Physics-Informed Deep Learning for Human Action Recognition: A Biomechanical Approach

Zineb Haimer
Advanced Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco
zineb.haimer@uit.ac.ma (corresponding author)

Khalid Mateur
Advanced Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco
khalidmateur@gmail.com

Youssef Farhan
Advanced Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco
youssef.farhan@uit.ac.ma

Abdessalam Ait Madi
Advanced Systems Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco
abdessalam.aitmadi@uit.ac.ma

Received: 10 December 2025 | Revised: 20 January 2026 | Accepted: 28 January 2026

 DOI: https://doi.org/10.48084/etasr.16856

ABSTRACT

Human action recognition systems traditionally rely on learning statistical patterns from visual data without explicit modeling of the physical laws governing human motion. This paper presents a physics-informed neural network architecture that integrates biomechanical modeling directly into the learning process. This approach computes kinematic features (joint angles) and kinetic features (torque, energy) from estimated poses and fuses them with visual motion features within a Transformer encoder. A multi-objective loss function encourages physically plausible representations by penalizing biomechanically infeasible poses and energetically unrealistic movements. Testing the proposed method in police traffic gesture recognition achieved 96.11% classification accuracy while maintaining biomechanical feasibility (0.998 average feasibility score). The integration of physics-based features enables the disambiguation of visually similar gestures through their underlying physical signatures. This approach produces interpretable physical measurements that can be validated against biomechanical principles, making it particularly suitable for safety-critical applications where model transparency is essential.

Keywords-physics-informed neural networks; action recognition; gesture recognition; biomechanics; transformer networks; computer vision

I. INTRODUCTION

Reliable interpretation of human actions represents a fundamental challenge in computer vision, with implications for autonomous systems, human-robot collaboration, and intelligent transportation. Although Deep Learning (DL) approaches have demonstrated impressive capabilities, they typically operate as statistical correlation engines without explicit knowledge of the physical principles governing human motion. Standard DL models learn correlations between visual patterns and action labels but lack mechanisms to explicitly represent physical distinctions. When faced with noisy sensor data, partial occlusions, or ambiguous visual cues, such models cannot distinguish plausible human motions from physically impossible artifacts. This limitation is particularly problematic

in safety-critical applications where prediction confidence must be grounded in interpretable, verifiable reasoning.

This study introduces a neural network architecture that integrates biomechanical modeling into action recognition. Rather than treating physical plausibility as a post-hoc validation criterion, it is embedded as a core learning component. The central hypothesis is that by endowing models with explicit representations of human biomechanics—the kinematic constraints and kinetic properties defining human movement—we can create systems that are more robust, interpretable, and reliable. The contributions of this study are:

1. Architectural Framework: An architecture that integrates a differentiable biomechanical physics engine within an end-to-end trainable network, computing kinematic (joint angles, angular velocities) and kinetic metrics (torques, energy expenditure) from estimated poses.
2. Feature Fusion Strategy: A novel fusion mechanism within a Transformer encoder that combines visual motion features with computed physical metrics, enabling learning from both appearance patterns and motion dynamics.
3. Physics-Informed Training: A multi-objective loss function supplements classification objectives with explicit penalties for biomechanically infeasible poses and energetically unrealistic motions.
4. Integrated System Design: A complete, self-contained pipeline that operates on raw video input, including a custom-trained pose estimator with differentiable coordinate extraction.
5. Empirical Validation: Comprehensive evaluation demonstrates that physics-informed learning yields models with high classification accuracy (96.11%) and verifiable physical coherence (0.998 feasibility score).

II. RELATED WORK

This work lies at the intersection of video-based action recognition, pose-driven gesture analysis, and physics-informed neural networks.

A. Spatio-Temporal Models for Video Action Recognition

The dominant paradigm applies deep learning directly to pixel data. A pioneering work [1] introduced Two-Stream networks that process spatial (RGB) and temporal (optical flow) information in parallel. This was extended to 3D convolutions in C3D [2] and I3D [3], learning unified spatio-temporal representations. Recent architectures such as SlowFast networks [4] and Vision Transformers [5, 6] have focused on efficiency and attention mechanisms. Similarly, a Dynamic Adaptation Convolutional Neural Network (DACNN) [7] enhanced the adaptability to different hand morphologies in gesture recognition. However, these methods operate as black-box pattern matchers, learning statistical correlations without explicit knowledge of human body structure or physical motion constraints, making them computationally expensive and difficult to interpret in safety-critical contexts.

B. Pose-Based Action Recognition

An alternative approach decouples action recognition into pose estimation followed by pose-sequence classification. Early methods applied RNNs and LSTMs [8] to keypoint sequences. More sophisticated approaches model the human skeleton as a graph structure, leading to Spatio-Temporal Graph Convolutional Networks (ST-GCN) [9]. PoseC3D [10] demonstrated that applying 3D convolutions to pose heatmaps achieves excellent efficiency. Building on this, in [11], a modified EfficientNetV2-B1 was employed to classify static hand gestures for Rock, Paper, Scissors. Although pose-based methods correctly focus on skeletal structure, they treat the skeleton as a purely geometric construct without modeling physical relationships, biomechanical constraints, or dynamic properties (mass, inertia, torque, energy) that govern keypoint motion.

C. Physics-Informed Machine Learning and Biomechanics

Physics-Informed Neural Networks (PINNs) [12] have emerged as a powerful paradigm in scientific computing, successfully solving partial differential equations by encoding physical laws as soft constraints in loss functions. In computer graphics and robotics, biomechanical models generate realistic human animations and control robotic systems [13]. Some pose estimation methods incorporate biomechanical constraints as post-processing corrections [14]. However, the application of physics-informed principles to end-to-end video action recognition remains largely unexplored. This work adapts the PIML paradigm to gesture recognition by formulating biomechanical principles as both features and soft constraints within an end-to-end learnable system, computing and integrating kinetic properties (torque, energy) as learnable features within the classification network.

D. This Work's Contribution

This work integrates biomechanical physics directly into action recognition learning, with Table I positioning it relative to existing methods. This approach uniquely provides explicit kinetic features such as torques and energy beyond geometric pose, end-to-end differentiable physics integration participating in gradient flow, and physics-informed loss encoding biomechanical constraints as soft regularization. Video methods [1-4, 6] achieve strong performance on large benchmarks but operate as black boxes without physical grounding. Pose methods [9, 10] model skeletal structure but treat it geometrically without dynamic forces governing motion. This is the first work to compute kinetic properties from video, fuse them with visual features in a trainable architecture, and employ physics-informed losses guiding learning toward biomechanically plausible solutions.

