GaussMedAct: Multivariate Gaussian based Representation Learning for Medical Action Evaluation

Anonymous ICCV submission

Paper ID 4288

Abstract

Fine-grained action evaluation in medical vision faces unique challenges due to the unavailability of comprehensive datasets, stringent precision requirements, and insufficient spatiotemporal dynamic modeling of very rapid actions. To support development and evaluation, we introduce CPREval-6k, a multi-view, multi-label medical action benchmark containing 6,372 expert-annotated videos with 22 clinical labels. Using this dataset, we present GaussMedAct, a multivariate Gaussian encoding framework, to advance medical motion analysis through adaptive spatiotemporal representation learning. Multivariate Gaussian Representation projects the joint motions to a temporally scaled multi-dimensional space, and decomposes actions into adaptive 3D Gaussians that serve as tokens. These tokens preserve motion semantics through anisotropic covariance modeling while maintaining robustness to spatiotemporal noise. Hybrid Spatial Encoding, employing a Cartesian and Vector dual-stream strategy, effectively utilizes skeletal information in the form of joint and bone features. The proposed method achieves 92.1% Top-1 accuracy with real-time inference on the benchmark, outperforming the ST-GCN baseline by +5.9% accuracy with only 10% FLOPs. Cross-dataset experiments confirm the superiority of our method in robustness.

1. Introduction

Recent advances in medical vision systems have underscored the critical need for fine-grained and rapid motion understanding in time-sensitive clinical scenarios, particularly cardiopulmonary resuscitation (CPR) [30, 36]. The quality of CPR directly determines survival outcomes in cardiac arrest emergencies [41], where compression depth and frequency are strongly correlated with survival rate [12, 41]. Using feedback techniques can effectively improve results [57]. However, the current manual CPR assessment suffers from fundamental limitations revealed in our experiments. The results showed that human evaluators achieve only 74.8% accuracy in detecting critical errors such as incomplete chest recoil and frequency deviations (tested in 23 certified practitioners). Existing computer vision systems fail to capture sub-centimeter motion deviations (*e.g.* 5 cm compression depth [44]) and millisecond-level frequency variations (*e.g.* 115 ± 5 bpm requirements) [46]. These challenges stem from persistent gaps in visual computing: modeling anatomical causality in spatiotemporal dynamics, preserving clinical interpretability, and achieving medical-grade temporal precision.

Current action recognition approaches that focus on images (RGB-based) or graphs (skeleton-based) face limitations. Methods based on RGB (*e.g.* TimeSformer [1]) that rely on pre-trained backbones suffer from drawbacks such as fundamentally lacking anatomical modeling capabilities while incurring prohibitive computational latency. Skeleton-based methods (*e.g.* ST-GCN [59]) discard motion semantics through rigid temporal pooling operations and remain vulnerable to noise. A good solution requires a balance between accuracy, explainability, and latency [19]. These limitations directly affect the reliability of any method in the real world, while suboptimal CPR guidance techniques do not improve lifesaving effectiveness [33].

Recent advances in Gaussian Splatting have demonstrated efficient and high-quality rendering in computer graphics, suggesting that sparse Gaussian distributions can effectively represent complex 3D point clouds [6, 23]. However, its potential remains unexplored for spatiotemporal representations. This gap presents an opportunity given two key observations: First, Gaussian Mixture Models [37] naturally align with the probabilistic nature of motion point-cloud distributions. Second, conjecture on spatiotemporal interest points reveals that human motion dynamics can be decomposed into temporal components [26]. Building on these insights, we propose a Multivariate Gaussian Representation (MGR) for robust spatiotemporal skeleton learning, which models temporal evolution of keypoints as probability distributions to achieve compact and noise-resistant action representations.

Our design is further motivated by fundamental principles of motion perception. Psychological studies show that sparse 2D point-light displays can convey strong action impressions through basic kinematic patterns [21]. Complementary motion semantics exist in both absolute joint positions and relative bone kinematics [40]. To fully exploit this dual nature, we introduce a Hybrid Spatial Encoding (HSE) through a Cartesian-Vector dual-stream architecture reconciles absolute anatomical positioning with relative kinematic patterns - an approach particularly effective for modeling rapid human motion.

- CPREVAL-6K: The largest multi-view clinical dataset featuring synchronized RGB-skeleton streams and expert-validated multi-label annotations. Our benchmark contains 6,372 chest compression clips in 22 categories, each video hierarchically annotated with a primary critical error and secondary factors.
- GAUSSMEDACT: An end-to-end framework that combines MGR and HSE. By generating precision action token tensors, GAUSSMEDACT enables multiple downstream tasks including real-time classification and the generation of evaluation reports.

2. Related Work

Action Recognition Datasets. The field of action recognition has witnessed rapid development in benchmark datasets, with the emergence of large-scale datasets such as HMDB-51 [25], UCF101 [43], Something-Something [18], AVA v2.2 [27], and Kitchens-100 [11]. Although multimodal datasets such as Penn Action [63] and NTU RGB+D [38] that incorporate skeleton data have been introduced, these general-purpose collections cover extensive action categories without specializing in domain-specific scenarios such as medical action recognition. Although medical-oriented datasets such as FLS-ASU [62], JIGSAWS [17] and Mistic-SL [13] have been proposed, they remain limited in scale and category diversity, failing to meet fine-grained classification standards or provide sufficient discriminative capacity. A critical challenge in constructing medical action datasets lies in acquiring domain expert-validated annotations to ensure clinical authenticity - existing resources generally lack both large-scale collection and fine-grained annotation tailored for specific medical tasks.

Recent Advances in Medical Action Evaluation. Recent years have witnessed significant progress in computer vision-based medical action evaluation, including CPR. Earlier methods used LSTM, CNN with wearable sensors [5, 51, 58]. Pandurangan *et al.* [34] demonstrated the potential of HAR in various domains, including sports and safety. Karácsony *et al.* [22] discussed the complexity and scarcity of fine-grained datasets in clinical motion recognition. Li *et al.* [28] achieved real-time feedback through OpenPose/YOLO integration, but introduced clinical impracticalities due to wristband dependency. The ImagineNet framework [48] addressed multi-label classification using single-class training paradigms, showing promising results, but exhibit granularity gaps in motion-phase labeling. The field's progress remains constrained by data fitness. While newer datasets (Ko *et al.* [24]) address recency, their clinical utility is limited by complex architecture and static image training paradigms. Despite a surge of publications in recent two years, fundamental challenges remain at dataset lacking clinically-validated annotations.

