error recognition experiments on existing HAR models. 60% of *Set-1* is used for training and 40% for testing. Table 3 summarizes the detailed settings and Top-1&3 accuracy of the models. Figure 6 visualizes some features generated by these models through the t-SNE algorithm [55].

Implementation Details. Three different types of action recognition models are implemented: video-based methods (TSN [58], TSM [31], TPN [64], I3D [12], C3D [52], TIN [47], SlowFast [21], TimeSFormer [9]), pose-based methods (ST-GCN [63], PoseC3D [19]), and multimodal fusion methods (Two-Stream [51], MMNet [65]). Detailed configurations of these models are summariaed in Table 3. All models are trained for 50 epochs through the SGD optimizer, except the SlowFast with the Cosine Annealing optimizer for 256 epochs, the PoseC3D for 240 epochs, and the MMNet for 80 epochs. The network input size is 224x224, while the coordinates of 2D poses remain unchanged at 4096x2160. All models are built on Pytorch and trained on an NVIDIA A800 GPU. Cross Entropy (CE), BCE, and Multi-Margin losses are adopted to compare the performance comprehensively.

Performance Analysis. Results in Table 3 suggest that different network-loss combinations will affect the final performance. Under the CE loss setting, TPN achieves the best performance, while the MMNet achieves the optimal performance with the BCE loss. Methods that integrates multimodal information has the most stable performance and can achieve superior performance under different losses. The performance under CE loss is superior to the other two losses, which is caused by the stronger label dependency assumption. Surprisingly, the performance of PoseC3D is inferior to ST-GCN. This is mainly because that the PoseC3D stacks 2D keypoints to form a 3D heatmap volume, while the repeatability and circularity of CPR hit its inherent flaw, leading to inferior results. In Figure 6, the scatters of four confusing classes are very close in TSN, TPN, and TSM, while the I3D, ST-GCN, and MMNet that pay more attention to temporal information can handle these situations well. Class-wise prediction results and the pros and cons of these HAR models are analyzed in the supplementary material. Overall, existing HAR models are able to handle single-error recognition tasks well. Next, we will explore composite error performance on these models.

Model	Config	Modality	Pre-training	CE Loss		BCE Loss		Multi-Margin Loss	
				mAP	mmit mAP	mAP	mmit mAP	mAP	mmit mAP
TSN [58]	1x1x8	RGB	Kinetics-400	0.5598	0.6143	0.4627	0.5629	0.4838	0.5579
TSM [31]	1x1x8	RGB	✗	0.5662	0.6618	0.5721	0.6688	0.5470	0.6255
ST-GCN [63]	1x1x300	Pose	✗	0.5776	0.6692	**0.5868**	0.6865	0.5874	0.6719
PoseC3D [63]	1x1x300	Pose	✗	0.5498	0.6393	0.5556	0.6241	0.5358	0.6142
MMNet [65]	Late-Fusion	RGB+Pose+RoI	✗	**0.5948**	**0.6735**	0.5871	**0.6973**	**0.5894**	**0.6830**

Table 4. Composite error action recognition performance on *Set-2* by direct migration. Only the results of four models in RGB and pose modality are reported due to the limited space. Significant performance degradation can be observed compared to the results in Table 3.

Model	mAP	Δ	mmit mAP	Δ
TSN [58]	0.5598	—	0.6143	—
w/ ImagineNet-FC	**0.6259**	↑ 6.61%	**0.6893**	↑ 8.50%
TSM [31]	0.5662	—	0.6618	—
w/ ImagineNet-FC	**0.7053**	↑ 13.91%	**0.7566**	↑ 9.48%
ST-GCN [63]	0.5776	—	0.6692	—
w/ ImagineNet-FC	**0.6404**	↑ 6.28%	**0.7115**	↑ 4.23%
MMNet [65]	0.5948	—	0.6735	—
w/ ImagineNet-FC	**0.6927**	↑ 9.79%	**0.7478**	↑ 7.43%

Table 5. Performance comparison between direct migration and ImagineNet-FC. All model settings are consistent with Table 4.