TABLE I. COMPARISON WITH RELATED APPROACHES

Method	Physics modeling	Feature types	End-to-end training	Temporal model	Key limitation
Two-Stream [1]	✗	RGB + Optical Flow	✓	2-Stream CNN	Purely visual patterns
C3D [2]	✗	3D Convolutions	✓	3D CNN	Black-box, no interpretability
I3D [3]	✗	Inflated 3D Conv	✓	3D CNN	Requires massive datasets
SlowFast [4]	✗	Multi-rate features	✓	Dual-pathway CNN	No physical constraints
Timesformer [6]	✗	Attention features	✓	Transformer	Ignores body structure
DACNN [7]	✗	Adaptive CNN	✓	CNN	Hand-specific, no dynamics
ST-GCN [9]	✗	Skeleton graph	✓	Graph CNN	Treats skeleton as abstract graph
PoseC3D [10]	✗	Pose heatmaps	✓	3D CNN	Geometric only, no kinetics
EfficientNet [11]	✗	CNN features	✓	CNN	Static gestures only
Biomech. Pose [14]	✓ (partial)	Kinematics	✗	Post-processing	Not trainable end-to-end
This method (GP-INN)	**✓ (full)**	**Kinematics + Kinetics**	**✓**	**Physics-Transformer**	**—**

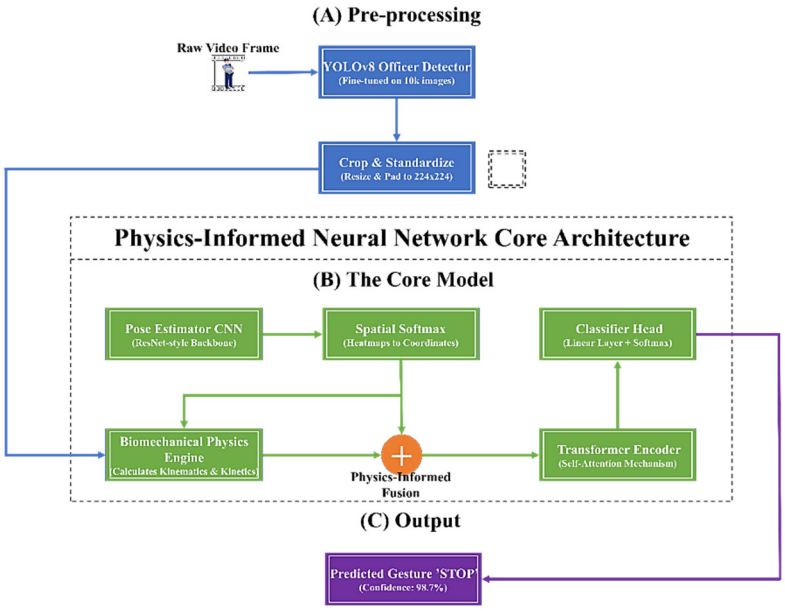

Fig. 1. The end-to-end architecture of the physics-informed neural network.

III. METHODOLOGY

Figure 1 illustrates the end-to-end architecture of the proposed physics-informed neural network. The system processes raw video input, performs officer localization and cropping, and feeds standardized frames into the network. The model internally estimates pose via a custom-trained pose estimator, extracts keypoint coordinates through a differentiable Spatial Softmax layer, calculates biomechanical physical metrics (kinematic and kinetic features), which are then fused with visual motion representations in a Transformer encoder, and produces final gesture classifications. The system is designed to be fully differentiable, enabling end-to-end training with gradient flow from the final classification loss back through the physics engine to the initial convolutional layers of the pose estimator.

A. *Preprocessing: Officer Localization*

For gesture recognition, the officer's motion constitutes the signal while the surrounding environment (vehicles, buildings, pedestrians) constitutes noise. An initial localization step is employed to ensure focus on relevant visual information and build robustness to background variations. This study employs YOLOv8 [15], a real-time object detection architecture known for speed and accuracy, fine-tuning it on a custom dataset of 9,601 police officer images collected from diverse global sources spanning various environmental conditions to ensure generalization. Each frame is processed through this detector to obtain a bounding box, which is then cropped and standardized to 224×224 pixels before feeding it to the pose estimator.

B. *Biomechanical Physics Engine*

The core innovation is a differentiable physics engine implemented in PyTorch that translates visual pose data into quantifiable physical metrics. This engine comprises two modules that model the kinematics and kinetics of human arm motion.

1) *Kinematics Module: Geometry and Constraints*

This study employs a 2-link planar arm model (shoulder-elbow-wrist), a standard biomechanics simplification appropriate for 2D video analysis. Given 2D coordinates of the left shoulder (P_S), elbow (P_S), and wrist (P_w), arm segment vectors are defined as:

$$\vec{v}_{upper} = P_e - P_S \; ; \; \vec{v}_{forearm} = P_w - P_e \quad (1)$$

The angle θ between vectors $\vec{a}$ and $\vec{b}$ is computed using the numerically stable dot product formula:

$$\theta = \arccos\left(\frac{\vec{a}\,.\vec{b}}{\|\vec{a}\|_2 \, \|\vec{b}\|_2}\right) \quad (2)$$

Two critical angles are computed: (1) Shoulder Flexion (θ_S), the angle between the upper arm vector and the downward vertical reference, representing arm elevation; (2) Elbow Flexion (θ_e), the interior angle between the upper arm and the forearm.

Based on established anthropometric data, valid ranges of motion are defined: Shoulder Flexion $\in [0°, 180°]$ and Elbow Flexion $\in [0°, 145°]$. A frame is feasible $F(t) = 1$ if all joint angles are within valid ranges. The overall feasibility score for a sequence of length T is:

$$Score_{feas} = \frac{1}{T}\sum_{t=1}^{T} F(t) \quad (3)$$

This quantifies how well estimated poses respect human anatomical constraints.