General Trends in Human Action Recognition. Human Action Recognition (HAR) surveys [35, 64] highlight broader methodological advances. In RGB-based recognition, CNN-based architectures such as C3D [45] and SlowFast [16] achieved good results. Attention-based architectures such as TimeSformer [1] and MMNet [3] then dominate the area. Graph convolutional networks (GCNs) remain prevalent for skeleton data analysis, with ST-GCN [59] and its variants [7–9, 14, 31, 40, 42, 54], pioneering GCN applications in HAR. However, these approaches exhibit inherent limitations in temporal-scale adaptability, particularly for very rapid subsecond motion semantics.

Emerging Methodologies and Insights. Recent innovations provide significant inspiration. While CPR-Coach [48] advances medical action recognition through its 13-class multimodal framework. However, since the error annotations are not verified by experts and have limited categories, the dataset inadequately addresses real-world requirements. JMDA [50] using Skeleton MixFormer [52] architectures validated the rationality of projecting 3D skeleton data into 2D latent spaces. PoseC3D [15] demonstrated alternative approaches to GCNs for robust spatiotemporal feature learning against pose estimation noise. Yu *et al.* [60] improved the applicability of medical AI through LLM finetuning, showcasing the potential of language models in clinical reasoning. These advances collectively shape our technical framework, particularly in addressing temporal resolution limitations and integrating anatomical semantics with clinical interpretability.

3. CPREVAL-6K

The proposed dataset CPREVAL-6K comprises 6,372 manually annotated videos documenting manual chest compression procedures, captured from multiple viewpoints. Through consultations with emergency physicians and team expertise, we identified 21 different error categories in six medically critical aspects of chest compression techniques, including ‘hand position’, ‘arm angle’, ‘body posture’, ‘compression depth’, ‘operation frequency’, and ‘body positioning’, as shown in Figure 1. Each CPR video is annotated with one primary error designation and multiple secondary error labels. See supplementary material Sec. S3 for details of the dataset.

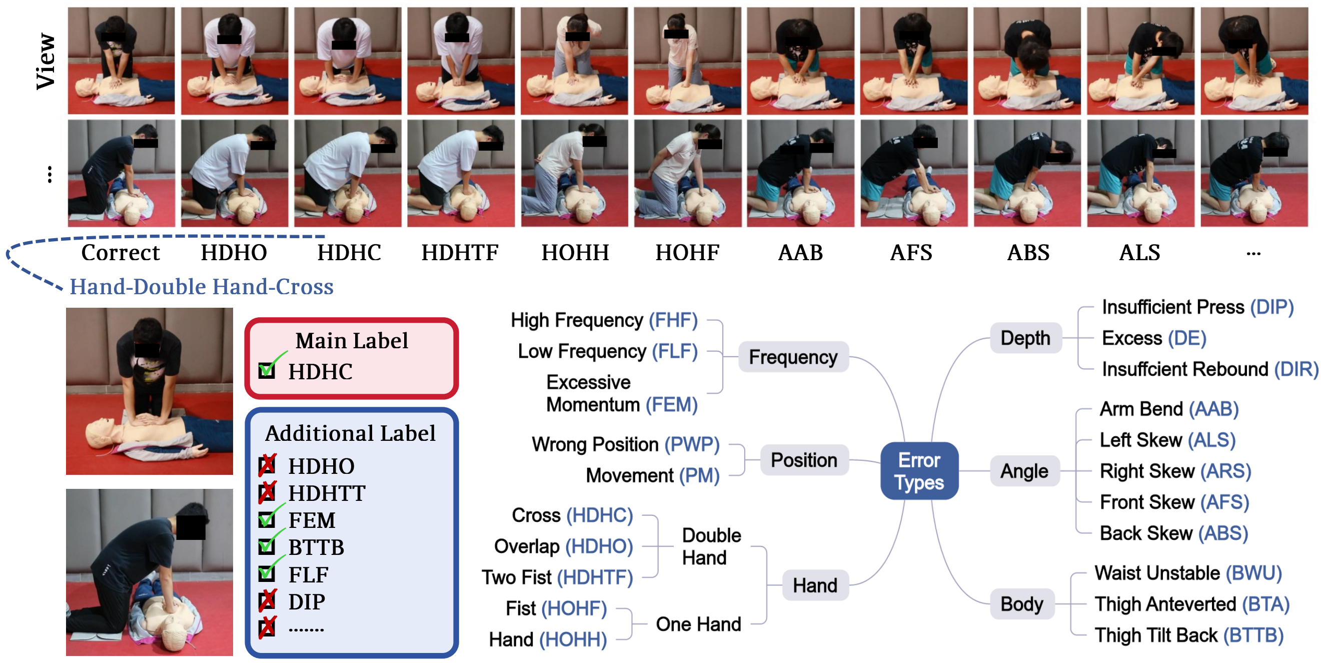

Figure 1. **Dataset Overview.** A multi-view CPR dataset with hierarchical error annotation, comprising six primary error classes and 21 fine-grained sub-classes. Each instance includes one primary error label and multiple secondary labels for compound error analysis.

Data Collection. We implemented a multiview camera recording system with synchronized initiation to ensure temporal alignment between perspectives. To validate dataset robustness, 6 volunteers who had not received professional CPR training and 6 American Heart Association (AHA) Heartsaver ® CPR-certified volunteers were recruited for standardized error simulation. The participants performed equal repetitions of each predefined error category, establishing a balanced distribution for the experimental validity. The entire dataset collection process is under the supervision of professionals to ensure that their actions meet the data collection standards.

Hierarchical Annotation. During data verification, we observed the involuntary co-occurrence of secondary errors during primary error simulations, necessitating our hierarchical annotation scheme of primary-secondary error labeling. A quality assurance protocol (QAP) involved 10 certified annotators who passed inter-rater reliability testing through rigorous Inter-Annotator Agreement evaluation. Annotation guidelines were strictly enforced under the supervision of a certified instructor who conducted final label verification, ensuring consistency and eliminating conspicuous annotation errors.

Data Analysis. The annotated dataset exhibits a significant class imbalance on the primary labels. As shown in Figure 7 in the supplementary materials, among the primary error labels, the *Depth-Insufficient Press* type accounts for 9.4% (298 instances), making it the most frequent error type. This distribution arises from the Explicit Error Prioritization annotation protocol, where each sample is assigned only one dominant primary error along with multiple secondary errors. Among the error categories (illustrated in the supplementary material Figure 8), *Depth-Insufficient Press* emerges as the predominant error type (29.1%, 298 instances), aligning with clinical studies [2] that identify monitoring challenges in chest compression quality.