Model	Variants	GFLOPs	mAP	mmit mAP
ImagineNet-FC	FC	0.001	0.7053	0.7566
ImagineNet-SA	SA	0.068	**0.7011**	0.7630
	SAx2	0.136	0.7007	**0.7656**
	SAx3	0.203	0.6995	0.7572
	w/o PosEmb	0.068	0.6822	0.7593
ImagineNet-CA	CA	0.068	0.6752	0.7346
	CA+SA	0.136	**0.6766**	**0.7406**
	CA+SAx2	0.203	0.6728	0.7377
	w/o PosEmb	0.068	0.6725	0.7339

Table 6. Performance and FLOPs comparison of the proposed three ImagineNet models and their variants based on the TSM.

5.2. Composite Error Action Recognition on *Set-2*

Taking *Set-1* as the training set and *Set-2* as the testing set, we can simulate the real CPR assessment. A naive approach is directly migrating the pre-trained model in single-class task to the composite error recognition task. Table 4 summarizes the performance of five selected models. All three losses cannot handle the huge gap between the two tasks, while the MMNet [65] achieves better migration performance through the fusion mechanism and bigger model size. The sharp decline in performance indicates that the new task has exceeded the representation capability of original models. Next, the deployment details and results of the ImagineNet will be introduced. Note that our core contribution is not to create a novel HAR model but to build a better composite error detector through existing models. Therefore, we adopt classic models (TSN, TSM, ST-GCN) and the SOTA model MMNet to instantiate ImagineNets for ensuring the reproducibility and stability, instead of those sophisticated methods such as TimeSFormer and PoseC3D.

Implementation Details. All ImagineNet models are trained for 60 epochs through the SGD optimizer. The learning rate is set to 0.001 initially and attenuated by 0.1 at 20 and 40-th epochs. The temporal length T is set to 8. Only the models trained with CE loss are explored. *mAP* and *mmit mAP* [35] are adopted as metrics for evaluation.

Quantitative Analysis. Table 5 compares the ImagineNet-FC model with the vanilla migration method. Through the *Imagine* mechanism, the ImagineNet-FC significantly improves the composite error recognition performance under restricted supervision, regardless of the input modality. In particular, the ImagineNet-FC brings 13.91% *mAP* and 9.48% *mmit mAP* improvement on TSM. The performance and computational complexity of ImagineNet-SA, ImagineNet-CA, and their variants are summarized in Table 6. The results reveal that the ImagineNet-SA outperforms the other two models. The CA mechanism does not improve performance as well as SA. More layers and computational complexity will lead to overfitting. The *Positional Embedding* module is essential in ImagineNets because chronological information is indispensable for distinguishing these fine-grained error actions. In Figure 8, we explore the relationship between the number of error combinations and the final performance on *Set-2*. The *mmit_mAP* of ImagineNet-FC gradually decreases as the number of composite errors increases, which is consistent with our intuition that more complex error combinations imply higher task difficulty.

Qualitative Analysis. To explore how ImagineNet impacts the network, we visualize and compare the features generated by TSM and TSM *w/* ImagineNet-FC on *Set-2* in Figure 9(a&b). Macroscopically, features obtained by the direct migration are messy, while the ImagineNet can help the network reduce intra-class distance and expand inter-class distance. The enhancement of feature clustering confirms the effectiveness of the proposed ImagineNet. High-resolution t-SNE figures of more ImagineNet models are demonstrated in the supplementary materials.

5.3. Combination of Perspectives

As shown in Figure 1(a), the video capture system includes four views. It is not practical to use all perspectives in deployment, which will cause redundant computation. Four-perspective settings can help us discover the best combination and achieve the optimal performance-computation trade-off. We evaluate the performance of the ImagineNet-FC on all different perspectives combinations. The results are shown in Figure 7. Overall, the performance increases with combing more perspectives. Perspective #3 provides more valuable information, while #4 is the opposite. This discovery is of great value for subsequent deployment.