2) Kinetics Module (Motion Physics): Dynamics, Forces, and Energy

This module models forces and energy required to produce observed motion. Angular velocity and acceleration are computed from joint angle sequences using finite differences, where $\Delta t = 1/30s$.

$$\omega(t) = \frac{\theta(t)-\theta(t-1)}{\Delta t}; \; \alpha(t) = \frac{\omega(t)-\omega(t-1)}{\Delta t} \quad (4)$$

Simplified 2-link arm inverse dynamics is implemented to estimate joint torques (τ)—rotational forces required to create motion. Based on Newton's Second Law for rotation ($\tau = I\alpha$, where I is the moment of inertia), the shoulder torque is:

$$\tau_S(t) = \underbrace{(I_1 + I_2)\alpha_s(t)}_{Inertial\ Torque} + \underbrace{(m_1 I_{c1} + m_2 I_1)\, g sin\big(\theta_s(t)\big)}_{Gravitational\ Torque} \quad (5)$$

where m_i, I_i, I_{ci}, and I_i are mass, length, center of mass position, and moment of inertia of segment i (1: upper arm, 2: forearm), and g is gravitational acceleration. This equation shows that the required torque depends not only on the acceleration but also on the arm's position relative to gravity.

The total metabolic energy (E) for a gesture is approximated by integrating absolute mechanical power ($P = |\tau\omega|$):

$$E = \int_0^T |P(t)|dt \approx \sum_{t=1}^{T}(|\tau_s(t)\omega_s(t)| + |\tau_e(t)\omega_e(t)|)\Delta(t) \quad (6)$$

This metric identifies unrealistic, jerky motions that exhibit unphysically high energy costs. Smooth, natural human motions are energetically efficient.

C. Network Architecture

This architecture is a fully end-to-end system comprising four stages, each serving a specific role in the physics-informed learning pipeline.

1) Stage 1: Custom Pose Estimator

To maintain self-containment and avoid dependence on external tools, a pose estimator is implemented and trained specifically on the data using a ResNet-style CNN backbone. Input 224×224 images pass through convolutional and residual blocks, progressively downsampling while increasing feature depth. A deconvolutional head upsamples output to produce 56×56 heatmaps for 33 body keypoints (COCO format). The network is trained with heatmap regression loss, comparing predictions to ground-truth Gaussian distributions centered at keypoint locations.

2) Stage 2: Differentiable Coordinate Extraction

A simple argmax operation to extract peak coordinates from heatmaps is non-differentiable, breaking gradient flow. A Spatial Softmax layer computes a weighted average of a predefined coordinate grid, where weights are softmax-normalized heatmap values. This operation is smooth and fully differentiable, creating an unbroken gradient highway from the final loss back to the first convolutional layer.

3) Stage 3: Physics-Informed Transformer Encoder

To disambiguate temporally similar gestures (e.g., stop vs. slow vs. neutral), a powerful temporal model is needed to capture long-range dependencies. A Transformer Encoder is employed, selected for its proven ability to model temporal relationships through self-attention. The key innovation lies in input construction. The sequence of 33 keypoints per frame (66-dimensional vector) is linearly projected to the embedding dimension $d_{model} = 256$. Scalar physics metrics (feasibility score, log-energy) are projected to the same dimension. The physics embedding is broadcast and added to keypoint embeddings at every timestep, injecting global physical context into each frame's representation:

$$h_t = ProjectKeypoints(keypoints_t) + ProjectPhysics\big(physics_{global}\big) \quad (7)$$

Sinusoidal positional encodings maintain temporal order. The fused sequence passes through a multi-layer Transformer Encoder (4 layers, 8 attention heads, 1024 feed-forward dimension). Self-attention allows the model to learn which frames are most discriminative. This fusion enables contextualizing each pose not only by geometric configuration but also by global physical characteristics of the entire motion. A stationary arm in "stop" is interpreted differently from the same pose in "neutral" because the energy and feasibility contexts differ.

4) Stage 4: Classification Head

The element-wise mean of the Transformer output sequence is computed along the time dimension, producing a fixed-size feature vector representing the entire gesture. This vector passes through a final linear layer producing logits for 6 gesture classes. Mean pooling ensures that the model distributes

information across the entire sequence rather than bottlenecking at a single point.

D. Physics-Informed Training Strategy

Training a complex multi-objective model with a randomly initialized pose estimator poses challenges. Initial chaotic keypoint predictions lead to extreme physics penalties that can destabilize learning.

1) Two-Stage Curriculum Learning

A curriculum strategy is employed to manage training complexity:

- Stage 1 (Warm-up, Epochs 1-10): Train using only Cross-Entropy classification loss. All physics loss weights are zero. This forces the pose estimator and Transformer to learn basic visual patterns, leading to stable keypoint estimations.
- Stage 2 (Physics Fine-tuning, Epochs 11-50): Activate full physics-informed loss. Now that keypoint estimations are reasonable, physics penalties no longer overwhelm the system. Instead, they act as a powerful regularizer, refining understanding and pushing toward solutions that are accurate and physically coherent.

2) Multi-Objective Physics-Informed Loss

The final Loss is a weighted sum of four components:

$$\mathcal{L}_{total} = \omega_c \mathcal{L}_c + \omega_f \mathcal{L}_f + \omega_v \mathcal{L}_v + \omega_e \mathcal{L}_e \quad (8)$$

where $\mathcal{L}_c$ is the standard cross-entropy loss for gesture classification, $\mathcal{L}_f = \mathbb{E}\left[1 - Score_{feas}\right]$ is the feasibility loss that penalizes low overall physical plausibility, $\mathcal{L}_v = (1/T)\sum V(t)$ is the violation loss that penalizes the mean ratio of frames that violate specific joint limits ($V(t) = 1$) if any joint angle violates range at time t, providing a more granular penalty signal, and $\mathcal{L}_e = \mathbb{E}\left[ReLU(E - E_{max}\right]$ is the energy loss (hinge loss), penalizing motions with energy expenditure (E) exceeding a reasonable threshold (E_{max}).

Loss weights were selected as $\omega_c = 1.0$, $\omega_f = 0.4$, $\omega_v = 0.2$, and $\omega_e = 0.1$. This multi-objective formulation implements physics-based regularization, incentivizing solutions that classify correctly while respecting fundamental constraints and dynamics of human motion, biasing learning toward the manifold of physically plausible solutions.

E. Technical Implementation Specifications

1) Pose Estimator Architecture

A custom lightweight ResNet-style CNN backbone was trained end-to-end as an integral component of the system. The architecture consists of an initial 7×7 convolutional layer (64 filters, stride 2) with batch normalization and ReLU, followed by four residual blocks with progressively increasing channels (64→128→256→512). A transpose convolutional decoder upsamples features to 56×56 resolution, producing 33 heatmaps for the COCO keypoint format. The pose estimator is trained implicitly through backpropagation from the gesture classification loss without explicit pose annotations, representing a weakly-supervised approach where the physics engine provides inductive bias toward anatomically plausible configurations.