The distribution of secondary labels under the primary label reveals frequent co-occurrence patterns between specific error types, as visualized in the supplementary materials Fig. 9. Preliminary correlation analysis identifies statistically significant relationships. *Freq-Excessive Momentum* and *Depth-Excess* exhibit a Pearson correlation coefficient of 0.52, while *Wrong Position* and *Position-Movement* show 0.39, consistent with biomechanical intuition.

To systematically uncover hierarchical error dependencies, we implement an enhanced Apriori algorithm ($min\_support = 0.025$, $min\_confidence = 0.25$) for the mining of association rules. From the supplementary material Tab. 4, we have the following key findings:

- **Strong Association:** Rule 5 (M:*Position-Movement* → A:*Position-Wrong Position*) demonstrates an exceptional association, achieving 77.6% confidence and 17.4× lift. This quantifies the linkage where compression point instability directly induces positional errors.
- **Kinematic Chain Coupling:** Upper-limb anomalies (M:*Angle-Arm Bend*) correlate with both posterior thigh

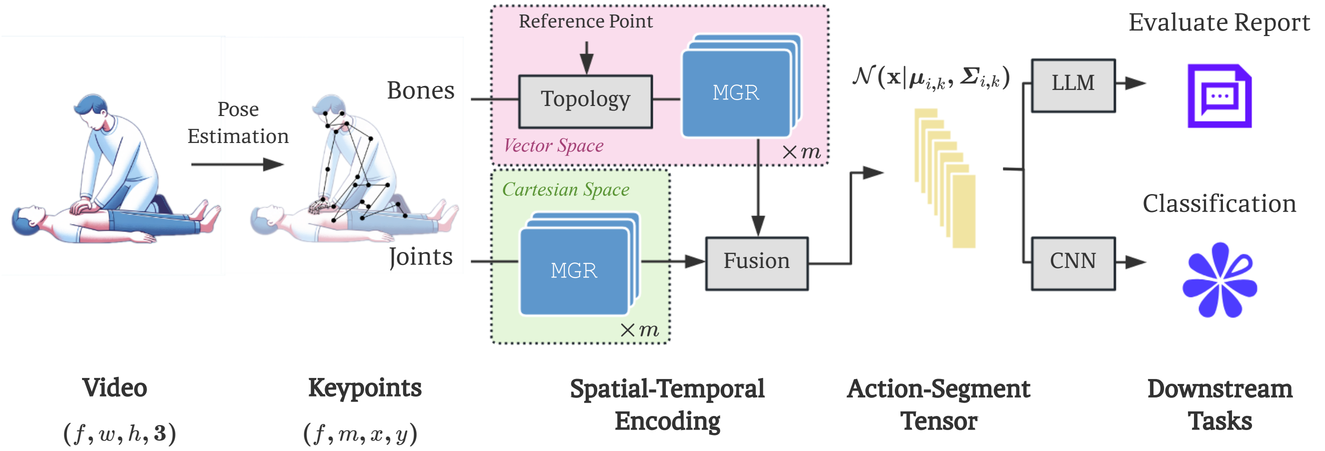

Figure 2. **Schematic of the proposed GAUSSMEDACT Pipeline.** Input data undergoes cartesian and vector based dual-stream encoding and pass through MGR to generate gaussians. Through feature fusion, action tokens are finally generated for downstream tasks.

tilt (A:*Body-Thigh Tilt Back*, 38.8% confidence) and insufficient compression depth (A:*Depth-Insufficient Press*, 38.4% confidence), revealing whole-body kinetic chain interactions during CPR execution.

- **Hierarchical Error Propagation:** M→A rules exhibit higher confidence (38.8–77.6%) compared to A→A co-occurrence rules (26.1–32.7%), indicating that primary errors exert strong causative drives on secondary errors.

Statistical data indicate that manual chest compressions constitute a coordinated technical action of the entire body. A single error in execution can propagate additional errors, ultimately resulting in suboptimal CPR outcomes. This highlights the critical importance of mastering the correct posture of CPR to ensure effective performance.

4. GAUSSMEDACT

We propose GAUSSMEDACT, a dual-stream spatiotemporal encoding framework for medical action recognition (see Fig. 2). The overall pipeline is summarized as follows.

- **Spatial dimension.** Human key points are extracted using pose estimation, and their characteristics are decoupled into complementary fluids from joint and bone. These two streams are processed separately and fused at a later stage to capture spatial dependencies and mitigate *collinearity* issues.
- **Temporal dimension.** MGR is introduced to process joints and bones independently. MGR captures action tokens along the temporal dimension and encodes them into Gaussian distributions, thereby modeling temporal dynamics and alleviating noise from pose estimation.
- **Feature Fusion.** Different fusion strategies used to integrate dual-stream features.
- **Downstream Tasks.** Two downstream tasks are introduced, including report generation and action classification. The use of label smoothing loss enhances the generalization performance.

4.1. Multivariate Gaussian Representation

Rationale. In the field of skeleton-based action recognition, GCN architectures (*e.g.* ST-GCN [59], CTR-GC [7]) typically rely on local convolution in temporal modeling. However, these methods often lack the ability to capture the temporal dynamics of human actions.

Inspired by the breakthroughs in Gaussian splatting [23] from the field of computer graphics, which has demonstrated remarkable performance in rendering tasks, we reexamined the temporal modeling problem for Gaussian splatting. In graphics rendering, Gaussian splatting can represent a highly dense point cloud in a vast spatial domain using only a small number of Gaussian distributions. This insight informs us that an intricate set of original spatial points can, in fact, be effectively described using significantly fewer key points.

We extend this idea to temporal encoding. In many human actions, the movements of keypoints are inherently continuous and can be represented by a compact set of temporal segments. Critical transition points, such as the start and end of accelerations or directional changes, constitute only a small subset relative to the overall temporal sampling space. This observation aligns with early discussions on spatiotemporal interest points (STIP) [26]. Building on these insights, we combined the innovations of STIP with Gaussian splatting, proposing MGR, enabling efficient and expressive temporal encoding.

