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

1. Supplementary for the CPR-Coach

Comparison with Other Datasets. Table 3 compares CPR-Coach with other existing medical action analysis datasets. Most of the traditional research is to rate the action of the subjects. For example, *Expert/Novice* in [35] and *Expert/Intermediate/Novice* in [37]. These methods model the assessment task as simple two or three-category classification problems. At the same time, the diversity of these datasets is limited. These research only contains specific two or three-type operations such as *Suturing*, *Knot-Tying* and *Needle-Passing* in [8]. Unfortunately, most researchers do not release the proposed datasets, which limits the development of the field of medical action assessment. Most open-source research [1, 11, 12, 16, 17, 19, 20, 22, 25, 27, 29, 33, 34] focus on surgical workflow recognition without assessing the quality of actions. The CPR-Coach dataset contains rich fine-grained incorrect action categories, various visual perspectives and sufficient video samples. The CPR-Coach dataset will be released later to enhance research in medical skill analysis tasks.

Details of the CPR-Coach Dataset. Figure 3 shows the filtering strategy in paired-composite errors by taking ten error actions as the main cases. All deleted combinations are marked. In Figure 3(a), errors about hands cannot co-occur, so two co-occurrences are deleted. In Figure 3(c), it is unlikely that errors such as *Excessive Pressing* and *Bending Arms* will occur when the *Single Hand* exists, and these combinations are deleted. Note that in Figure 3(h), the *Insufficient Pressing* error may combine with any other errors so that all combinations can be received. All deletions have been carefully considered and carefully reviewed by emergency doctors. Figure 5 shows all combinations of the 10 triple- and 5 quadruple-composite error actions studied in this paper. Figure 4 shows all the composite error actions in detail. Note that temporal-related errors such as *Insufficient Pressing*, *Slow Frequency*, and *Random Position Pressing* are not evident in images.

2. Results on SOTA Video Backbones

The core contribution of this study is **NOT** to create a novel/SOTA HAR model but to build a better composite error detector through existing models under the *Single-class Training & Multi-class Testing* settings. Therefore, we focus on exploring the performance of some classic action recognition frameworks in the main text. These frameworks are concise and easy to replicate. For the completeness of the research, we supplement the experiments with Video Swin Transformer [15], Vi-ViT [2], and MViTv2 [13] as SOTA video backbones. All models are trained with *Cross-Entropy* loss. Table 1 lists the performance of three SOTA video backbones under the single-error setting and direct migration setting. These powerful backbones are able to handle error recognition tasks well, but they cannot achieve satisfactory performance under composite error settings. This is are consistent with the conclusions in Table 3&4 in the main text. Table 2 shows that with the help of the proposed ImagineNet, all three backbones achieve significant performance improvements. Especially, performance of the Video Swin Transformer has improved by 13.86% in *mAP* and 9.37% in *mmit mAP*, respectively. The improvement in performance confirms the effectiveness of the proposed framework.

Model	Config	Pre-train	Single-class Recogn.	
			Top-1	Top-3
Vi-ViT [2]	base-16x2	Kinetics-400	0.9814	1.0000
MViTv2 [13]	base-32x3x1	Kinetics-400	0.9867	0.9980
Video Swin [15]	base-32x2x1	Kinetics-400	0.9918	1.0000
Model	**Config**	**Pre-train**	**Direct Migration**	
			mAP	mmit mAP
Vi-ViT [2]	base-16x2	Kinetics-400	0.5582	0.6651
MViTv2 [13]	base-32x3x1	Kinetics-400	0.5715	0.6740
Video Swin [15]	base-32x2x1	Kinetics-400	0.5696	0.6701

Table 1. Composite error action recognition performance on SOTA video backbones.

Model	mAP	Δ	mmit mAP	Δ
Vi-ViT [2]	0.5582	—	0.6651	—
w/ ImagineNet-FC	**0.6587**	↑ 10.05%	**0.7523**	↑ 8.72%
MViTv2 [13]	0.5715	—	0.6740	—
w/ ImagineNet-FC	**0.6869**	↑ 11.54%	**0.7461**	↑ 7.21%
Video Swin [15]	0.5696	—	0.6701	—
w/ ImagineNet-FC	**0.7082**	↑ 13.86%	**0.7638**	↑ 9.37%

Table 2. Performance comparison between direct migration and ImagineNet-FC on SOTA video backbones.

Research Theme	Dataset	#Actions	Modality	#Videos	#Views	Evaluation Type	Available
Skills in Laparoscopic Surgery	FLS-ASU [35]	1	RGB	28	2	Skill Rating	✘
	Zhang *et al.* [36]	1	RGB	546	1	Skill Rating	✘
	Chen *et al.* [5]	3	RGB	720	2	Skill Rating	✘
Basic Surgical Skills Assessment	Sharma *et al.* [23]	2	RGB	33	1	OSATA Score	✘
	Bettadapura *et al.* [4]	3	RGB	64	2	Skill Rating	✘
	Zia *et al.*[37]	2	RGB	104	1	Skill Rating	✘
Skills on *Da Vinci* Surgical Systems	MISTIC-SL [6]	4	RGB+Kinematics	49	1	Skill Rating	✘
	JIGSAWS [8]	3	RGB+Kinematics	103	1	Skill Rating	✔
Exercise Rehabilitation Assessment	UI-PRMD [26]	10	RGB+Kinematics	1,000	1	Skill Rating	✔
Surgical Workflow Recognition	Cataract-101 [22]	10	RGB	101	1	Workflow Recogn.	✔
	Hei-Chole [29]	7	RGB	33	1	Workflow Recogn.	✔
	HeiCo [17]	0	RGB	30	1	Workflow Recogn.	✔
	RARP45 [27]	8	RGB	45	1	Workflow Recogn.	✔
	Cholec80 [25]	7	RGB	80	1	Workflow Recogn.	✔
	GastricBypass337 [33]	10	RGB	337	1	Workflow Recogn.	✘
	Gastrectomy461 [34]	8	RGB	461	1	Workflow Recogn.	✘
	Nephrec9 [19]	10	RGB	1,262	1	Workflow Recogn.	✔
	CATARACTS [1]	21	RGB	50	1	Tools Recogn.	✔
	CholecT50 [20]	10	RGB	50	1	Triplet Recogn.	✔
	Laparo425 [12]	9	RGB	425	1	Early Recogn.	✘
	PETRAW [11]	6	RGB+Kinematics	90	1	Workflow Recogn.	✔
	DESK [16]	7	RGB+Kinematics	2,897	1	Workflow Recogn.	✔
Cardiopulmonary Resuscitation	CPR-Coach (Ours)	14+74	RGB+Flow+Pose	5,664	4	Error Recogn.	✔

Table 3. Comparison with existing medical action analysis datasets. Due to the inheritance of these research, we classify these datasets according to different research themes.