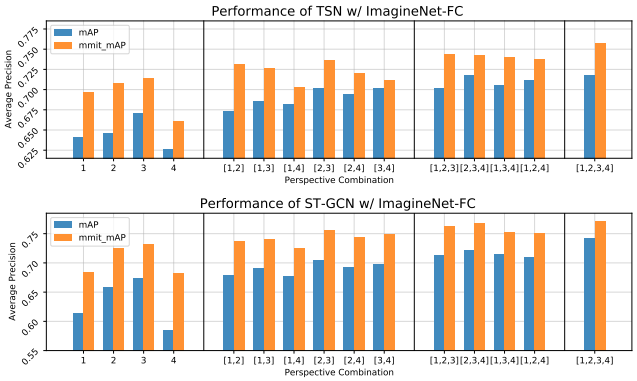


Figure 7. Performance of combining different perspectives. Different numbers of views are grouped by black dividing lines.

Model	Agg-1	Agg-2	Agg-3	mAP	mmit mAP
TSN [58]	–	–	–	0.5598	0.6143
w/ ImagineNet-FC	✘	✘	✘	0.6198	0.6738
	✔	✘	✘	**0.6259**	**0.6893**
	✘	✔	✘	0.6019	0.6775
	✘	✘	✔	0.6033	0.6725
TSM [31]	–	–	–	0.5662	0.6618
w/ ImagineNet-FC	✘	✘	✘	0.6871	0.7353
	✔	✘	✘	**0.7053**	**0.7566**
	✘	✔	✘	0.6434	0.7308
	✘	✘	✔	0.6569	0.7219
ST-GCN [63]	–	–	–	0.5776	0.6692
w/ ImagineNet-FC	✘	✘	✘	0.6374	0.7089
	✔	✘	✘	**0.6404**	**0.7115**
	✘	✔	✘	0.5783	0.6877
	✘	✘	✔	0.6159	0.6864

Table 7. Ablation studies on three feature aggregation strategies.

5.4. Ablation Studies

Ablation studies are conducted to explore the effectiveness of feature aggregation strategies. Table 7 summarizes the results of ImagineNet-FC and its variants based on TSN, TSM, and ST-GCN. Performance of the random weighted summation mechanism surpasses the vanilla method and other two bilinear pooling aggregation methods both in RGB and pose modes. This reveals that the proposed mechanism can generate richer feature combinations concisely and effectively, thus enabling ImagineNet to achieve better generalization performance on unseen error combinations.

5.5. Cross Modality Studies

In previous experiment settings, inputs of the ImaginNet belong to different categories but the same modality. The structure of ImagineNet inherently supports multi-modal data fusion. Taking TSM and ST-GCN as basic models, Table 8 compares the ImagineNet-CA with the Two-Stream fusion method, two bilinear pooling fusion methods, and MMNet under cross modality settings. The latency of these fusion models is reported by averaging 1,000 running times, while basic models are excluded. Results show that the ImagineNet-CA surpasses the other three multimodal fusion methods. Although BLOCK performs similarly to ImagineNet-CA, its latency is nearly 7.8x longer, which is

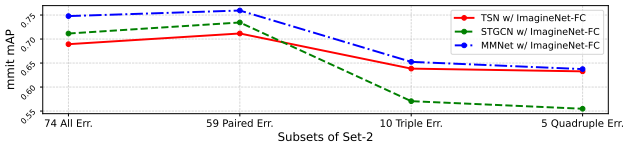
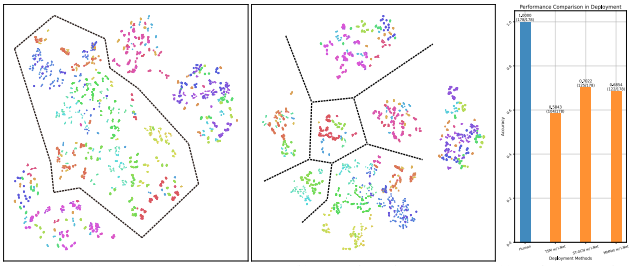


Figure 8. *mmit mAP* Performance on different subsets of *Set-2*.



(a) TSM (b) TSM w/ ImagineNet-FC (c) Deployment testing

Figure 9. (a&b) Feature visualization comparison via t-SNE on *Set-2*. Black auxiliary lines are marked for clarity. (c) System deployment and testing performance comparison.