Input Construction. For each joint i, its temporal trajectory is defined as a spatiotemporal point set $\mathcal{X}_i = \{\mathbf{x}_{i,t}\}_{t=1}^{T}$, where $\mathbf{x}_{i,t} = (x_{i,t}, y_{i,t}, t) \in \mathbb{R}^3$ contains 2D coordinates and normalized timestamps. To balance spatial and tempo-

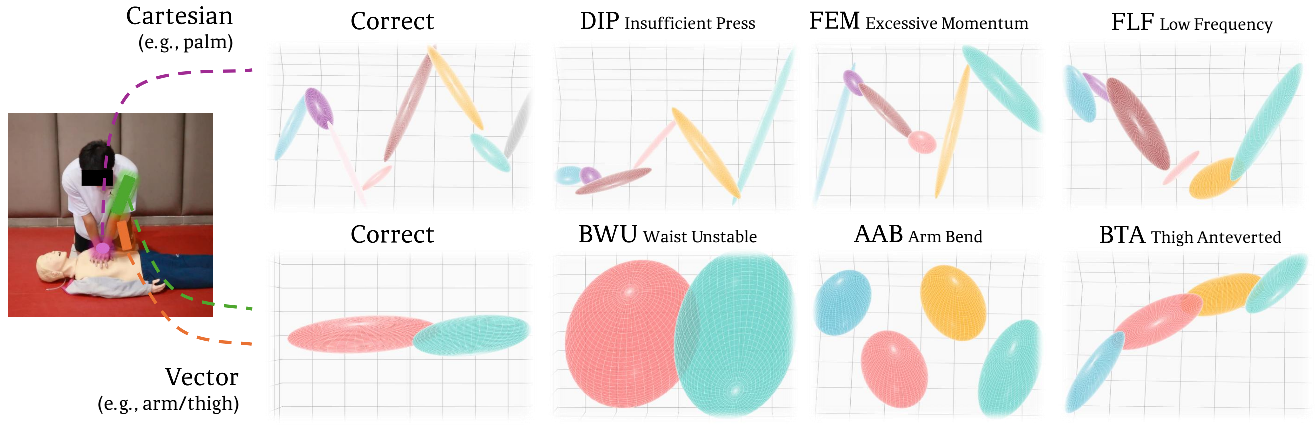

Figure 3. **Dynamics visualization of MGR.** Each ellipsoid represents a 3D Gaussian distribution (mean μ, covariance Σ) of a single joint/bone dynamics over time. From the perspective of stream-specific strengths, Joint Cartesian Stream excels at capturing global trajectory consistency (*e.g.*, palm trajectory); Bone vector stream better encodes kinematic transitions (*e.g.*, arm bend).

ral scales, we introduce a time-axis scaling factor α:

$$\mathbf{x}_{i,t} \leftarrow (x_{i,t}, y_{i,t}, \alpha \cdot t) \quad (\alpha \in \mathbb{R}^{+}). \tag{1}$$

Gaussian Modeling. We conjecture that $\mathcal{X}_i$ is generated by a mixture of K Gaussian distributions [37]. The probability density function is:

$$p(\mathbf{x}|\boldsymbol{\theta}_i) = \sum_{k=1}^{K} \pi_{i,k} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}), \tag{2}$$

where $\boldsymbol{\theta}_i = \{\pi_{i,k}, \boldsymbol{\mu}_{i,k}, \boldsymbol{\Sigma}_{i,k}\}_{k=1}^{K}$, with mixture weights $\pi_{i,k}$ satisfying $\sum_{k=1}^{K} \pi_{i,k} = 1$. Parameters are optimized using the Expectation Maximization (EM) algorithm (see supplementary material Sec. S4).

Action Token. For each Gaussian component k, we extract compressed features from $\boldsymbol{\mu}_{i,k}$ and $\boldsymbol{\Sigma}_{i,k}$. Each Gaussian represents an action token. Although we did not use higher dimensions in our architecture (discussed in Sec. 4.2), MGR can be extended to more dimensions, such as incorporating limb angles, states, etc. For the x-y-t or r-θ-t point set, we have a 3D Gaussian distribution:

$$\mathcal{N}(\mathbf{x}|\mu, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mu)}, \tag{3}$$

where, $\boldsymbol{\mu}$ represents the average position of the joints within this segment, while the covariance matrix $\boldsymbol{\Sigma}$ represents the scaling (motion) and rotation (direction) across different dimensions.

Covariance Decomposition. TTo achieve a representation that is both differentiable and interpretable, we transform the covariance matrix into scale and rotation components that have physical significance [23].

$$\boldsymbol{\Sigma}_{i,k} = \mathbf{R}_{i,k}\mathbf{S}_{i,k}\mathbf{S}_{i,k}^{\top}\mathbf{R}_{i,k}^{\top}, \tag{4}$$

where $\mathbf{S}_{i,k} = \text{diag}(s_{i,k}^{x}, s_{i,k}^{y}, s_{i,k}^{t}) \in \mathbb{R}^{3\times3}$ is derived from a 3D vector $\mathbf{s}_{i,k} \in \mathbb{R}^{3}$, $\mathbf{R}_{i,k} \in \mathbb{R}^{3\times3}$ is computed from a unit quaternion $\mathbf{q}_{i,k} \in \mathbb{R}^{4}$.

This decomposition disentangles the motion dynamics into scale (magnitude of movement) and orientation (direction of movement), aligning with human biomechanical constraints. Finally, we concatenate $\boldsymbol{\mu}_{i,k}$, $\mathbf{s}_{i,k}$, and $\mathbf{q}_{i,k}$ into a compact 10D tensor:

$$\mathbf{f}_{i,k} = [\boldsymbol{\mu}_{i,k};\ \mathbf{s}_{i,k};\ \mathbf{q}_{i,k}] \in \mathbb{R}^{3+3+4=10}. \tag{5}$$

4.2. Hybrid Spatial Encoding

Rationale. Human action recognition, especially in medical scenarios, is based on modeling anatomical structures from sparse keypoints. Research has shown that 2D point and line combinations provide a strong impression of the type of action [21]. Both absolute joint positions and relative bone kinematics encode complementary motion semantics [40]. The features have two streams of information including direct use of Cartesian coordinates as input (coordinate-based encoding, (x, y, t)); and angles or vectors as input (vector-based encoding (r, θ, t)). In terms of anatomy, if bones are projected into a 2D space, information such as angles can be expressed in polar coordinates. These features are critically important in the recognition of medical actions. See Fig. 4 for the sensitivity analysis of the two streams.

Let us discuss the limitations of prior encoding schemes:

- Joint Cartesian *w/* Bone Cartesian: Widely adopted in GCNs and variants using a dynamic graph or attention mechanism [49, 55, 56, 59]. This approach forces networks to learn positions, resulting in inefficient feature redundancy.
- Joint Polar *w/* Bone Vector: Despite their anatomical grounding, angular wraparound artifacts (*e.g.*, $-\pi \leftrightarrow \pi$

discontinuities) induce gradient instability, as evidenced by our polar-based model in ablation studies; see Tab. 2.