2) Video Preprocessing Pipeline

Each video is uniformly sampled to extract exactly 30 frames using *np.linspace* for temporal consistency. Frames are cropped using YOLOv8n bounding boxes and resized to 224×224 pixels via bilinear interpolation. RGB values are normalized to [0, 1] by dividing by 255. During training, stochastic augmentation is employed: horizontal flipping (probability 0.5), brightness/contrast adjustment (±15%), and random rotation (±10°). These augmentations enhance model robustness to viewpoint variations and lighting conditions.

3) Biomechanical Model Parameters

The 2-link planar arm physics engine employs standard anthropometric constants from biomechanics literature (Table II). The energy threshold for hinge loss $E_{max} = 50$ arbitrary units was selected through preliminary validation experiments. Numerical stability measures include: epsilon addition (1×10^{-8}) in vector norm calculations to prevent division by zero, and clamping dot products to a strict range [-1.0, 1.0] before arccos operations to handle floating-point rounding errors. Angular derivatives use direct finite differences (*torch.diff*) with replicate boundary padding, while the PoseRefinementBlock's learnable 1D temporal convolution provides adaptive smoothing.

TABLE II. BIOMECHANICAL PARAMETERS

Parameter	Value	Source
Upper arm mass	2.8 kg	[16]
Forearm mass	1.6 kg	[16]
Upper arm length	0.36 m	[16]
Forearm length	0.27 m	[16]
Upper arm CoM	43.6% of length	[17]
Forearm CoM	43.0% of length	[17]
E_{max} threshold	50 AU	Empirically tuned

4) Transformer Encoder Specifications

PyTorch's *nn.TransformerEncoderLayer* is employed with the following configuration: embedding dimension $d_{model} = 256$, 8 attention heads (each operating on 32-dimensional subspace), feed-forward network dimension = 1024, 4 encoder layers, activation function = ReLU, normalization = Layer Normalization applied after attention and feed-forward blocks (post-LN), and dropout = 0.1 in both attention and feed-forward paths. Positional encodings use the standard sinusoidal formulation from [18]:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (9)$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \quad (10)$$

For temporal aggregation, mean pooling is applied across the sequence dimension: $final\_features = encoder\_output.mean(dim = 1)$, producing a 256-dimensional vector fed to the classification head.

5) Training Protocol

Optimization used AdamW [19] (learning rate 1×10^{-4}, weight decay 0.01, β=(0.9, 0.999)) with gradient norm clipping at 1.0. Physical batch size was 4 with gradient accumulation over 4 steps, yielding an effective batch size of 16. Training involved 50 epochs in total (10 warm-up with classification loss only, then 40 with full physics-informed loss). Validation was monitored every epoch, saving the best model based on validation accuracy. Physics loss weights ($\omega_c = 1.0$, $\omega_f = 0.4$, $\omega_v = 0.2$, $\omega_e = 0.1$) were selected through preliminary experiments. Training completed in approximately 8 hours; inference achieved 42 FPS, including all pipeline stages (detection, pose estimation, physics computation, classification), enabling real-time deployment.

IV. EXPERIMENTAL SETUP AND RESULTS

A. Datasets

This study utilized two custom-curated datasets, designed to address the distinct challenges of officer localization and gesture recognition.

1) Officer Localization Dataset

To train a robust and generalizable police officer detector, a custom dataset was created due to the absence of suitable benchmark datasets for this specific task. Although general person detection benchmarks such as MS COCO and CrowdHuman exist, they are designed for generic pedestrian detection and lack the specialized characteristics essential for identifying uniformed officers. Specifically, these datasets do not adequately represent: (i) diverse police uniforms across different countries and jurisdictions, (ii) professional gear such as safety vests, helmets, and duty belts, (iii) officer-specific poses and gestures used in traffic control and law enforcement operations, and (iv) the variety of environmental conditions under which officers operate. A preliminary survey of existing datasets revealed only small-scale collections (fewer than 600 images) focusing on police equipment rather than officer detection, making them unsuitable for training a robust detector. Consequently, benchmark comparisons are not feasible for this specialized application, necessitating the development of a task-specific dataset and evaluation framework. Images were systematically gathered from public online sources, primarily through Google Images and stock photography websites, ensuring broad representation across multiple dimensions. The collection prioritizes global diversity, featuring police officers from numerous countries, including the United States, the United Kingdom, China, and Morocco, capturing wide variations in uniform styles, colors, safety vest designs, helmets, and equipment configurations. This international scope ensures that the model generalizes beyond region-specific uniform patterns. The dataset encompasses diverse and challenging operational scenarios: officers performing traffic control gestures, crowd management, routine patrols, and emergency responses. The images were collected across varied environmental conditions, including bright daylight, nighttime illumination, fog, rain, and snow, to ensure robustness in real-world deployment. Scene complexity varies from single-officer scenarios to multi-officer formations, with officers captured from multiple viewpoints (front, side, and back angles) and occlusion levels. A significant portion features officers wearing high-visibility safety vests, which are standard in many traffic control situations worldwide. The raw collection of 1,829 images was meticulously annotated using the Roboflow platform to create precise bounding boxes around each officer instance. To enhance dataset size and diversity, a strategic augmentation pipeline was implemented using the Albumentations library, designing seven distinct transformation sequences that simulate real-world challenges such as varying lighting conditions, perspective changes, and image quality degradation. This process expanded the dataset to 9,601 images, partitioned into 70% training (6,721 images), 15% validation (1,440 images), and 15% testing (1,440 images) subsets. To facilitate reproducibility and enable future research in this domain, the complete annotated dataset is publicly available [20].

2) Gesture Recognition Dataset

For the gesture recognition task, an existing public dataset of Chinese police traffic gestures [21] was adapted. Although this dataset provides valuable source material, it presents several limitations for the classification objectives of this work. The original videos contained long, unsegmented sequences with multiple consecutive actions, background activities, and transitions between gestures, which makes them unsuitable for direct use in training a classification model. Furthermore, the source videos exhibited limited environmental variability, as they were primarily recorded in consistent daylight conditions from similar camera distances and angles.