Model	Modality	Latency (ms)↓	mAP	mmit mAP
TSM [31]	RGB	–	0.5662	0.6618
ST-GCN [63]	Pose	–	0.5776	0.6692
Two-Stream [51]	RGB+Pose	**0.1501**	0.6003	0.6815
CBP [23]	RGB+Pose	0.3043	0.7089	0.7506
BLOCK [7]	RGB+Pose	1.294	0.7107	**0.7675**
MMNet [65]	RGB+Pose+RoI	0.2479	0.6927	0.7478
w/ ImagineNet-CA	RGB+Pose	0.1642	**0.7110**	0.7515

Table 8. Cross modality studies on *RGB* and *Pose* information.

mainly caused by the complex approximate outer product computation. The Two-Stream fusion model can reduce latency but has poor performance.

5.6. System Deployment and Testing

To verify the performance of the ImagineNet in real deployment, we collaborate with the skill training center of a hospital to collect CPR videos in real training scenarios. After video collection and selection, we obtain a set contains 187 videos with various errors. Note that these videos are outside of the CPR-Coach. The comparison between human and ImagineNets is shown in Figure 9(c). Results show that the skeleton-based method has better generalization performance than the RGB-based method. The proposed system has preliminary capabilities to assist decision-making.

6. Conclusion

This paper proposes the CPR-Coach dataset, which supports fine-grained action recognition and composite error action recognition tasks in CPR training under restricted supervision. We extensively evaluate and compare the existing HAR models and propose different ImagineNet frameworks inspired by human cognition to improve the performance of the model under the composite error settings. Sufficient comparison and actual deployment experimental results verified the effectiveness of the proposed framework.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv:1609.08675*, 2016. 2

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. The kinetics human action video dataset. *arXiv:1705.06950*, 2017. 2

[3] Narges Ahmidi, Piyush Poddar, Jonathan Jones, Swaroop Vedula, Lisa Ishii, Gregory Hager, and Masaru Ishii. Automated objective surgical skill assessment in the operating room from unstructured tool motion in septoplasty. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 10(6):981–991, 2015. 2

[4] Hassan Alhajj, Mathieu Lamard, Pierre-henri Conze, Béatrice Cochener, and Gwenolé Quellec. Cataracts, 2021. 2

[5] American Heart Association. Highlights of the 2020 AHA guidelines for CPR and ECC. *Acesso em*, 4(07), 2021. 2

[6] Vivek Singh Bawa, Gurkirt Singh, Francis Kaping'a, Inna Skarga-Bandurova, and et. al. The SARAS endoscopic surgeon action detection (ESAD) dataset: Challenges and methods. *arXiv:2104.03178*, 2021. 2

[7] Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8102–8109, 2019. 5, 8

[8] Gedas Bertasius, Hyun Soo Park, Stella X. Yu, and Jianbo Shi. Am i a baller? basketball performance assessment from first-person videos. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 2

[9] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proc. International Conference on Machine Learning (ICML)*, 2021. 2, 6

[10] Vinay Bettadapura, Grant Schindler, Thomas Ploetz, and Irfan Essa. Augmenting bag-of-words: Data-driven discovery of temporal and structural information for activity recognition. In *CVPR*, 2013. 3

[11] Matthew Boutell, Jiebo Luo, Xipeng Shen, and Christopher Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004. 2

[12] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 2, 6

[13] Lin Chen, Qiang Zhang, Qiongjie Tian, and Baoxin Li. Learning skill-defining latent space in video-based analysis of surgical expertise - A multi-stream fusion approach. In *Medicine Meets Virtual Reality (MMVR)*, 2013. 2

[14] Lin Chen, Qiang Zhang, Peng Zhang, and Baoxin Li. Instructive video retrieval for surgical skill coaching using attribute learning. In *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, 2015. 3

[15] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 933–942, 2021. 2

[16] Krzysztof Dembczyński, Weiwei Cheng, and Eyke Hüllermeier. Bayes optimal multilabel classification via probabilistic classifier chains. In *Proc. International Conference on Machine Learning (ICML)*, 2010. 2

[17] Robert DiPietro, Colin Lea, Anand Malpani, Narges Ahmidi, Swaroop Vedula, Gyusung Lee, Mija Lee, and Gregory Hager. Recognizing surgical activities with recurrent neural networks. In *Proc. International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016. 2, 3