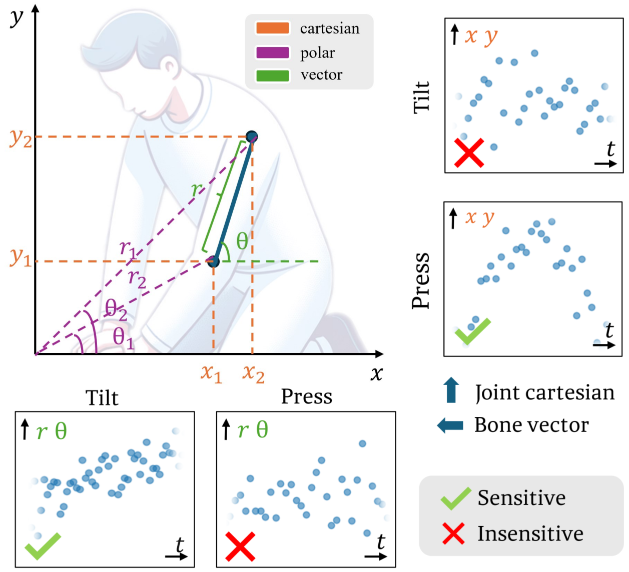

Figure 4. **Feature Discriminability of Hybrid Spatial Encoding.** The figure illustrates three information modes: cartesian-based, polar-based, and vector-based. Using chest compression and limb tilt as prototypes, the analysis reveals distinct signal-formative capabilities: Sensitive modes generate structured point clusters that fit to kinematic functions (*e.g.*, slopes for compression, waves for tilt), while insensitive modes exhibit noise distributions. The hybrid architecture orchestrates dual-stream processing to adaptively harness these geometric discriminators.

We propose the HSE solution that takes advantage of the joint Cartesian coordinates (x, y, t) and the bone vector (r, θ, t), and decouples absolute localization from relative kinematics, contextualizing fine-grained bone dynamics.

Semantics Disentangling. Joint stream (Cartesian encoding) is represented in the xyt-space to preserve absolute spatial-temporal positions. Bone stream (vector encoding) is parameterized by dynamic vector features in $r\theta t$-space.

$$\mathbf{J}_i = [x_i, y_i, t] \in \mathbb{R}^3, \mathbf{B}_{ij} = [\Delta r_{ij}, \theta_{ij}, t] \in \mathbb{R}^3, \qquad (6)$$

where $\Delta r_{ij} = \|\mathbf{J}_j - \mathbf{J}_i\|_2$ and $\theta_{ij} = \arctan\left(\frac{y_j - y_i}{x_j - x_i}\right)$. Concatenating raw joint and bone features (x, y, r, θ, t) as input to MGE may risk *multicollinearity*, as Cartesian coordinates and polar parameters are geometrically interdependent ($x = r\cos\theta, y = r\sin\theta$). Our MGE module first processes joints and bones in isolated embedding spaces, disentangling absolute and relative semantics. Subsequent fusion operates on decorrelated high-level features.

Feature Fusion Strategy. We adopt fusion with multiple variants: Concatenation (Eq. (7)), Cross-Attention (Eq. (8)), and others. See supplementary material Sec. S5 for details.

$$\mathbf{F} = [\mathbf{F}_J; \mathbf{F}_B] \in \mathbb{R}^{2M \times K \times 10}. \qquad (7)$$

$$\mathbf{F} = \text{LayerNorm}(\mathbf{F}_J + \text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V})). \qquad (8)$$

Different fusion strategies can work differently on different data complexities and dataset types.

4.3. Temporal Dynamics with MGR

As illustrated in Fig. 3, our temporal encoding module generates compact yet discriminative representations through MGR. The temporal evolution of actions (*e.g.*, chest compression) is encoded as compact sequences of Gaussian components that represent motion primitives. Remarkably, complex 60 frame motions can only be represented by ≈ 6 Gaussians, demonstrating MGR's ability to distill high-level motion primitives. Gaussian parameters (mean μ, covariance Σ) create separable clusters in feature space, enabling classifiers to achieve 92.1% accuracy with minimal fine-tuning (see Tab. 1).

This analysis confirms that Gaussian-based representation learning bridges the gap between raw pose dynamics and specific semantics, an advantage for medical applications that require both precision and interpretability.

4.4. Downstream Tasks

Loss Function. Taking into account the fine-grained nature of the medical model and the fact that some labels have few or uneven samples, we use MixUp [61] to comprehensively improve the effectiveness of the model in processing medical data. Through linear interpolation of inputs and labels, the model can be generalized by generating synthetic samples. For a given pair of samples (x_i, y_i) and (x_j, y_j), the mixed input and label are computed as:

$$\tilde{x} = \lambda x_i + (1 - \lambda) x_j, \quad \tilde{y} = \lambda y_i + (1 - \lambda) y_j, \qquad (9)$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ for $\alpha \in (0, +\infty)$ is sampled from a Beta distribution, and $\alpha > 0$ controls the interpolation strength. The MixUp loss is defined as:

$$\mathcal{L}_{\text{MixUp}} = \lambda \mathcal{L}(f(\tilde{x}), y_i) + (1 - \lambda) \mathcal{L}(f(\tilde{x}), y_j). \qquad (10)$$

By combining pairs of inputs and labels, smoother decision boundaries and reduced overfitting are encouraged, which helps to enhance the robustness of the model. For other loss strategies such as CE and Label Smoothing [32], see supplementary material Sec. S6.

Classifier. The fused features $\mathbf{F}_{\text{fused}}$ are mapped to category scores through multilayer CNN with spatiotemporal pooling and network outputs $\mathbf{y}_{\text{pred}}$. The architecture is shown in the supplementary material Sec. S7.

Evaluation Report Generation. For generating a report of the evaluation, MGR-encoded skeletal tensors undergo spatiotemporal tokenization to bridge visual and textual modalities. Specifically, we first discretize continuous kinematic

Model	Modality	Backbone	Dual-stream	Pre-trained	**Accuracy**			**GFLOPs**
					Top-1	Top-5	Mean	Per sample
TSM [29]	RGB	ResNet-50	✗	✗	0.8665	0.9820	0.8469	131.83
TSN [47]	RGB	ResNet-50	✗	Kinetics-400	0.9009	0.9867	0.8899	251.12
TPN [53]	RGB	ResNet-50	✗	✗	0.8111	0.9727	0.7235	168.00
TIN [39]	RGB	ResNet-50	✗	✗	0.8306	0.9766	0.7951	131.83
C3D [45]	RGB	3D ConvNet	✗	Sports-1M	0.9165	0.9906	0.9069	308.92
I3D [4]	RGB	ResNet-50	✗	✗	0.8462	0.9789	0.8331	266.80
SlowFast [16]	RGB	ResNet-50	✓	Kinetics-400	0.9144	0.9874	0.9107	97.27
TimeSformer [1]	RGB	ViT	✗	✗	0.8283	0.9781	0.8043	360.89
ST-GCN [59]	Skeleton	GCN	✗	✗	0.8618	0.9758	0.8356	43.76
2s-aGCN [40]	Skeleton	GCN	✓	✗	0.8907	0.9703	0.8812	50.01
STGCNPP [14]	Skeleton	GCN	✗	✗	0.8938	0.9742	0.8777	21.69
PoseC3D [15]	Skeleton	3D ConvNet	✗	✗	0.8798	0.9727	0.8582	24.51
Ours *(MGR w/ HSE)*	Skeleton	CNN	✓	✗	0.9212	0.9836	0.9082	4.45
Ours *(MGR only)*	Skeleton	CNN	✗	✗	0.8954	0.9602	0.8836	2.23