Manual temporal segmentation and curation were performed to create a clean action-specific dataset suitable for robust gesture classification. Each source video was carefully analyzed, and segments were cropped to isolate distinct individual gesture performances with clear start and end points. This meticulous process yielded 300 base videos with balanced class distribution: 50 videos, each for six gesture classes (stop, pass, turn left, turn right, slow, and neutral) captured at 30 FPS. To enhance the generalizability and simulate diverse real-world viewing conditions not present in the original dataset, a comprehensive augmentation strategy was implemented using the Albumentations library. Seven distinct augmentation pipelines were designed, incorporating temporal variations (speed adjustments between 0.85× and 1.15×) and spatial transformations (rotations ±8°, scaling ±5%, perspective shifts, brightness and contrast adjustments). This expanded the dataset to 1,200 videos (200 per class). To ensure a rigorous evaluation without data leakage, the 300 source videos were split chronologically into training (70%, 210 videos), validation (15%, 45 videos), and testing (15%, 45 videos) subsets before augmentation. All augmented versions of each source video were assigned exclusively to the same split as their original, guaranteeing that test subjects remained unseen during training. This protocol resulted in a final distribution of 840 videos for training, 180 for validation, and 180 for testing.

B. Implementation Details

The complete pipeline was implemented in PyTorch 1.12.0 with CUDA 11.6 on an NVIDIA RTX 3080 GPU (10 GB VRAM). The YOLOv8n localization model was fine-tuned for 50 epochs using the SGD optimizer with default

hyperparameters on the 9,601-image static dataset. The main gesture recognition model was trained using the proposed two-stage curriculum for 50 total epochs (10 warm-up, 40 physics-informed) with an effective batch size of 16 (4 physical batches × 4 gradient accumulation steps). Table III presents the detailed configuration for the three key experiments.

C. Evaluation Metrics

1) For Officer Localization

The fine-tuned YOLOv8 detector was evaluated using standard object detection metrics [22]. The primary metric is mean Average Precision (mAP), calculated as the mean of Average Precision (AP) values across IoU thresholds from 0.5 to 0.95 with a step of 0.05. In addition, Precision, Recall, and IoU are reported at the standard 0.5 threshold.

2) For Gesture Recognition

Model performance is assessed using standard multi-class classification metrics derived from the confusion matrix. Overall Accuracy is reported, along with per-class Precision, Recall, and F1-score. Summary statistics include macro-averaged (unweighted mean) and weighted-averaged (by class support) metrics across all six gesture classes.

To assess biomechanical plausibility, two novel metrics were introduced:

- Feasibility Score: Proportion of frames exhibiting anatomically valid joint angles based on human kinematic constraints.
- Average Energy: Mean estimated energy expenditure per gesture sequence, computed using inverse dynamics modeling.

TABLE III. DETAILED HYPERPARAMETER AND ARCHITECTURE CONFIGURATIONS

Parameter group	Hyperparameter	Model A	Model B	Model C
Architecture	Pose refinement Block	Yes	Yes	Yes
	Temporal encoder type	Transformer	Bidirectional LSTM	Transformer
	LSTM specific			
	hidden_dim	N/A	256	N/A
	num_layers	N/A	2	N/A
	dropout	N/A	0.2	N/A
	Transformer specific			
	d_{model} (Embedding dimension)	256	N/A	256
	nhead (Attention heads)	8	N/A	8
	dim_feedforward	1024	N/A	1024
	num_encoder_layers	4	N/A	4
	dropout	0.1	N/A	0.1
	Physics fusion	No	Yes	Yes
Loss function	Primary loss	Cross-entropy	Cross-entropy	Cross-entropy
	Physics losses enabled	No	Yes	Yes
	ω_c	1.0	1.0	1.0
	ω_f	0.0	0.4	0.4
	ω_v	0.0	0.2	0.2
	ω_e	0.0	0.1	0.1

D. Results

1) Officer Localization Performance

A robust gesture recognition system requires a high-fidelity front-end to accurately isolate the subject of interest. To this end, a YOLOv8n object detection model was fine-tuned on the custom-curated dataset of 9,601 diverse static images. Table IV summarizes the performance of this specialized police officer detector, evaluated on a held-out test set of images.

TABLE IV. PERFORMANCE OF THE FINE-TUNED YOLOV8N OFFICER DETECTOR ON TEST SET

Metric	Value
Precision	0.983
Recall	0.935
mAP@0.5	0.981
mAP@0.5:0.95	0.863
F1-score	0.958
Inference speed	3.8 ms

The mAP of 98.1% confirms high accuracy, while the 3.8 ms inference time ensures real-time capability without creating a pipeline bottleneck.

2) Gesture Recognition Performance

The complete Physics-Informed Neural Network achieved the following performance on the unseen test set of 180 videos:

- Test Accuracy: 96.11%
- Average Biomechanical Feasibility: 0.998
- Average Energy Expenditure: 11.42 J

The high accuracy demonstrates an effective classification, while the near-perfect feasibility score (0.998) indicates that the predicted poses respect anatomical constraints. The low energy value suggests smooth, realistic motion representations. Table V presents a detailed per-class breakdown of the model's performance. The model achieves perfect or near-perfect performance on distinct gestures (turn left, turn right, slow, pass) while showing slightly lower but still strong performance on the most ambiguous classes (stop, neutral). Figure 2 presents confusion matrices that compare a baseline model (trained without physics) against the proposed full physics-informed model. The baseline exhibits significant confusion between visually similar classes, particularly misclassifying

neutral and slow gestures as "stop." The proposed physics-informed model dramatically reduces these confusions, with predictions concentrated along the diagonal. This visual evidence confirms that physics integration provides crucial disambiguating information.

TABLE V. CLASSIFICATION REPORT OF THE PHYSICS-INFORMED MODEL ON TEST SET

Class	Precision	Recall	F1-Score	Support
neutral	0.88	0.90	0.89	31
pass	1.00	0.97	0.98	30
slow	1.00	1.00	1.00	31
stop	0.87	0.87	0.87	23
turn_left	1.00	1.00	1.00	35
turn_right	1.00	1.00	1.00	30
Weighted Avg	0.96	0.96	0.96	180
Macro Avg	0.96	0.96	0.96	180