[18] Konstantin Dmitriev and Arie E. Kaufman. Learning multi-class segmentations from single-class datasets. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9493–9503, 2019. 2

[19] Haodong Duan, Yue Zhao, Kai Chen, Dian Shao, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. *arXiv:2104.13586*, 2021. 2, 6

[20] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 4

[21] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. 2, 6

[22] Yixin Gao, Swaroop Vedula, Carol Reiley, and *et al.* JHU-ISI gesture and skill assessment working set (JIGSAWS): A surgical activity dataset for human motion modeling. In *Proc. Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*, 2014. 1, 2, 3

[23] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 317–326, 2016. 5, 8

[24] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 2

[25] Ke-Wei Huang and Zhuolun Li. A multilabel text classification algorithm for labeling risk factors in sec form 10-k. *ACM Transactions on Management Information Systems*, 2 (3), 2011. 2

[26] Arnaud Huaulmé, Duygu Sarikaya, Kévin Le Mut, Fabien Despinoy, Yonghao Long, Qi Dou, Chin-Boon Chng, Wenjun Lin, Satoshi Kondo, Laura Bravo-Sánchez, Pablo Arbeláez, Wolfgang Reiter, Manoru Mitsuishi, Kanako Harada, and Pierre Jannin. Micro-surgical anastomose workflow recognition challenge report. *arXiv:2103.13111*, 2021. 2

[27] Gazi Islam, Kanav Kahol, John Ferrara, and Richard Gray. Development of computer vision algorithm for surgical skill assessment. In *Proc. International Conference on Ambient Media and Systems (ICAMS)*, 2011. 2

[28] Gazi Islam, Baoxin Li, and Kanav Kahol. An affordable real-time assessment system for surgical skill training. In *Proc. International Conference on Intelligent User Interfaces (IUI)*, 2013. 2

[29] Feng Kang, Rong Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1719–1726, 2006. 2

[30] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. Large-scale video classification with convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1725–1732, 2014. 2

[31] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 7082–7092, 2019. 2, 6, 7, 8

[32] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W. Tsang. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(11):7955–7974, 2022. 2

[33] Anand Malpani, Swaroop Vedula, Chi Chiung Grace Chen, and Gregory Hager. Pairwise comparison-based objective score for automated skill assessment of segments in a surgical task. *Information Processing in Computer-Assisted Interventions (IPCAI)*, 2014. 2

[34] J. A. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown. Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery*, 1997. 2

[35] Mathew Monfort, Bowen Pan, Kandan Ramakrishnan, Alex Andonian, Barry A Mcnamara, Alex Lascelles, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: learning and interpreting models for multi-action video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2021. 7

[36] Hirenkumar Nakawala, Roberto Bianchi, Laura Erica Pescatori, Ottavio De Cobelli, Giancarlo Ferrigno, and Elena De Momi. "deep-onto" network for surgical workflow and context recognition. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 2019. 2

[37] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78, 2022. 2

[38] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017. 2

[39] Paritosh Parmar and Brendan Tran Morris. Action quality assessment across multiple actions. In *Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[40] Paritosh Parmar and Brendan Tran Morris. What and how well you performed? a multitask learning approach to action quality assessment. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[41] Judea Pearl and Dana Mackenzie. The book of why : the new science of cause and effect. 361(6405):855.2–855, 2018. 2

[42] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Proc. European Conference on Computer Vision (ECCV)*, 2014. 2

[43] Duygu Sarikaya, Jason J Corso, and Khurshid A Guru. Detection and localization of robotic tools in robot-assisted surgery videos using deep neural networks for region proposal and detection. *IEEE transactions on medical imaging*, 36(7):1542–1549, 2017. 2

[44] Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. Cataract-101: Video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM Multimedia Systems Conference*, page 421–425, 2018. 2, 3

[45] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1010–1019, 2016. 2

[46] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, 2020. 2

[47] Hao Shao, Shengju Qian, and Yu Liu. Temporal interlacing network. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, pages 11966–11973, 2020. 2, 6

[48] Yachna Sharma, Vinay Bettadapura, Thomas Plotz, Nils Hammerla, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew McCaskie, and Irfan Essa. Video based assessment of OSATS using sequential motion textures. In *MMCAI*, 2014. 3