3. Supplementary Experimental Results

Supplementary Experiment of TSN *w/* ImagineNet. Table 4 lists the performance and FLOPs comparison of the proposed three ImagineNet models and their variants based on the TSN [30]. The ImagineNet-SA outperforms the other two models, which is consistent with the results in Table 6 in the main text. Table 5 lists the cross modality results on RGB and pose information based on the TSN. The performance and latency are consistent with the results on TSM.

t-SNE Visualization on *Set-2*. Page 7 summarizes the experimental results of TSN and TSM, and Page 8 summarizes the results of TPN and ST-GCN. Results in Table 6&7 suggest that the ImagineNet-FC significantly improves the network's performance on composite error action recognition tasks. Figure 6&7 show the t-SNE visualization of TSN and TSM, respectively. Large intervals are marked in dotted lines for clarity. Apparent margins reveal the effectiveness of the proposed ImagineNet-FC. Figure 8&9 show the t-SNE visualization of TSN and TSM, respectively. The performance improvement can be observed on TPN but is not apparent on ST-GCN. This is consistent with the performance comparison in Table 7.

System Demonstration *Set-2*. The proposed CPR composite error action recognition system is shown in Figure 10, 11, 12, 13. The demonstration video was also uploaded as part of the supplementary materials.

4. CBP and BLOCK Models

As the representative of bilinear pooling aggregation methods, CBP [7] and BLOCK [3] models are equipped with the natural characteristics of aggregating features. We compared the above two methods in *5.4 Ablation Studies* with the proposed random weighted summation mechanism (Figure 2). In the *5.5 Cross Modality Studies*, the performance of ImagineNet-CA is compared with these models. Limited by space, these methods are not introduced in detail in the main text.

The brief introduction and implementation details of these methods are as follow.

Compact Bilinear Pooling. Bilinear pooling is the merging operation of a series of local image descriptors. Given a set of local descriptors $\mathcal{X} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{|\mathcal{X}|}, \boldsymbol{x}_s \in \mathbb{R}^C)$, the bilinear pooling generates a global representation through

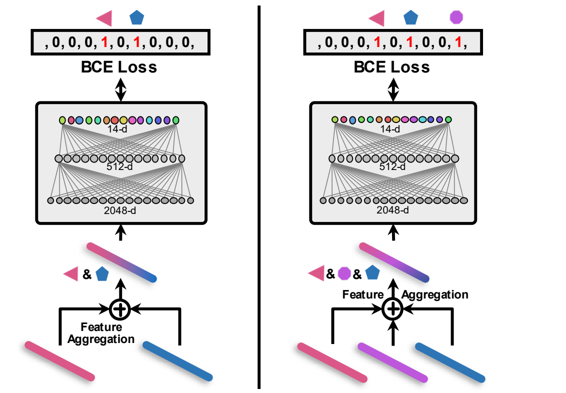
$$B(\mathcal{X}) = \sum_{s\in\mathcal{S}} \boldsymbol{x}_s \boldsymbol{x}_s^T. \tag{1}$$

Given two sets of local descriptors: $\mathcal{X}$ and $\mathcal{G}$, the dot product of two features is represented as

$$\langle vec(B(\mathcal{X})), vec(B(\mathcal{G}))\rangle = \sum_{s\in\mathcal{S}}\sum_{g\in\mathcal{G}} \langle \boldsymbol{x}_s, \boldsymbol{y}_g\rangle^2. \tag{2}$$

The CBP method [7] aims to find a low dimensional projection function $\Phi(x) \in \mathbb{R}^d$, where $d \ll c^2$ and satisfy

$$\sum_{s\in\mathcal{S}}\sum_{g\in\mathcal{G}} \langle \boldsymbol{x}_s, \boldsymbol{y}_g\rangle^2 \approx \sum_{s\in\mathcal{S}}\sum_{g\in\mathcal{G}} \langle \Phi(\boldsymbol{x}_s), \Phi(\boldsymbol{y}_g)\rangle. \tag{3}$$



(a) Paired-composite errors. **(b)** Triple-composite errors.

Figure 1. Demonstration of the ImagineNet-FC handles two and three inputs.

The low-dimensional approximation operation dramatically reduces the computational complexity. Tensor Sketch Projection [21] is adopted as the dimension reduction method. **Block-superdiagonal Tensor Decomposition**. In [3], Ben-younes *et al.* introduced bilinear pooling methods to perform multimodal fusion in the VQA and VRD tasks. A bilinear model takes two features as input and projects them into a k-dimensional space with tensor products

$$\boldsymbol{b} = \boldsymbol{\mathcal{T}} \times \boldsymbol{x} \times \boldsymbol{y}, \tag{4}$$

where $\boldsymbol{x} \in \mathbb{R}^{C_1}$, $\boldsymbol{y} \in \mathbb{R}^{C_2}$, and $\boldsymbol{b} \in \mathbb{R}^{K}$. $\forall k \in [1, K]$,

$$b_k = \sum_{i=1}^{C_1} \sum_{j=1}^{C_2} \boldsymbol{\mathcal{T}}_{ijk} \cdot x_i \cdot y_j. \tag{5}$$

To reduce the number of parameters and computational complexity, $\boldsymbol{\mathcal{T}}$ is decomposed through block-term decomposition in $rank(L, M, N)$ terms:

$$\boldsymbol{\mathcal{T}} = \sum_{r=1}^{R} \boldsymbol{\mathcal{D}}_r \times \boldsymbol{A}_r \times \boldsymbol{B}_r \times \boldsymbol{C}_r, \tag{6}$$

where $\forall r \in [1, R]$, $\boldsymbol{\mathcal{D}}_r \in \mathbb{R}^{L \times M \times N}$, $\boldsymbol{A}_r \in \mathbb{R}^{C_1 \times L}$, $\boldsymbol{B}_r \in \mathbb{R}^{C_2 \times M}$, and $\boldsymbol{C}_r \in \mathbb{R}^{K \times N}$. By adopting structural constraint to $\boldsymbol{\mathcal{T}}$, the projection process is parametrized through a block-superdiagonal tensor $\boldsymbol{\mathcal{D}}^{bd} \in \mathbb{R}^{LR \times MR \times NR}$.

5. Evaluation Metrics

Due to space limitations, we did not provide a specific introduction to metrics in the main text. The *mAP* adopted in this paper refers to the *macro mAP* in [18], which denotes the average of the mean average precision for each class:

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C}. \tag{7}$$

The *mmit mAP* refers to the *micro mAP* in [18], which denotes the mean average precision over all videos:

$$mmit\ mAP = \frac{\sum_{j=1}^{N} AP_j}{N}. \tag{8}$$

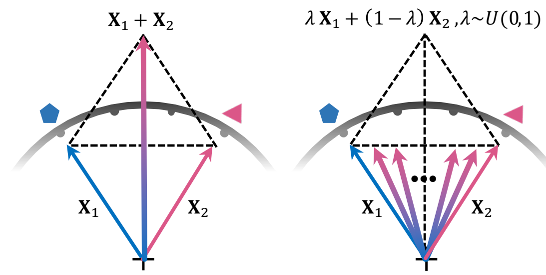


Figure 2. Visualization of the vanilla additive mechanism and the proposed weighted feature summation mechanism.