Table 1. **Model Performance Across Modalities.** Blue indicates the best result and Orange indicates the second-best result.

features (depth, frequency, posture angles) into clinically-grounded linguistic descriptors ("insufficient 4cm compression","optimal 110bpm rhythm") using quantization bins derived from AHA guidelines, then made final inferences (metrics→manifestation→consequence) through a logical chain reasoning mechanism generated based on dataset analysis. Effectiveness score calculation and causal analysis is shown in the supplementary material Sec. S8.

5. Experiments

All experiments were implemented in PyTorch and use the MMPose framework [10]. To generate a pose input for GAUSSMEDACT from RGB modality, we adopted RTMpose [20]. The image sequences are uniformly sampled to 32-frame clips with a spatial resolution of 224×224. To ensure fairness, all models shared identical training-test splits (80%-20% random partition) and underwent rigorous adjustments for confounding factors, as detailed in Tab. 1. We trained models using early stopping with a maximum of 300 epochs. Memory-intensive models were deployed on NVIDIA A100 GPUs, while other models utilized NVIDIA A6000 GPUs.

5.1. Comparative Analysis

We conduct comprehensive evaluations of multimodal approaches on the CPREVAL-6K dataset, comparing our skeleton-based GAUSSMEDACT with state-of-the-art methods that use RGB and skeleton modalities. Four metrics are adopted: Top-1/5 accuracy, class-wise mean accuracy, and computational complexity (GFLOPs). The top-1 accuracy receives the primary emphasis due to its clinical relevance as it critically affects CPR effectiveness. There may be multiple relatively reasonable answers (*e.g.*, when both primary and secondary labels exist, and the model identifies the secondary category). In such cases, using the Top-5 accuracy metric would provide a fairer evaluation of the model's performance. Class-wise mean accuracy ensures balanced performance across CPR errors, particularly for low-frequency but high-risk categories (*e.g.*, depth-excess). As shown in Tab. 1 and Fig. 5, three key observations emerge:

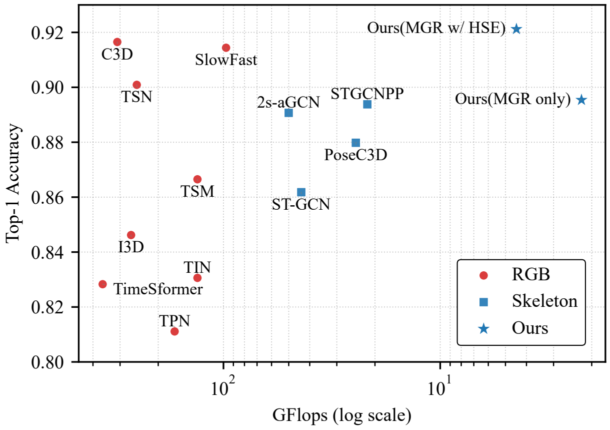

Figure 5. **Efficiency-Accuracy Trade-off Comparison.** Each point represents a model, with coordinates indicating computational complexity (GFLOPs) in log scale and top-1 accuracy.

Modality. Skeleton-based models demonstrate significantly lower computational requirements (6.13× fewer GFLOPs on average) compared to RGB counterparts. Even when accounting for pose extraction costs (4.03-5.45 GFLOPs for RTMpose), the total computation remains substantially lower than RGB methods. However, RGB approaches achieve higher accuracy ceilings (91.65% vs. 89.38% for prior skeleton methods), indicating that the RGB modality has specific advantages in feature richness.
Superiority of GAUSSMEDACT. Our method establishes new state-of-the-art performance with 92.12% Top-1 accu-

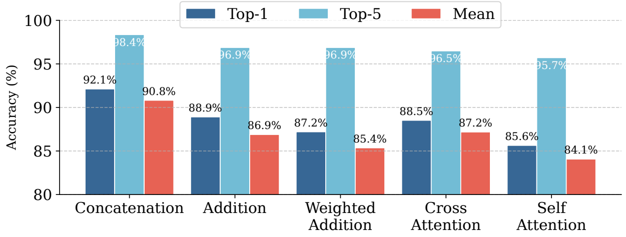

Figure 6. **Performance Comparison of Feature Fusion Strategies.** Simpler strategy achieved the highest accuracy.

racy, surpassing all existing models, including RGB-based approaches. In particular, this is achieved with only 4.45 GFLOPs, which is 10% of the computational cost required by ST-GCN (43.76 GFLOPs). The accuracy improvement over RGB models confirms the underexploited potential of skeleton data in medical action recognition when coupled with effective representation learning.

Efficiency-Performance Trade-off. The *MGR-only* variant (89.54% Top-1) already outperforms all skeleton baselines while requiring only half the computational complexity of full model, validating the efficacy of MGR. The complete model further improves performance by 2.58%, demonstrating synergistic effects between MGR and HSE.

5.2. Component Analysis

To validate the effectiveness of MGR and HSE, we systematically evaluated five configuration variants with and without MGR under identical experimental settings in Tab. 2.

Key observations reveal that: (1) The removal of MGR or HSE leads to significant performance degradation (2.58-5.31% Top-1 drop), confirming their complementary roles in motion representation. (2) The polar coordinate representation for the joints demonstrates limited learnability through the MGR-CNN pipelines, aligning with our theoretical analysis in Section 3.2. (3) Although the bone vector alone achieves suboptimal Top-1 accuracy (86.81%), its competitive Top-5 performance (96.10%>96.02%) suggests enhanced model confidence in candidate selection.

To evaluate the effectiveness of different feature fusion strategies in HSE, we conducted a comparison of five fusion methods. Our quantitative results (Fig. 6) reveal performance variations between the fusion strategies. More complex strategies lead to worse results, which may indicate that the low semantic complexity of tokens processed by MGR, so advanced fusion mechanisms are unnecessary.