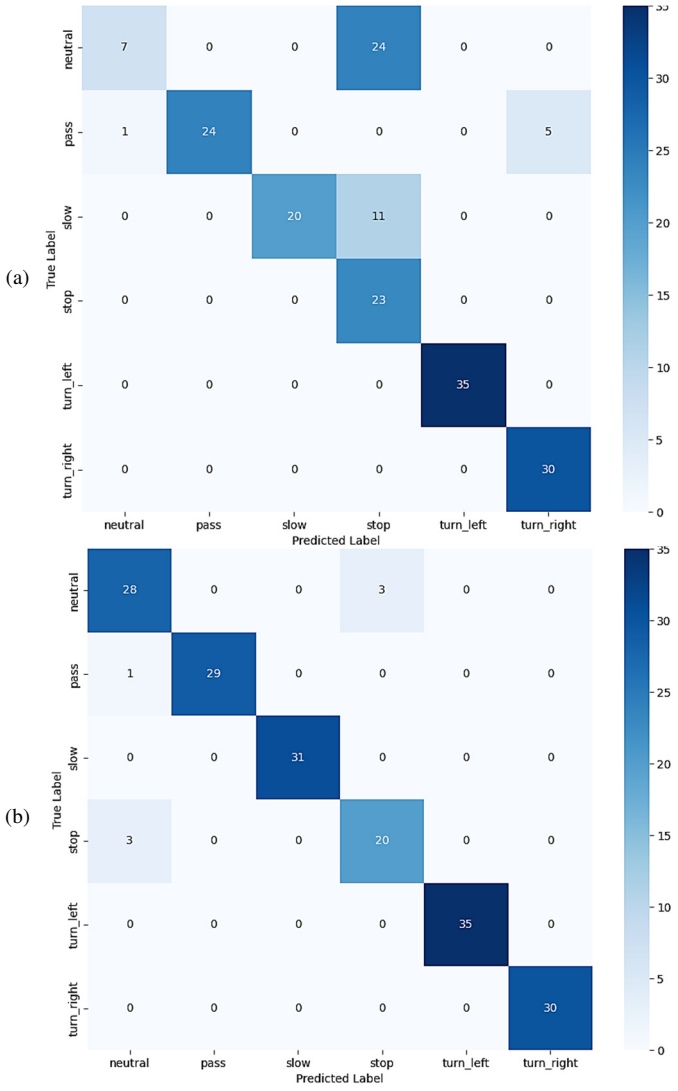

Fig. 2. Confusion matrices on the test set: (a) Baseline model (No Physics), (b) Full model.

In addition to quantitative metrics, qualitative results were also examined. Figure 3 shows sample predictions with overlaid physics metrics. The model correctly classifies diverse gestures while simultaneously computing interpretable physical measurements (joint angles, energy) that can be validated against biomechanical expectations in real time.

Fig. 3. Qualitative results of the physics-informed model in action: (a) "Turn_Left" gesture, (b) "Turn_Right" gesture, (c) "Neutral" gesture, (d) "Stop" gesture, (e) "Slow" gesture, (f) "Pass" gesture.

V. ABLATION STUDY

To validate that physics integration is the key driver of performance, a systematic ablation study was conducted, comparing three model configurations:

- Model A (Baseline): The Transformer-based architecture trained on keypoint sequences alone, with both the proposed physics feature fusion and physics loss terms disabled.
- Model B (Physics+LSTM): The full physics-informed method with LSTM for the temporal encoder instead of a Transformer.

- Model C (Physics+Transformer): The complete proposed architecture.

The results in Table VI form the core of the findings. The performance gap between Model A and Model B (+16.67%) demonstrates that physics-informed learning provides substantial benefits. The baseline struggles with temporal ambiguity—gestures that look similar frame-by-frame but differ in dynamics. By incorporating kinematic and kinetic features, Model B distinguishes gestures based on physical signatures (energy profiles, torque patterns) rather than visual appearance alone. Model C achieves an additional +2.22% improvement, suggesting that Transformer self-attention effectively leverages physics-informed representations. This possibly indicates that the Transformer's ability to compute global relationships across all timesteps complements the global physics metrics (total energy, sequence feasibility), enabling more sophisticated integration than LSTM's sequential processing.

TABLE VI. PERFORMANCE COMPARISON ACROSS MODEL CONFIGURATIONS

Model	Physics Fusion	Physics Loss	Temporal Encoder	Test Accuracy	Improvement (over Baseline)
A. Baseline	No	No	Transformer	77.22%	-
B. Physics+LSTM	Yes	Yes	LSTM	93.89%	+16.67%
C. Physics+Transformer	**Yes**	**Yes**	**Transformer**	**96.11%**	**+18.89%**

Beyond accuracy, the physics-informed models (B and C) produce representations with verifiable physical properties. Table VII shows that the baseline achieves reasonable anatomical plausibility (feasibility 0.9307), but the physics-informed models show measurable improvement, reaching near-perfect feasibility scores (>0.995). Interestingly, the baseline exhibits lower energy (8.97 J), which may indicate overly simplified or static pose sequences that lack the dynamic complexity of realistic gestures. The physics-informed models produce slightly higher but more realistic energy profiles (11.4-13.6 J), suggesting that they capture the full kinematic and kinetic richness of natural arm movements. This confirms that the physics-informed loss successfully guides the model toward representations that are both anatomically valid and dynamically realistic.

TABLE VII. PHYSICAL COHERENCE COMPARISON ACROSS MODELS

Model	Avg. Feasibility	Avg. Energy (J)
Baseline (A)	0.9307	8.97
Physics-LSTM (B)	0.9953	13.6
Physics-Transformer (C)	0.998	11.4

The most dramatic improvements occur for the ambiguous stop/neutral/slow classes, as shown in the results in Table VIII, which reveal the specific value of physics-informed learning. The baseline struggles most severely with "neutral" (F1=0.36), likely because a neutral stance involves minimal motion and lacks distinctive visual features. The physics-informed model achieves a remarkable +0.53 improvement by recognizing neutral's characteristic low energy signature and minimal joint movement. Similarly, "stop" improves by +0.30, as the model learns to identify the physical profile of rapid deceleration followed by sustained hold against gravity. Even "slow," which the baseline handles reasonably well (F1=0.78), reaches perfect performance (F1=1.00) when physics features enable recognition of its smooth, controlled motion pattern.

The Transformer baseline represents a strong architecture with self-attention mechanisms, yet its performance on ambiguous classes reflects the inherent difficulty of distinguishing temporally similar gestures using visual features alone. Although this study does not provide direct comparisons with ST-GCN or PoseC3D due to requiring substantial re-implementation and fair hyperparameter tuning, these methods fundamentally lack kinetic modeling, such as torques and energy, which form the core contribution of the proposed method. The results of this study demonstrate complementary value, particularly for limited training data scenarios with 840 videos versus 100K+ videos in large benchmarks, safety-critical applications requiring interpretable predictions, and tasks where dynamics differentiate classes more than appearance.