Note that AP_i denotes the average precision over the i-th class, while AP_j denotes the average precision for the j-th sample.

6. Limitations and Discussions

As the first study on fine-grained error action recognition and AQA in CPR training, this work inevitably has some limitations. The diversity and complexity of the CPR-Coach dataset remains to be improved. Standard CPR consists of several stages (*e.g.*, electric defibrillation, artificial respiration), while only the external cardiac compression is studied due to the time and scale limitation. Nevertheless, the CPR-Coach has reached 450GB and 2.2M frames, which allows us to make some preliminary algorithm exploration. We look forward to some valuable and promising research directions in the future. We hope these prospects will bring some inspiration to the readers.

• **Diversity & Complexity of the Dataset**. The CPR-Coach dataset only considers the external cardiac compression action in CPR. In the future, we will continue to co-operate with the training center of the hospital to enrich the dataset. There is still huge potential exploration space for complex and multi-stage medical action analysis.

• **Data Generation**. The data acquisition of medical action datasets is highly professional, which makes it challenging to expand the scale of datasets. Deep Generative Models (DGMs) such as GAN [9] and Diffusion Models [10] have achieved excellent performance on image/video generation tasks. It will be very interesting to combine these generative models with medical action analysis scenarios to generate high-quality, large-scale datasets in different modalities.

• **Combination with Language Models**. Based on CPR-Coach, this paper design a discriminative *Coach* with the ability to identify single and composite errors. However, a real coach can give verbal guidance and advice to beginners. By combining language models [28] with the assessment tasks, we will design a more perfect and human-centered system like a real coach.

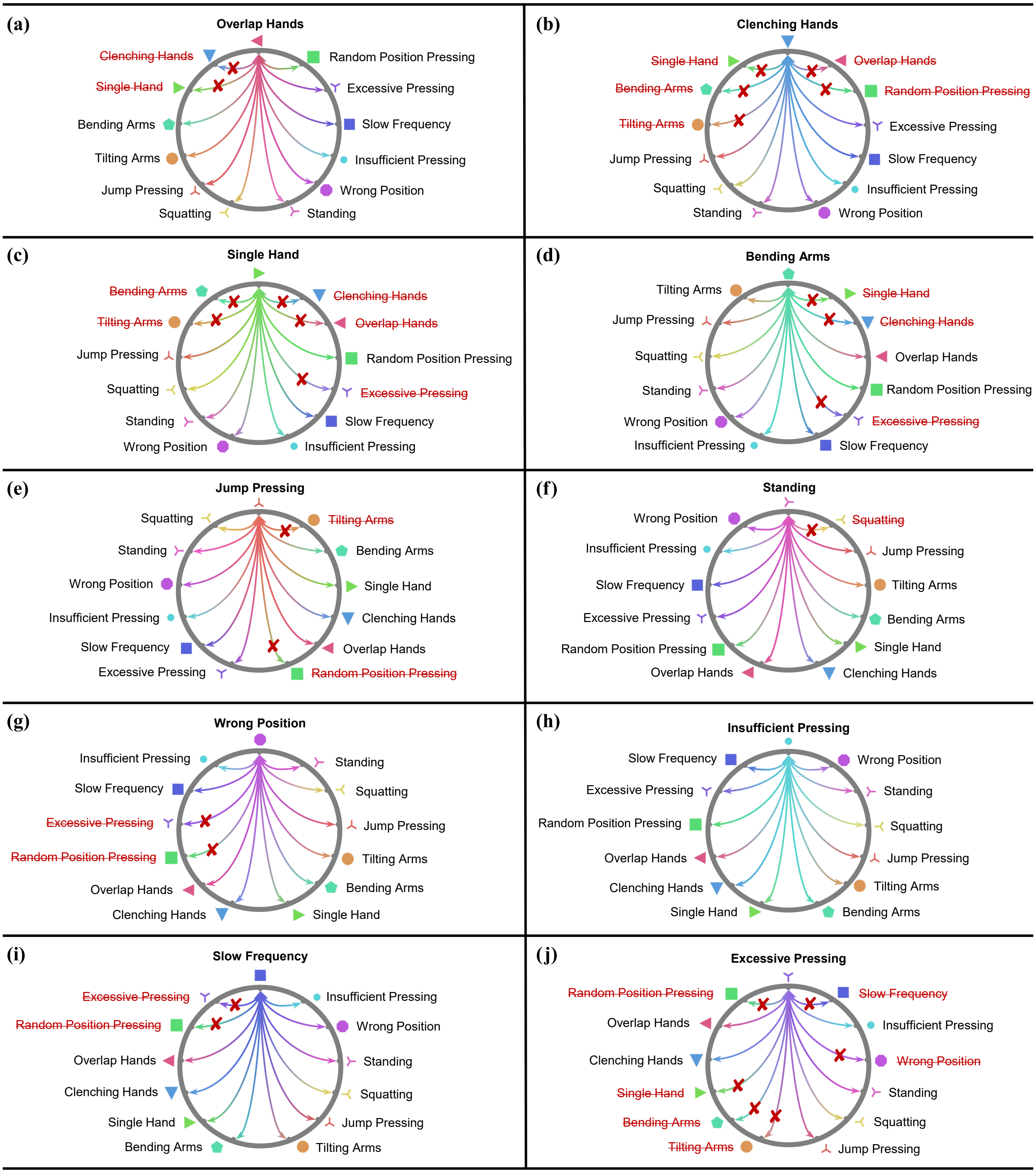
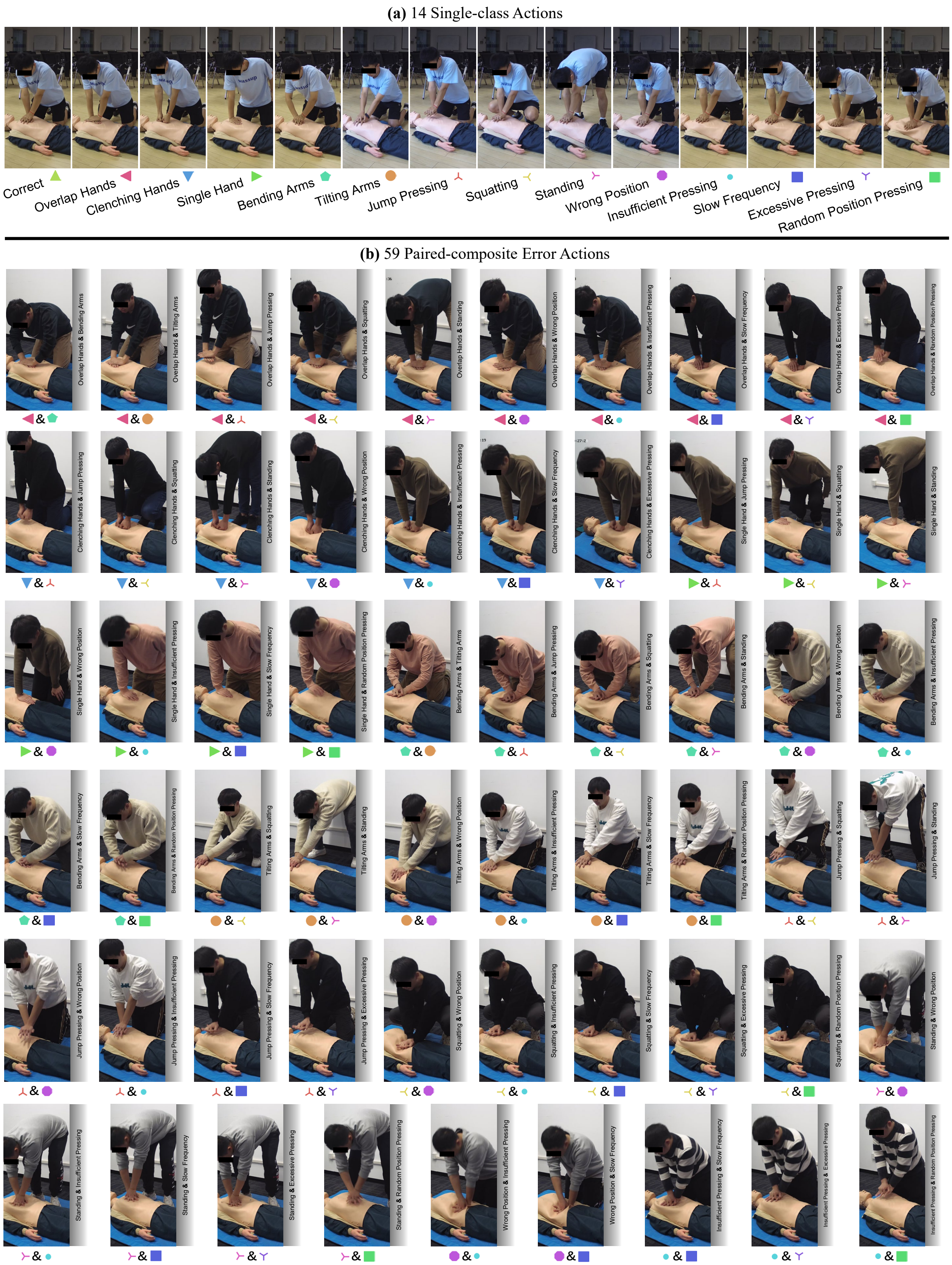
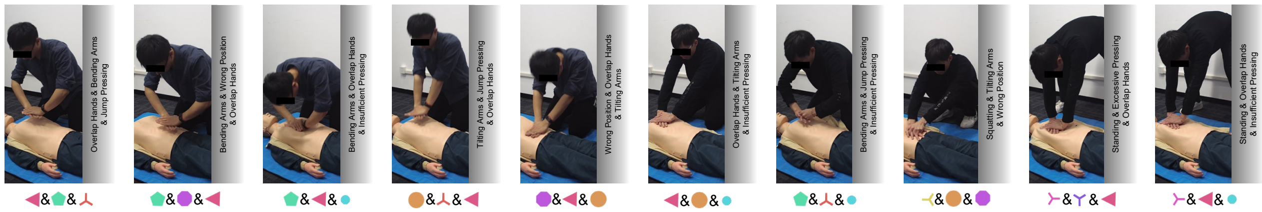