5.3. Cross-Dataset Evaluation

We conduct rigorous cross-dataset validation using the recently released medical CPR-Coach benchmark [48], which contains 14 classes. Following the official (60/40) split protocol, we compare GAUSSMEDACT with four approaches in different modality. Training configurations strictly follow original papers for all methods. All models use the best results reported under different settings in the original paper. As shown in Tab. 3, our method achieves 2.78% absolute improvement in Top-1 accuracy while maintaining real-time performance. The superior performance across datasets demonstrates the robustness of MGR.

MGR	Info Space		Accuracy			
	Joint	Bone	Top-1	Δ	Top-5	Mean
✓	Carte.	Vector	0.9212	-	0.9836	0.9082
✗	Carte.	Vector	0.8197	↓ 10.2	0.8915	0.7988
✓	Carte.	/	0.8954	↓ 2.6	0.9602	0.8836
✓	/	Vector	0.8681	↓ 5.3	**0.9610**	0.8524
✓	Polar	Vector	0.8813	↓ 4.0	0.9516	0.8652

Table 2. **Ablation Study on Different Information Mode and Space.** *Carte.* represents cartesian. *Bone-Vector Only* showed a high top-5 accuracy **(Bold)** demonstrating bone feature enhance the ability to perceive the feature ambiguity and identify the potential correct range. Polar-space is harder for neural networks to learn. *Joint-Cartesian w/ Bone-Vector* showed the best result.

Model	Modality	Epoch	**Accuracy**	
			Top-1	Top-3
TSN-pretrained	RGB	50	0.9067	0.9921
TSN	Flow	50	0.8304	0.9851
STGCN-best	Skeleton	50	0.9246	0.9970
PoseC3D	Skeleton	240	0.9208	0.9922
Ours	Skeleton	100	0.9524	0.9950

Table 3. **Cross-dataset evaluation on *Coach*.** The dataset is a medical action dataset with 14 categories. Our model achieves 2.78% accuracy gain while maintaining real-time capability. Blue indicates the best result and Orange indicates the second-best.

6. Conclusions

This paper addresses critical dataset and methodology gaps in medical action evaluation. We introduce CPREVAL-6K, a much-needed fine-grained medical dataset that incorporates expert-validated hierarchical annotations which capture subtle error patterns. Through comprehensive comparative studies, we demonstrate that our proposed framework using spatiotemporal Gaussian mixture representation in decoupled joint and bone spaces outperforms both RGB- and skeleton- based models in accuracy while achieving significant computational cost reduction. The framework also exhibits robustness to cross-dataset generalization. These dual contributions establish a new foundation for the real-time evaluation of CPR, with potential applications that extend to other important medical and clinical assessments.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 1, 2, 7

[2] Bentley J Bobrow, Tyler F Vadeboncoeur, Uwe Stolz, Annemarie E Silver, John M Tobin, Scott A Crawford, Terence K Mason, Jerome Schirmer, Gary A Smith, and Daniel W Spaite. The influence of scenario-based training and real-time audiovisual feedback on out-of-hospital cardiopulmonary resuscitation quality and survival from out-of-hospital cardiac arrest. *Annals of emergency medicine*, 62 (1):47–56, 2013. 3, 2

[3] XB Bruce, Yan Liu, Xiang Zhang, Sheng-hua Zhong, and Keith CC Chan. Mmnet: A model-based multimodal network for human action recognition in rgb-d videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3522–3538, 2022. 2

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 7

[5] Nishanth Adithya Chandramouli, Sivaramakrishnan Natarajan, Amal H Alharbi, Subhash Kannan, Doaa Sami Khafaga, Sekar Kidambi Raju, Marwa M Eid, and El-Sayed M El-Kenawy. Enhanced human activity recognition in medical emergencies using a hybrid deep cnn and bi-directional lstm model with wearable sensors. *Scientific Reports*, 14(1): 30979, 2024. 2

[6] Guikun Chen and Wenguan Wang. A survey on 3d gaussian splatting. *arXiv preprint arXiv:2401.03890*, 2024. 1

[7] Yuxin Chen, Ziqi Zhang, Chunfeng Yuan, Bing Li, Ying Deng, and Weiming Hu. Channel-wise topology refinement graph convolution for skeleton-based action recognition. In *Proceedings of the ICCV*, pages 13359–13368, 2021. 2, 4

[8] Zhan Chen, Sicheng Li, Bing Yang, Qinghan Li, and Hong Liu. Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. In *AAAI*, pages 1113–1122, 2021.

[9] Hyung-gun Chi, Myoung Hoon Ha, Seunggeun Chi, Sang Wan Lee, Qixing Huang, and Karthik Ramani. Infogcn: Representation learning for human skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20186–20196, 2022. 2

[10] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. `https://github.com/open-mmlab/mmpose`, 2020. 7

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2022. 2

[12] Clara Daudre-Vignier, Declan G Bates, Timothy E Scott, Jonathan G Hardman, and Marianna Laviola. Evaluating current guidelines for cardiopulmonary resuscitation using an integrated computational model of the cardiopulmonary system. *Resuscitation*, 186:109758, 2023. 1

[13] Robert DiPietro, Colin Lea, Anand Malpani, Narges Ahmidi, S Swaroop Vedula, Gyusung I Lee, Mija R Lee, and Gregory D Hager. Recognizing surgical activities with recurrent neural networks. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part I 19*, pages 551–558. Springer, 2016. 2

[14] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7351–7354, 2022. 2, 7

[15] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2969–2978, 2022. 2, 7

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the CVPR*, pages 6202–6211, 2019. 2, 7

[17] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, page 3, 2014. 2

[18] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2

[19] Julia Gruber, Dominik Stumpf, Bernhard Zapletal, Stephanie Neuhold, and Henrik Fischer. Real-time feedback systems in cpr. *Trends in Anaesthesia and Critical Care*, 2(6):287–294, 2012. 1

[20] Tao Jiang, Peng Lu, Li Zhang, Ningsheng Ma, Rui Han, Chengqi Lyu, Yining Li, and Kai Chen. Rtmpose: Real-time multi-person pose estimation based on mmpose. *arXiv preprint arXiv:2303.07399*, 2023. 7

[21] Gunnar Johansson. Visual perception of biological motion and a model for its analysis. *Perception & psychophysics*, 14:201–211, 1973. 2, 5

[22] Tamás Karácsony, László Attila Jeni, Fernando De la Torre, and João Paulo Silva Cunha. Deep learning methods for single camera based clinical in-bed movement action recognition. *Image and Vision Computing*, 143:104928, 2024. 2

[23] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1, 4, 5

[24] Seongji Ko, Yoongeol Lee, Mingi Choi, Daun Choi, Choung Ah Lee, and Jong-Uk Hou. Harnessing optical flow in deep learning framework for cardiopulmonary resuscitation training. *Expert Systems with Applications*, 238:121775, 2024. 2