TABLE VIII. PER-CLASS IMPACT OF PHYSICS-INFORMED LEARNING

Class	Baseline F1	Physics-Transformer F1	Improvement
stop	0.57	0.87	+0.30
neutral	0.36	0.89	+0.53
slow	0.78	1.00	+0.22

VI. DISCUSSION

A. Physics-Informed Learning: Impact and Mechanisms

The experimental results demonstrate that explicit integration of biomechanical modeling provides substantial and measurable benefits for action recognition. The proposed physics-informed approach achieves an 18.89% accuracy improvement over a baseline Transformer model, with the most dramatic gains occurring precisely where expected: on temporally ambiguous gestures that are visually similar but physically distinct. The neutral gesture, which the baseline identifies with only a 0.36 F1-score, improves remarkably to 0.89 when physics features are integrated, a 0.53 gain that represents the difference between system failure and reliable performance. This is not coincidental. Neutral stance is characterized by minimal visual motion, making it nearly invisible to appearance-based features, yet it possesses a clear physical signature: low energy expenditure, stable joint configurations, and minimal torque. Similarly, the stop gesture improves by +0.30 (F1: 0.57→0.87) as the model learns to recognize its distinctive kinematic profile of rapid deceleration followed by sustained hold against gravity, requiring continuous muscular effort visible in torque calculations.

Beyond disambiguation, the physics-informed models produce representations with verifiable physical coherence. The near-perfect feasibility scores (0.9953 for LSTM, 0.998 for Transformer) indicate that predicted poses overwhelmingly respect anatomical constraints, compared to the baseline's 0.9307. More subtly, the energy profiles reveal that physics-informed models capture dynamic realism: while the baseline exhibits lower energy (8.97 J), suggesting overly static or simplified motion representations, the physics-informed models produce realistic energy values (11.4-13.6 J) that reflect the full kinematic and kinetic complexity of natural arm gestures. This confirms that the multi-objective physics-informed loss successfully regularizes the solution space, guiding learning toward the manifold of physically plausible motions. Furthermore, the ablation study reveals important architectural synergy. While physics integration with LSTM yields substantial improvement (+16.67%), the Transformer architecture achieves additional gains (+18.89%), suggesting that self-attention mechanisms effectively leverage global physical context. We hypothesize that the Transformer's ability to compute relationships across all timesteps complements our sequence-wide physics metrics (total energy, average feasibility), enabling more sophisticated integration than LSTM's sequential processing. Critically, this entire pipeline is end-to-end trainable through the proposed differentiable physics engine, avoiding the need for pre-trained pose estimators or post-hoc corrections and ensuring that all components optimize jointly toward the classification objective.

1) Methodological Contributions and Broader Applicability

Beyond the specific application domain, this work contributes a generalizable framework for integrating domain knowledge into data-driven action recognition systems. The proposed differentiable biomechanical engine, implemented entirely in PyTorch with standard tensor operations, demonstrates that complex physical calculations—joint angle through dot products, inverse dynamics modeling, and energy integration—can be made fully compatible with automatic differentiation. This provides a template that researchers can adapt to other domains in which physical principles govern observed phenomena. The engine computes kinematic features (angles, velocities, accelerations) and kinetic features (torques, energy) that capture both the geometry and dynamics of motion, going beyond the purely geometric representations used in prior pose-based methods like ST-GCN or PoseC3D.

The proposed feature fusion strategy offers a simple yet principled approach to combining global physical context with local spatiotemporal features. By projecting physics metrics into the same embedding space as visual features and broadcasting them across all timesteps, a sequence-wide physical context is injected into the representation of each individual frame. This differs from concatenation or late fusion approaches; the physics embedding modulates how the Transformer attends to and interprets each pose, effectively providing a "physical lens" through which visual information is processed. The multi-objective physics-informed loss formulation demonstrates how domain knowledge can be encoded as soft constraints that guide rather than rigidly constrain learning. By combining classification loss with granular penalties for anatomical violations and energetic implausibility, the training objective optimizes for accuracy and physical coherence. The two-stage curriculum—warm-up with classification loss alone, then physics fine-tuning—proves essential for stable convergence, preventing extreme initial physics penalties from destabilizing early learning.

A crucial advantage of physics-informed models extends beyond accuracy metrics to interpretability and safety. The proposed system produces not only class predictions but also quantifiable physical measurements: joint angles that can be validated against anthropometric literature, torques that reflect biomechanical effort, and energy values that indicate motion efficiency. These metrics provide transparency that is particularly valuable in safety-critical applications. For autonomous vehicles interpreting traffic officer gestures, a prediction accompanied by feasibility scores and energy profiles offers multiple validation signals. If a gesture is classified with high confidence but exhibits physically implausible joint angles or unrealistic energy, this discrepancy flags potential sensor malfunction, adversarial conditions, or model uncertainty—enabling fail-safe behaviors. Domain experts (biomechanists, safety engineers) can inspect these physical metrics to audit model behavior in ways that are impossible with black-box predictions. This interpretability makes physics-informed approaches particularly suitable for deployment scenarios where stakeholder trust and regulatory compliance demand explainable AI systems.

2) Limitations and Future Research Directions

This study acknowledges several limitations that contextualize these contributions and suggest directions for future work. The proposed biomechanical engine employs a 2D planar arm model, a reasonable simplification for frontal gesture analysis but inadequate for actions with significant depth variation or out-of-plane motion. Extension to full 3D biomechanics would require depth sensing (stereo cameras, LiDAR) or multi-view inputs, enabling modeling of complex 3D joint rotations and whole-body coordination. The proposed physics model captures essential dynamics through a 2-link arm with standard inverse dynamics, but more sophisticated biomechanical models could incorporate muscle activation patterns, multi-segment coordination, contact forces, or fatigue effects—potentially offering even richer physical signals for complex action recognition tasks. The computed kinetic features (torque and energy) are most informative for gestures that involve significant limb motion with clear physical signatures. For subtle actions like facial expressions or fine finger movements where forces are minimal, physics-based features may provide limited additional information beyond geometric pose features.

The evaluation focused primarily on a single domain—police traffic gestures—demonstrating the feasibility and benefits of physics-informed learning in a controlled setting. While this establishes proof of concept, broader validation across diverse action recognition benchmarks (e.g., NTU RGB+D for full-body actions, Kinetics for everyday activities, domain-specific datasets in sports or sign language) would strengthen claims about generalization and reveal which action

categories benefit most from physics integration. The physics computation adds approximately 12% inference overhead compared to pure pose-based classification; for ultra-low-latency applications, investigating selective physics computation (e.g., periodic rather than per-frame calculation) could reduce this cost while maintaining benefits.