Figure 3. Ten error actions are selected as the main class for demonstrating the selection strategy. All combinations of each main class are enumerated and listed in detail. Impossible co-occurrences in each subfigure are flagged via red delete symbols. Three omitted actions also follow this selection strategy.


(a) 14 Single-class Actions
Correct
Overlap Hands
Clenching Hands
Single Hand
Bending Arms
Tilting Arms
Jump Pressing
Squatting
Standing
Wrong Position
Insufficient Pressing
Slow Frequency
Excessive Pressing
Random Position Pressing
(b) 59 Paired-composite Error Actions

(c) 10 Triple-composite Error Actions



(d) 5 Quadruple-composite Error Actions

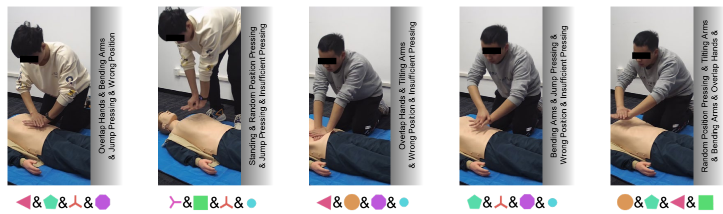


Figure 4. All single-class and composite error examples studied in this paper. Marks and annotations are also listed in detail.

Multi-Comp.	Composite Error Actions	Marks
10 Triple-Composite Errors	Overlap Hands **&** Bending Arms **&** Jump Pressing	& &
	Bending Arms **&** Wrong Position **&** Overlap Hands	& &
	Bending Arms **&** Overlap Hands **&** Insufficient Pressing	& &
	Tilting Arms **&** Jump Pressing **&** Overlap Hands	& &
	Wrong Position **&** Overlap Hands **&** Tilting Arms	& &
	Overlap Hands **&** Tilting Arms **&** Insufficient Pressing	& &
	Bending Arms **&** Jump Pressing **&** Insufficient Pressing	& &
	Squatting **&** Tilting Arms **&** Wrong Position	& &
	Standing **&** Excessive Pressing **&** Overlap Hands	& &
	Standing **&** Overlap Hands **&** Insufficient Pressing	& &
5 Quadruple-Comp. Errors	Overlap Hands **&** Bending Arms **&** Jump Pressing **&** Wrong Position	& & &
	Standing **&** Random Position Pressing **&** Jump Pressing **&** Insufficient Pressing	& & &
	Overlap Hands **&** Tilting Arms **&** Wrong Position **&** Insufficient Pressing	& & &
	Bending Arms **&** Jump Pressing **&** Wrong Position **&** Insufficient Pressing	& & &
	Tilting Arms **&** Bending Arms **&** Overlap Hands **&** Random Position Pressing	& & &

Figure 5. All combinations of the 10 triple- and 5 quadruple-composite error actions studied in this paper.