[25] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2

[26] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64:107–123, 2005. 1, 4

[27] Ang Li, Meghana Thotakuri, David A Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *arXiv preprint arXiv:2005.00214*, 2020. 2

[28] Yongyuan Li, Mingjie Yin, Wenxiang Wu, Jiahuan Lu, Shangdong Liu, and Yimu Ji. A deep-learning-based cpr action standardization method. *Sensors (Basel, Switzerland)*, 24(15):4813, 2024. 2

[29] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7083–7093, 2019. 7

[30] Siobhán Masterson, Tatsuya Norii, Mio Yabuki, Takanari Ikeyama, Ziad Nehme, Janet Bray, et al. Real-time feedback for cpr quality–a scoping review. *Resuscitation Plus*, 19:100730, 2024. 1

[31] Shuangyan Miao, Yonghong Hou, Zhimin Gao, Mingliang Xu, and Wanqing Li. A central difference graph convolutional operator for skeleton-based action recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4893–4899, 2021. 2

[32] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019. 6

[33] Boulos S Nassar and Richard Kerber. Improving cpr performance. *Chest*, 152(5):1061–1069, 2017. 1

[34] Shalini Pandurangan, Michela Papandrea, and Mirko Gelsomini. Fine-grained human activity recognition-a new paradigm. In *Proceedings of the 7th International Workshop on Sensor-based Activity Recognition and Artificial Intelligence*, pages 1–8, 2022. 2

[35] Preksha Pareek and Ankit Thakkar. A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. *Artificial Intelligence Review*, 54 (3):2259–2322, 2021. 2

[36] Kaustubha D Patil, Henry R Halperin, and Lance B Becker. Cardiac arrest: resuscitation and reperfusion. *Circulation Research*, 116(12):2041–2049, 2015. 1

[37] Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663):3, 2009. 1, 5

[38] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2

[39] Hao Shao, Shengju Qian, and Yu Liu. Temporal interlacing network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11966–11973, 2020. 7

[40] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12026–12035, 2019. 2, 5, 7

[41] Koichiro Shinozaki, Hiroshi Nonogi, Ken Nagao, and Lance B Becker. Strategies to improve cardiac arrest survival: a time to act. *Acute Medicine & Surgery*, 3(2):61, 2016. 1

[42] Yi-Fan Song, Zhang Zhang, Caifeng Shan, and Liang Wang. Constructing stronger and faster baselines for skeleton-based action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 45(2):1474–1488, 2022. 2

[43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[44] Ian G Stiell, Siobhan P Brown, James Christenson, Sheldon Cheskes, Graham Nichol, Judy Powell, Blair Bigham, Laurie J Morrison, Jonathan Larsen, Erik Hess, et al. What is the role of chest compression depth during out-of-hospital cardiac arrest resuscitation? *Critical Care Medicine*, 40(4): 1192–1198, 2012. 1

[45] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the CVPR*, pages 4489–4497, 2015. 2, 7

[46] Andrew H Travers, Thomas D Rea, Bentley J Bobrow, Dana P Edelson, Robert A Berg, Michael R Sayre, Marc D Berg, Leon Chameides, Robert E O'Connor, and Robert A Swor. Part 4: Cpr overview: 2010 american heart association guidelines for cardiopulmonary resuscitation and emergency cardiovascular care. *Circulation*, 122(18_suppl_3): S676–S684, 2010. 1

[47] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2740–2755, 2018. 7

[48] Shunli Wang, Shuaibing Wang, Dingkang Yang, Mingcheng Li, Haopeng Kuang, Xiao Zhao, Liuzhen Su, Peng Zhai, and Lihua Zhang. Cpr-coach: Recognizing composite error actions based on single-class training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18782–18792, 2024. 2, 8

[49] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 5

[50] Linhua Xiang and Zengfu Wang. Joint mixing data augmentation for skeleton-based action recognition. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024. 2

[51] Leiyu Xie, Yuxing Yang, Zeyu Fu, and Syed Mohsen Naqvi. One-shot medical action recognition with a cross-attention mechanism and dynamic time warping. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 2

[52] Wentian Xin, Qiguang Miao, Yi Liu, Ruyi Liu, Chi-Man Pun, and Cheng Shi. Skeleton mixformer: Multivariate topology representation for skeleton-based action recogni-

tion. In *Proceedings of the 31st ACMMM*, pages 2211–2220, 2023. 2

[53] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020. 7

[54] Hao Yang, Dan Yan, Li Zhang, Yunda Sun, Dong Li, and Stephen J Maybank. Feedback graph convolutional network for skeleton-based action recognition. *IEEE Transactions on Image Processing*, 31:164–175, 2021. 2

[55] Fanfan Ye, Shiliang Pu, Qiaoyong Zhong, Chao Li, Di Xie, and Huiming Tang. Dynamic gcn: Context-enriched topology learning for skeleton-based action recognition. In *Proceedings of the 28th ACMMM*, pages 55–63, 2020. 5

[56] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 649–665. Springer, 2020. 5

[57] Joyce Yeung, Reylon Meeks, Dana Edelson, Fang Gao, Jasmeet Soar, and Gavin D Perkins. The use of cpr feedback-/prompt devices during training and cpr performance: a systematic review. *Resuscitation*, 80(7):743–751, 2009. 1

[58] Jun Yin, Jun Han, Chenghao Wang, Bingyi Zhang, and Xiaoyang Zeng. A skeleton-based action recognition system for medical condition detection. In *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2019. 2

[59] Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018. 1, 2, 4, 5, 7

[60] Hongzhou Yu, Tianhao Cheng, Ying Cheng, and Rui Feng. Finemedlm-o1: Enhancing the medical reasoning ability of llm from supervised fine-tuning to test-time training. *arXiv preprint arXiv:2501.09213*, 2025. 2

[61] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 6

[62] Qiang Zhang and Baoxin Li. Video-based motion expertise analysis in simulation-based surgical training using hierarchical dirichlet process hidden markov model. In *Proceedings of the 2011 international ACM workshop on Medical multimedia analysis and retrieval*, pages 19–24, 2011. 2

[63] Weiyu Zhang, Menglong Zhu, and Konstantinos G Derpanis. From actemes to action: A strongly-supervised representation for detailed action understanding. In *Proceedings of the IEEE international conference on computer vision*, pages 2248–2255, 2013. 2

[64] Lanfei Zhao, Zixiang Lin, Ruiyang Sun, and Aili Wang. A review of state-of-the-art methodologies and applications in action recognition. *Electronics*, 13(23):4733, 2024. 2