Several promising directions emerge for extending this work. Multi-modal physics integration could combine visual pose estimation with Inertial Measurement Units (IMUs), force plates, or Electromyography (EMG) sensors, providing ground-truth physical measurements for even stronger supervision and enabling applications in sports biomechanics or rehabilitation monitoring. Hierarchical physics modeling could extend from isolated limbs to full-body coordination, capturing higher-level principles, such as center of mass dynamics, momentum conservation, balance maintenance, and whole-body energy optimization—relevant for complex activities such as dancing, martial arts, or gymnastics. Incorporating uncertainty quantification through Bayesian neural networks could provide not only classification predictions and physical metrics but also confidence estimates for both, crucial for safe decision-making in autonomous systems. Transfer learning strategies could pre-train physics-informed models on large-scale general action datasets, then fine-tune on specialized domains with limited data, potentially improving data efficiency in applications where annotated training examples are scarce. Finally, investigating adversarial robustness could reveal whether grounding in physical principles provides inherent defense against perturbations, as physically impossible adversarial examples might be detectable through violation of biomechanical constraints. These directions collectively point toward a future where action recognition systems understand human motion not merely as visual patterns but as physical phenomena governed by interpretable principles.

VII. CONCLUSION AND FUTURE WORK

This study presented a physics-informed neural network architecture for human action recognition that integrates biomechanical modeling directly into the learning process. By computing kinematic and kinetic features from estimated poses and fusing them with visual motion representations within a Transformer encoder, physics-aware models can learn to recognize gestures based on both appearance and underlying physical principles. The proposed method addresses a fundamental limitation of purely data-driven approaches: the lack of explicit grounding in the physical laws that govern human motion. Through a multi-objective physics-informed loss function, the model is guided toward solutions that are not only accurate but also physically coherent and interpretable.

Evaluation in police traffic gesture recognition demonstrated that this approach achieved 96.11% classification accuracy while maintaining high biomechanical feasibility (0.998) and realistic energy profiles (11.4 J). Ablation studies confirm that physics integration is the primary driver of performance, providing an 18.89% improvement over a baseline without. The benefits are most pronounced for temporally ambiguous gestures that are visually similar but physically distinct.

Beyond accuracy metrics, this approach produces interpretable physical measurements that can be validated against biomechanical principles, making it particularly suitable for safety-critical applications where model transparency and reliability are paramount. The differentiable physics engine and fusion strategy provide a general framework applicable to diverse action recognition tasks.

Although this evaluation focuses on a specific gesture recognition domain, the proposed method establishes a foundation for integrating domain knowledge into data-driven learning systems. By explicitly encoding the constraints and dynamics of human motion, this approach moves toward models that understand not merely what actions look like, but how and why they unfold as they do.

ACKNOWLEDGMENT

The authors acknowledge their colleagues at the Advanced Systems Engineering Laboratory (Ibn Tofail University) for valuable discussions.

REFERENCES

[1] K. Simonyan and A. Zisserman, "Two-Stream Convolutional Networks for Action Recognition in Videos." arXiv, 2014, https://doi.org/10.48550/ARXIV.1406.2199.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks." arXiv, 2014, https://doi.org/10.48550/ARXIV.1412.0767.

[3] J. Carreira and A. Zisserman, "Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset." arXiv, 2017, https://doi.org/10.48550/ARXIV.1705.07750.

[4] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "SlowFast Networks for Video Recognition." arXiv, 2018, https://doi.org/10.48550/ARXIV.1812.03982.

[5] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." arXiv, 2020, https://doi.org/10.48550/ARXIV.2010.11929.

[6] G. Bertasius, H. Wang, and L. Torresani, "Is Space-Time Attention All You Need for Video Understanding?" arXiv, 2021, https://doi.org/10.48550/ARXIV.2102.05095.

[7] A. O. Hashi, S. Z. M. Hashim, and A. B. Asamah, "Dynamic Adaptation in Deep Learning for Enhanced Hand Gesture Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15836–15841, Aug. 2024, https://doi.org/10.48084/etasr.7670.

[8] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Aug. 1997, https://doi.org/10.1162/neco.1997.9.8.1735.

[9] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition." arXiv, 2018, https://doi.org/10.48550/ARXIV.1801.07455.

[10] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting Skeleton-based Action Recognition." arXiv, 2021, https://doi.org/10.48550/ARXIV.2104.13586.

[11] C. Prabha, R. Singh, M. Malik, M. R. Pradhan, and B. Acharya, "Advanced Gesture Recognition in Gaming: Implementing EfficientNetV2-B1 for 'Rock, Paper, Scissors,'" *Engineering, Technology & Applied Science Research*, vol. 15, no. 3, pp. 23386–23392, June 2025, https://doi.org/10.48084/etasr.10373.

[12] M. Raissi, P. Perdikaris, N. Ahmadi, and G. E. Karniadakis, "Physics-Informed Neural Networks and Extensions." arXiv, 2024, https://doi.org/10.48550/ARXIV.2408.16806.

[13] Y. Xia, X. Zhou, E. Vouga, Q. Huang, and G. Pavlakos, "Reconstructing Humans with a Biomechanically Accurate Skeleton," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025, pp. 5355–5365, https://doi.org/10.1109/CVPR52734.2025.00504.

[14] T. Xiao and Y. F. Fu, "Biomechanical Modeling of Human Body Movement," *Journal of Biometrics & Biostatistics*, vol. 7, no. 3, 2016, https://doi.org/10.4172/2155-6180.1000309.

[15] G. Jocher, J. Qiu, and A. Chaurasia, "Ultralytics YOLO." Jan. 2023, [Online]. Available: https://github.com/ultralytics/ultralytics.

[16] D. A. Winter, *Biomechanics and Motor Control of Human Movement*. John Wiley & Sons, 2009.

[17] P. De Leva, "Adjustments to Zatsiorsky-Seluyanov's segment inertia parameters," *Journal of Biomechanics*, vol. 29, no. 9, pp. 1223–1230, Sept. 1996, https://doi.org/10.1016/0021-9290(95)00178-6.

[18] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[19] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization." arXiv, 2017, https://doi.org/10.48550/ARXIV.1711.05101.

[20] "police_officers_dataset." Kaggle, [Online]. Available: https://www.kaggle.com/datasets/zaynabh/police-officers-dataset.

[21] "zc402/traffic-gesture-datasets." May 07, 2025, [Online]. Available: https://github.com/zc402/traffic-gesture-datasets.

[22] R. Padilla, S. L. Netto, and E. A. B. Da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," in *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, July 2020, pp. 237–242, https://doi.org/10.1109/IWSSIP48289.2020.9145130.