Model	Variants	GFLOPs	mAP	mmit mAP
ImagineNet-FC	FC	0.001	0.6259	0.6893
ImagineNet-SA	SA	0.068	0.6426	0.7049
	SAx2	0.136	**0.6450**	**0.7131**
	SAx3	0.203	<u>0.6436</u>	<u>0.7086</u>
	w/o PosEmb	0.068	0.6305	0.6906
ImagineNet-CA	CA	0.068	0.6307	0.6933
	CA+SA	0.136	**0.6347**	<u>0.7005</u>
	CA+SAx2	0.203	<u>0.6335</u>	**0.7046**
	w/o PosEmb	0.068	0.6281	0.6953

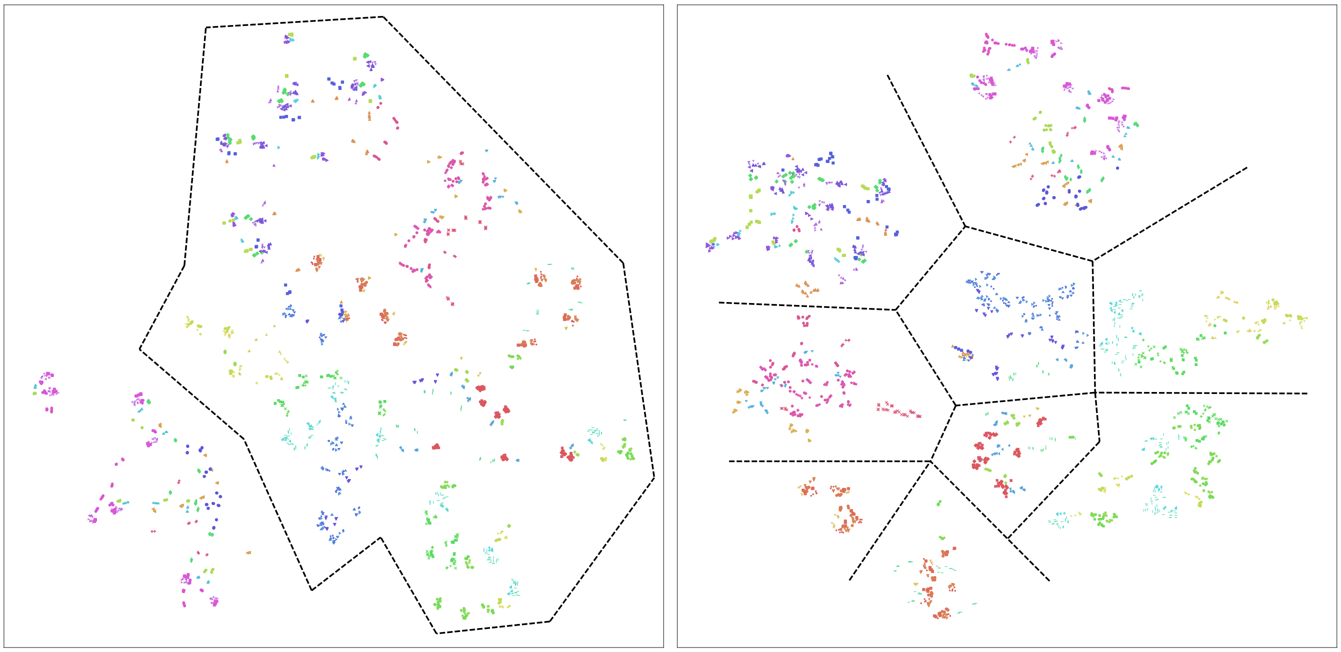
Table 4. Performance and FLOPs comparison of the proposed three ImagineNet models and their variants based on the TSN.

Model	Modality	Latency (ms)↓	mAP	mmit mAP
TSN [30]	RGB	–	0.5598	0.6143
ST-GCN [31]	Pose	–	0.5776	0.6692
Two-Stream [24]	RGB+Pose	**0.1426**	0.5915	0.6823
CBP [7]	RGB+Pose	0.3032	0.7066	0.7460
BLOCK [3]	RGB+Pose	1.254	<u>0.7094</u>	<u>0.7597</u>
w/ ImagineNet-CA	RGB+Pose	<u>0.1612</u>	**0.7133**	**0.7641**

Table 5. Cross modality studies on *RGB* and *Pose* information.

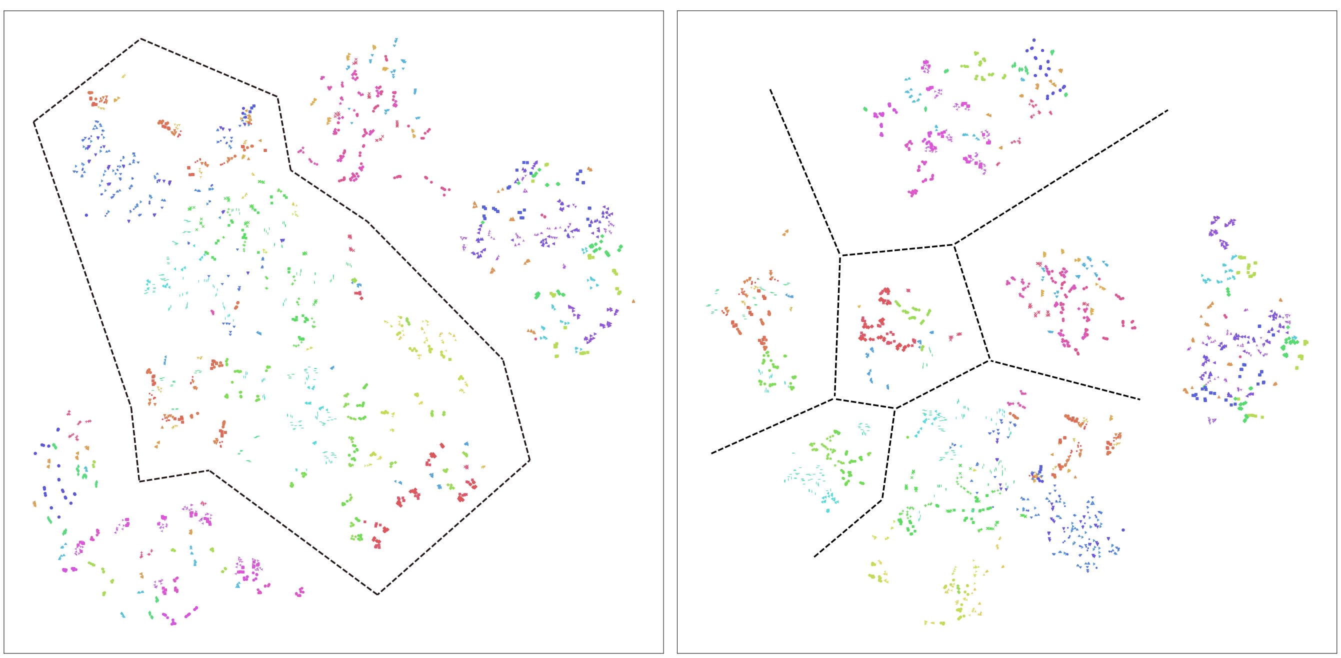
Model	mAP	Δ	mmit mAP	Δ
TSN [30]	0.5598	—	0.6143	—
w/ ImagineNet-FC	**0.6259**	↑ 6.61%	**0.6893**	↑ 8.50%
TSM [14]	0.5662	—	0.6618	—
w/ ImagineNet-FC	**0.7053**	↑ 13.91%	**0.7566**	↑ 9.48%

Table 6. Performance comparison between direct migration and ImagineNet-FC on TSN and TSM.