[49] Yachna Sharma, Thomas Plötz, Nils Hammerld, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew McCaskie, and Irfan Essa. Automated surgical osats prediction from videos. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2014. 2

[50] Yachna Sharma, Vinay Bettadapura, Thomas Plötz, Nils Hammerld, Sebastian Mellor, Roisin McNaney, Patrick Olivier, Sandeep Deshmukh, Andrew McCaskie, and Irfan Essa. Video based assessment of OSATS using sequential motion textures. In *Proc. Modeling and Monitoring of Computer Assisted Interventions (M2CAI)*, 2016. 2

[51] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. 2, 6, 8

[52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015. 2, 6

[53] Andru P. Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36 (1):86–97, 2017. 2

[54] Aleksandar Vakanski, Hyung pil Jun, David Paul, and Russell Baker. A data set of human body movements for physical rehabilitation exercises. page 1–15, 2018. 2, 3

[55] Laurens van der Maaten and Geoffrey Hinton. Viualizing data using t-sne. *Journal of Machine Learning Research*, 9: 2579–2605, 2008. 6

[56] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, and *et al.* Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *arXiv:2109.14956*, 2021. 2, 3

[57] Hua Wang, Chris Ding, and Heng Huang. Multi-label classification: Inconsistency and class balanced k-nearest neighbor. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2010. 2

[58] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proc. European Conference on Computer Vision (ECCV)*, pages 20–36, 2016. 2, 4, 5, 6, 7, 8

[59] Shunli Wang, Dingkang Yang, Peng Zhai, Chixiao Chen, and Lihua Zhang. TSA-Net: Tube self-attention network for action quality assessment. In *Proc. ACM International Conference on Multimedia (ACM-MM)*, 2021. 2

[60] Shunli Wang, Dingkang Yang, Peng Zhai, Qing Yu, Tao Suo, Zhan Sun, Ka Li, and Lihua Zhang. A survey of video-based action quality assessment. In *Proc. International Conference on Networking Systems of AI (INSAI)*, pages 1–9, 2021. 1, 2

[61] Chengming Xu, Yanwei Fu, Bing Zhang, Zitian Chen, Yu-Gang Jiang, and Xiangyang Xue. Learning to score figure skating sport videos. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 2020. 2

[62] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2939–2948, 2022. 2

[63] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2018. 2, 6, 7, 8

[64] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 591–600, 2020. 6

[65] Bruce X.B. Yu, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith C.C. Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(3):3522–3538, 2023. 2, 6, 7, 8

[66] Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l1 optical flow. In *Proc. of DAGM Conference on Pattern recognition*, 2007. 4

[67] Luca Zappella, Benjamín Béjar, Gregory Hager, and René Vidal. Surgical gesture classification from video and kinematic data. *Medical image analysis*, 2013. 2

[68] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2018. 5

[69] Min-Ling Zhang and Kun Zhang. Multi-label learning by exploiting label dependency. In *Proc. ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, page 999–1008, 2010. 2

[70] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 26(8):1819–1837, 2014. 2

[71] Qiang Zhang and Baoxin Li. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In *International ACM workshop on Medical Multimedia Analysis and Retrieval*, 2011. 2, 3

[72] Qiang Zhang and Baoxin Li. Relative hidden markov models for evaluating motion skill. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2, 3

[73] Qiang Zhang and Baoxin Li. Relative hidden markov models for video-based evaluation of motion skills in surgical training. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(6):1206–1218, 2015. 2

[74] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric Sarin, Thomas Ploetz, Mark Clements, and Irfan Essa. Automated video-based assessment of surgical skills for training and evaluation in medical schools. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 11(9): 1623–1636, 2016. 2, 3

[75] Aneeq Zia, Chi Zhang, Xiaobin Xiong, and Anthony Jarc. Temporal clustering of surgical activities in robot-assisted surgery. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 12(7):1171–1178, 2017. 2

[76] Aneeq Zia, Yachna Sharma, Vinay Bettadapura, Eric Sarin, and Irfan Essa. Video and accelerometer-based motion analysis for automated surgical skills assessment. *International Journal of Computer Assisted Radiology and Surgery (IJCARS)*, 13(3):443–455, 2018. 2