(a) TSN (b) TSN *w/* ImagineNet

Figure 6. t-SNE feature visualization comparison of TSN on CPR-Coach *Set-2*.

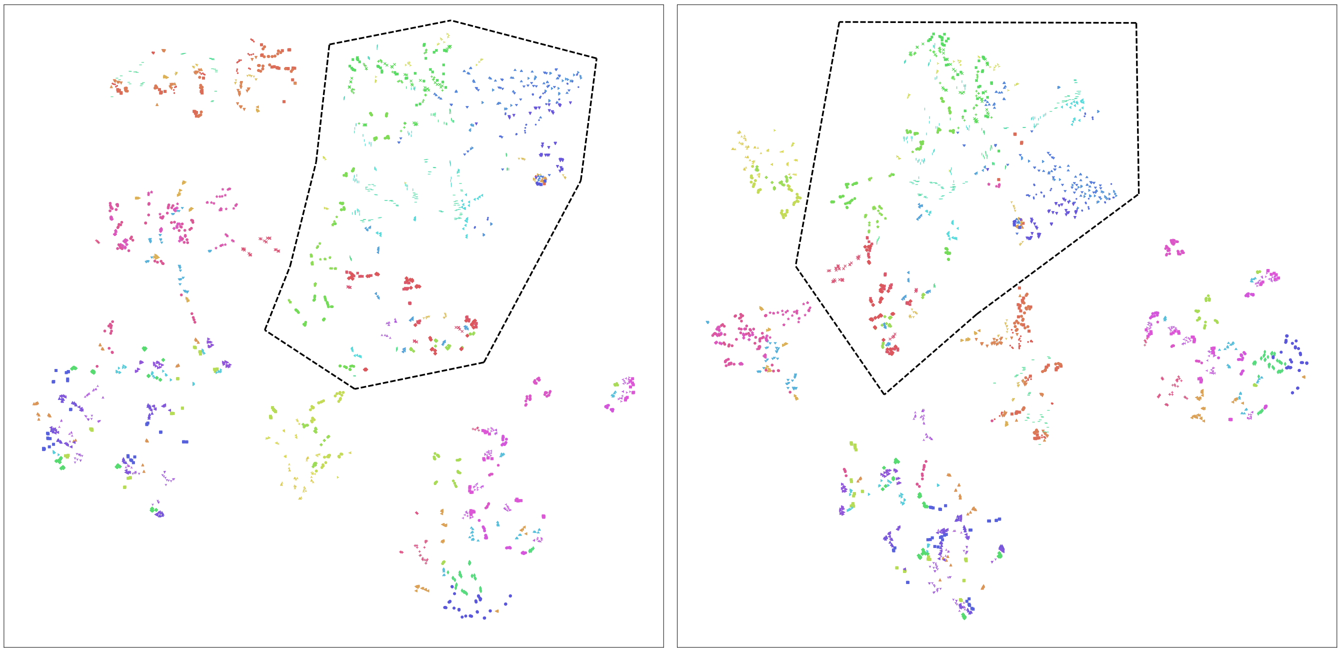


(a) TSM (b) TSM *w/* ImagineNet.

Figure 7. t-SNE feature visualization comparison of TSM on CPR-Coach *Set-2*.

Model	mAP	Δ	mmit mAP	Δ
TPN [32]	0.6250	—	0.7016	—
w/ ImagineNet-FC	**0.7094**	↑ 8.44%	**0.7620**	↑ 6.04%
ST-GCN [31]	0.5776	—	0.6692	—
w/ ImagineNet-FC	**0.6404**	↑ 6.28%	**0.7115**	↑ 4.23%

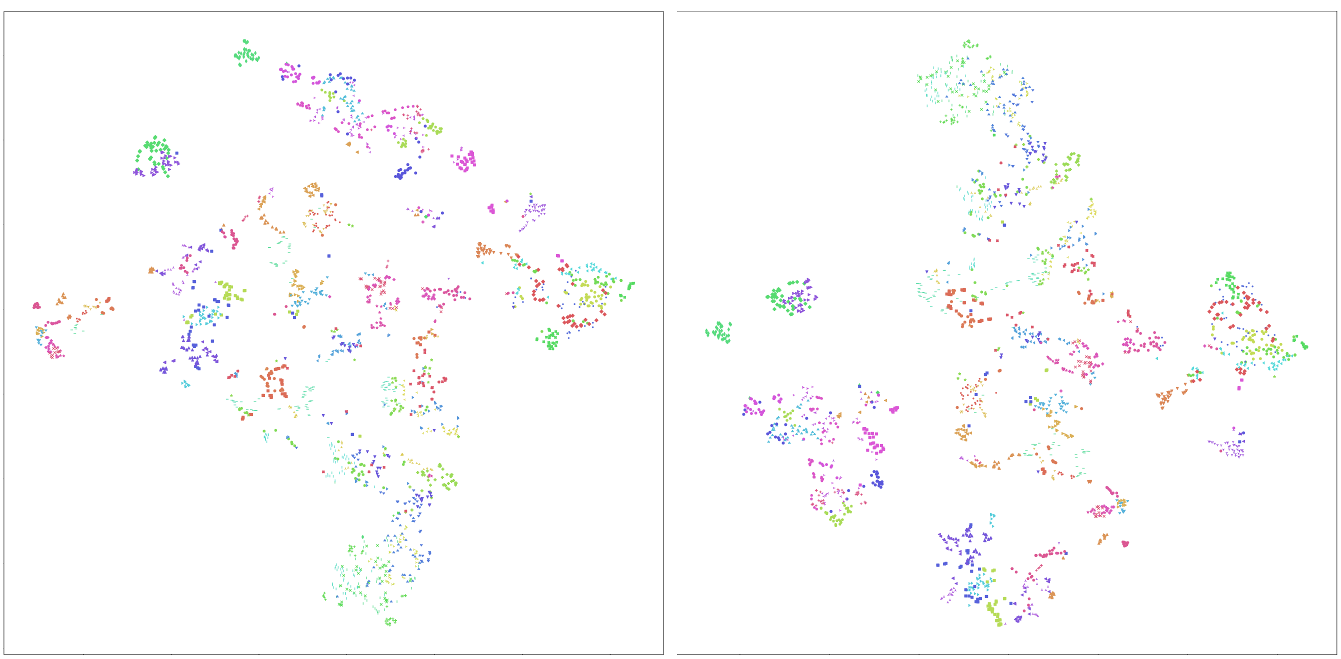
Table 7. Performance comparison between direct migration and ImagineNet-FC on TPN and ST-GCN.



(a) TPN

(b) TPN *w/* ImagineNet

Figure 8. t-SNE feature visualization comparison of TPN on CPR-Coach *Set-2*.



(a) ST-GCN

(b) ST-GCN *w/* ImagineNet

Figure 9. t-SNE feature visualization comparison of ST-GCN on CPR-Coach *Set-2*.

Figure 10. Single error actions recognition results.

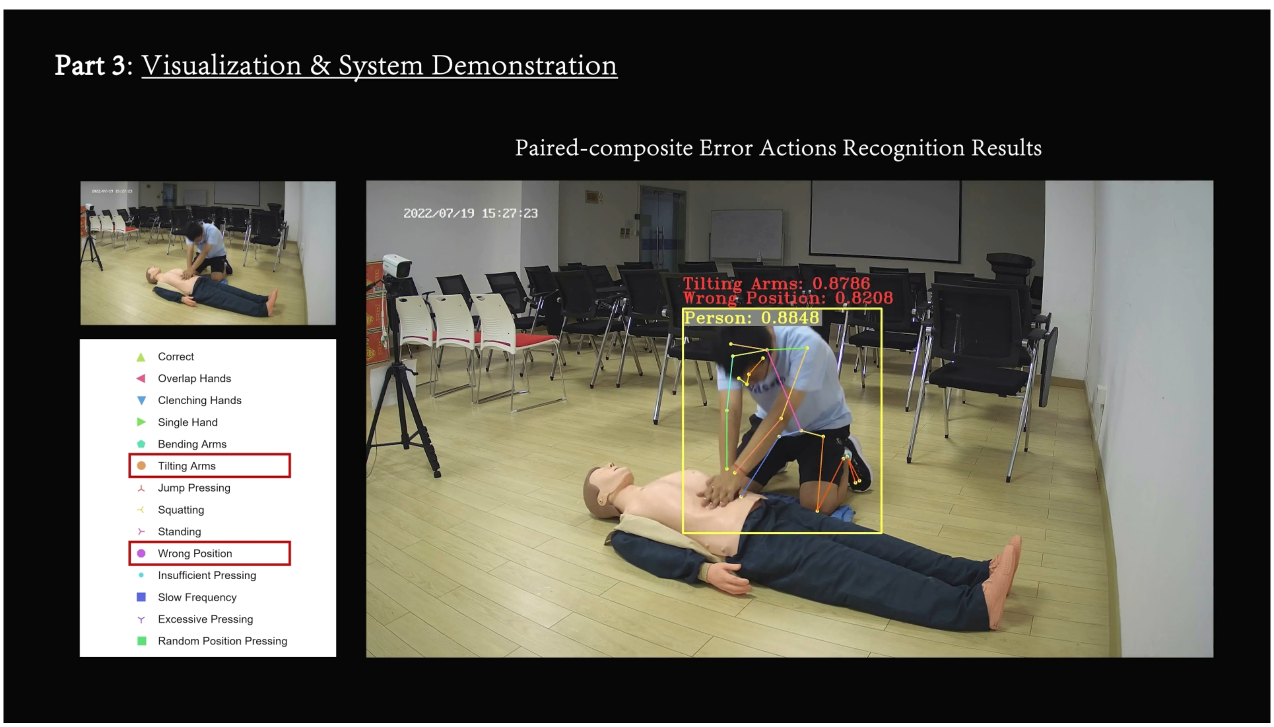


Figure 11. Paired-composite error actions recognition results.

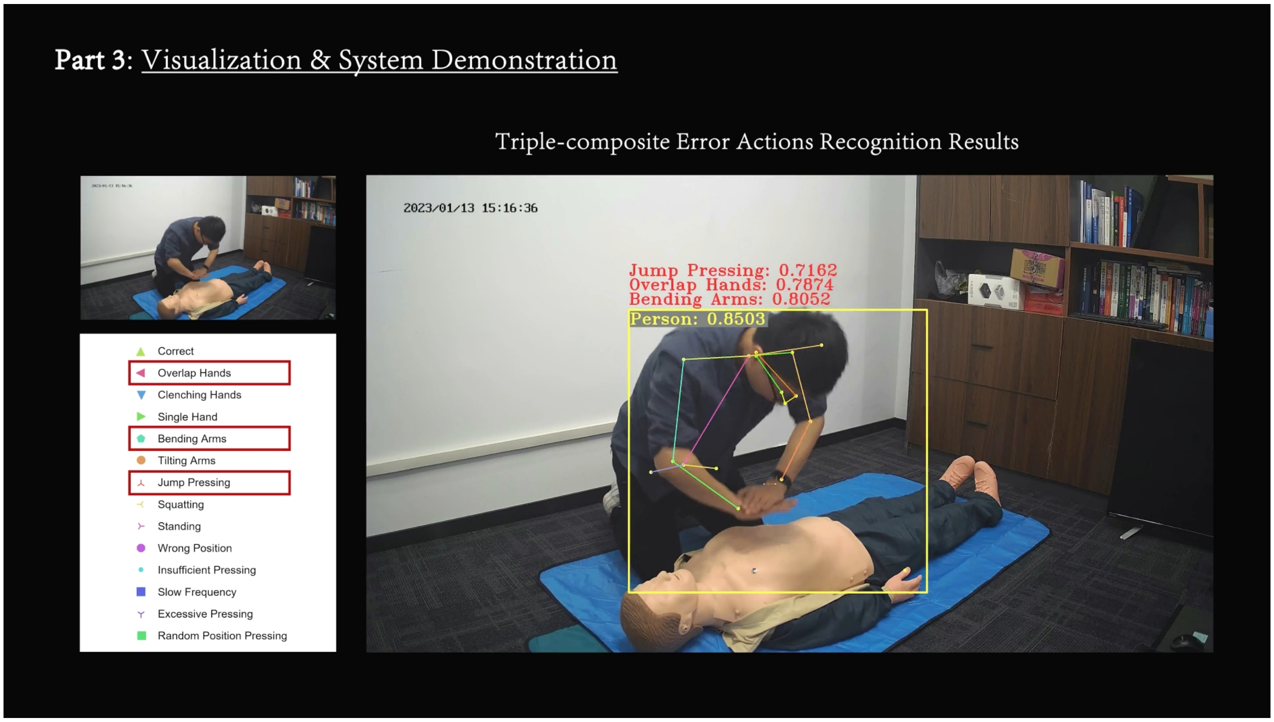


Figure 12. Triple-composite error actions recognition results.

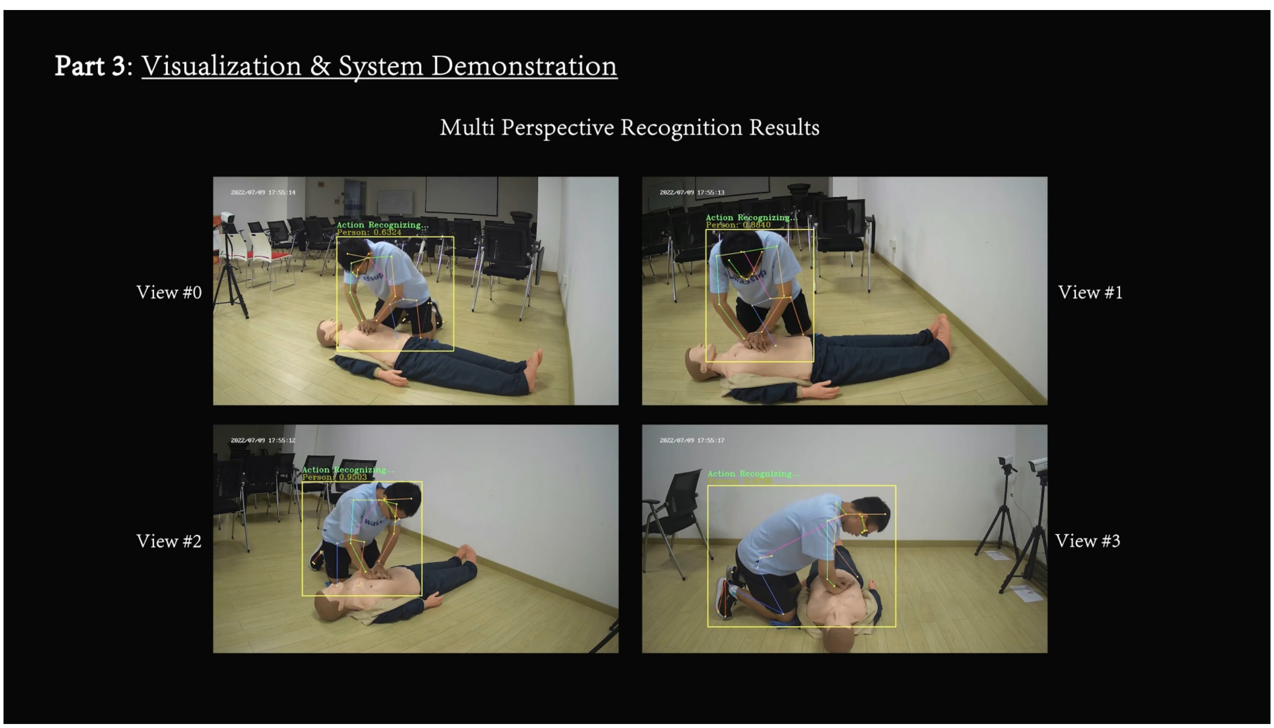


Figure 13. Multi perspective recognition results.

References

[1] Hassan Al Hajj, Mathieu Lamard, Pierre-Henri Conze, Soumali Roychowdhury, Xiaowei Hu, Gabija Maršalkaitė, Odysseas Zisimopoulos, Muneer Ahmad Dedmari, Fenqiang Zhao, Jonas Prellberg, et al. Cataracts: Challenge on automatic tool annotation for cataract surgery. *Medical Image Analysis*, 52:24–41, 2019. 1, 2

[2] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6816–6826, 2021. 1

[3] 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*, volume 33, pages 8102–8109, 2019. 2, 3, 6

[4] 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. 2

[5] Lin Chen, Qiang Zhang, Peng Zhang, and Baoxin Li. Instructive video retrieval for surgical skill coaching using attribute learning. In *ICME*, 2015. 2

[6] 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 *MICCAI*, 2016. 2

[7] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *CVPR*, pages 317–326, 2016. 2, 6

[8] 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 *M2CAI*, 2014. 1, 2

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[11] Arnaud Huaulmé, Kanako Harada, Quang-Minh Nguyen, Bogyu Park, Seungbum Hong, Min-Kook Choi, Michael Peven, Yunshuang Li, Yonghao Long, Qi Dou, et al. Peg transfer workflow recognition challenge report: Does multi-modal data improve recognition? *arXiv preprint arXiv:2202.05821*, 2022. 1, 2

[12] Siddharth Kannan, Gaurav Yengera, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Future-state predicting lstm for early surgery type recognition. *IEEE Transactions on Medical Imaging*, 39(3):556–566, 2019. 1, 2

[13] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, pages 4794–4804, 2022. 1

[14] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 2019. 7

[15] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3192–3201, 2022. 1

[16] Naveen Madapana, Md Masudur Rahman, Natalia Sanchez-Tamayo, Mythra V Balakuntala, Glebys Gonzalez, Jyothsna Padmakumar Bindu, LN Vishnunandan Venkatesh, Xingguang Zhang, Juan Barragan Noguera, Thomas Low, et al. Desk: A robotic activity dataset for dexterous surgical skills transfer to medical robots. In *IROS*, pages 6928–6934, 2019. 1, 2

[17] Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, et al. Heidelberg colorectal data set for surgical data science in the sensor operating room. *Scientific data*, 8(1):101, 2021. 1, 2

[18] 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. *TPAMI*, 2021. 3

[19] 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. *IJCARS*, 14:685–696, 2019. 1, 2

[20] 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*, 78:102433, 2022. 1, 2

[21] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *ACM SIGKDD*, page 239–247, 2013. 3

[22] Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. Cataract-101: video dataset of 101 cataract surgeries. In *Proc. of ACM Multimedia Systems Conf.*, pages 421–425, 2018. 1, 2

[23] 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. 2

[24] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014. 6

[25] 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 TMI*, 36(1):86–97, 2016. 1, 2

[26] 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

[27] Beatrice Van Amsterdam, Isabel Funke, Eddie Edwards, Stefanie Speidel, Justin Collins, Ashwin Sridhar, John Kelly, Matthew J Clarkson, and Danail Stoyanov. Gesture recogni-

tion in robotic surgery with multimodal attention. *IEEE TMI*, 41(7):1677–1687, 2022. 1, 2

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[29] Martin Wagner, Beat-Peter Müller-Stich, Anna Kisilenko, Duc Tran, Patrick Heger, Lars Mündermann, David M Lubotsky, Benjamin Müller, Tornike Davitashvili, Manuela Capek, et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Medical Image Analysis*, 86:102770, 2023. 1, 2

[30] 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 *ECCV*, 2016. 2, 6, 7

[31] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 6, 8

[32] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *CVPR*, 2020. 8

[33] Bokai Zhang, Julian Abbing, Amer Ghanem, Danyal Fer, Jocelyn Barker, Rami Abukhalil, Varun Kejriwal Goel, and Fausto Milletarì. Towards accurate surgical workflow recognition with convolutional networks and transformers. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 10(4):349–356, 2022. 1, 2

[34] Bokai Zhang, Amer Ghanem, Alexander Simes, Henry Choi, Andrew Yoo, and Andrew Min. Swnet: Surgical workflow recognition with deep convolutional network. In *Medical Imaging with Deep Learning*, pages 855–869, 2021. 1, 2

[35] Qiang Zhang and Baoxin Li. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In *MMAR*, 2011. 1, 2

[36] Qiang Zhang and Baoxin Li. Relative hidden markov models for evaluating motion skill. In *CVPR*, 2013. 2

[37] 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. *IJCARS*, 11(9):1623–1636, 2016. 1, 2