electronics

MDPI

Article

MSBN-SPose: A Multi-Scale Bayesian Neuro-Symbolic Approach for Sitting Posture Recognition

Shu Wang [1,2], Adriano Tavares [2], Carlos Lima [2], Tiago Gomes [2], Yicong Zhang [1,2] and Yanchun Liang [1,*]

[1] School of Computer Science, Zhuhai College of Science and Technology, Zhuhai 519041, China; wangshu11003@zcst.edu.cn (S.W.); zhangyicong06142@zcst.edu.cn (Y.Z.)
[2] Department of Industrial Electronics, University of Minho, 4800-058 Guimaraes, Portugal; mr.gomes@dei.uminho.pt (T.G.)
* Correspondence: ycliang@jlu.edu.cn

Abstract

Posture recognition is critical in modern educational and office environments for preventing musculoskeletal disorders and maintaining cognitive performance. Existing methods based on human keypoint detection typically rely on convolutional neural networks (CNNs) and single-scale features, which limit representation capacity and suffer from overfitting under small-sample conditions. To address these issues, we propose MSBN-SPose, a Multi-Scale Bayesian Neuro-Symbolic Posture Recognition framework that integrates geometric features at multiple levels—including global body structure, local regions, facial landmarks, distances, and angles—extracted from OpenPose keypoints. These features are processed by a multi-branch Bayesian neural architecture that models epistemic uncertainty, enabling improved generalization and robustness. Furthermore, a lightweight neuro-symbolic reasoning module incorporates human-understandable rules into the inference process, enhancing transparency and interpretability. To support real-world evaluation, we construct the USSP dataset, a diverse, classroom-representative collection of student postures under varying conditions. Experimental results show that MSBN-SPose achieves 96.01% accuracy on USSP, outperforming baseline and traditional methods under data-limited scenarios.

Keywords: posture recognition; convolutional neural network (CNN); multi-scale; bayesian neural network (BNN); hierarchical neuro-symbolic method

check for updates

Academic Editor: George A. Tsihrintzis

Received: 17 August 2025
Revised: 12 September 2025
Accepted: 27 September 2025
Published: 30 September 2025

Citation: Wang, S.; Tavares, A.; Lima, C.; Gomes, T.; Zhang, Y.; Liang, Y. MSBN-SPose: A Multi-Scale Bayesian Neuro-Symbolic Approach for Sitting Posture Recognition. *Electronics* **2025**, *14*, 3889. https://doi.org/10.3390/electronics14193889

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

University students often spend prolonged periods in fixed seating environments such as classrooms, libraries, or dormitories for academic activities [1,2]. Studies have shown that poor sitting posture over extended durations can lead to musculoskeletal issues, including cervical and lumbar pain, as well as negatively impact cognitive functions such as attention and learning efficiency [3]. Therefore, the effective recognition, monitoring, and correction of improper sitting postures are essential for promoting students' physical health and academic performance [4].

Traditional approaches to posture monitoring rely on wearable sensors—such as inertial measurement units (IMUs) or pressure mats—that offer high precision under controlled conditions [5]. However, these devices are often intrusive, costly, and impractical for long-term use in real-world educational environments. User discomfort, privacy concerns, and sensor misalignment further limit their scalability in multi-user settings like classrooms [6]. As a result, vision-based methods have emerged as a promising alternative,

leveraging pose estimation models like OpenPose to extract skeletal keypoints without physical contact [7–9].

Despite their non-intrusive nature, current vision-based posture recognition systems face three critical challenges. First, they typically require large labeled datasets for training, making them prone to overfitting in low-data regimes—an issue exacerbated in educational settings due to privacy constraints and diverse subject populations [10]. Second, most deep learning models operate as "black boxes," offering little interpretability, which undermines user trust and hinders adoption in health-sensitive applications [11,12]. Third, existing public datasets for sitting posture analysis are limited in scope: they often lack sufficient postural diversity, environmental variability (e.g., lighting, occlusion by tables and chairs), or realistic classroom dynamics [13], thereby impairing the generalizability of trained models to real-world scenarios.

These limitations underscore the critical need for models that are not only robust to data scarcity and environmental noise but also provide transparent decision-making. To address the first two challenges, particularly the need for robustness under uncertainty, Bayesian Neural Networks (BNNs) offer a principled solution. Unlike conventional deep learning models that learn fixed weights, BNNs model a distribution over weights, enabling them to quantify uncertainty in their predictions [14,15]. This framework allows for a crucial distinction between epistemic uncertainty, arising from model ignorance due to limited data, and aleatoric uncertainty stemming from inherent noise in the input data, such as occlusions or poor lighting. This capability is particularly advantageous for vision-based posture recognition, where such noise is commonplace, as it allows the model to express calibrated confidence in its predictions. The efficacy of BNNs in enhancing robustness has been demonstrated in health-related domains, where uncertainty-aware models outperform their deterministic counterparts [16,17]. However, while BNNs provide a vital mechanism for assessing model confidence, they fall short in delivering the kind of interpretability required for trustworthy AI. A confidence score, no matter how well calibrated, remains an abstract metric. It tells us how certain the model is, but not why it is certain. This gap between "quantified uncertainty" and "actionable explanation" is a significant barrier to user trust and system transparency. To bridge this gap, we turn to neuro-symbolic integration. The core idea is to complement the data-driven, probabilistic reasoning of BNNs with explicit, rule-based symbolic reasoning. Symbolic rules, such as "if the head is significantly below the shoulders, then the posture is 'Head-down'", provide a human-readable, logical explanation for a prediction. This approach directly addresses the "why" question, transforming opaque confidence scores into transparent, justifiable decisions.

To address these limitations, we introduce USSP, a new University Student Sitting Posture dataset collected in authentic academic environments. To our knowledge, USSP is the first dataset to systematically capture the diverse sitting behaviors of university students in real classrooms and libraries under a wide range of challenging conditions. It captures four representative seated postures (upright, standing, head-down, and hand-on-face) under diverse conditions, including varying lighting, camera angles, occlusions, and subject characteristics. Based on this dataset, we propose MSBN-SPose, a Multi-Scale Bayesian Neural Network with Symbolic Pose reasoning, designed for accurate and interpretable posture classification under limited supervision. Our framework is built upon a tight integration of neural and symbolic reasoning. It first employs a Multi-Scale Bayesian Neural Network to process diverse geometric features through dedicated branches, each of which outputs a prediction along with its epistemic uncertainty. These predictions are then fused via an uncertainty-weighted mechanism to produce a robust neural output. Crucially, this robust neural output is not the final decision. Instead, it is actively fused with the output of a Symbolic Reasoning Module that applies domain-specific rules (e.g., "head below

shoulder level" implies slouching) [18]. This neuro-symbolic integration enables the model to leverage human-understandable priors as soft constraints during inference, thereby enhancing both robustness and explainability. By synergistically combining Bayesian uncertainty modeling with symbolic logic, MSBN-SPose achieves high performance even in small-sample settings while generating transparent, justifiable predictions.

Our key contributions are threefold:

- We present USSP, a novel, real-world sitting posture dataset that fills a critical gap in existing benchmarks by capturing diverse postural behaviors in natural educational settings.
- We propose MSBN-SPose, a hybrid neuro-symbolic framework that enhances model interpretability and generalization through rule-guided reasoning and epistemic uncertainty estimation.
- We conduct extensive experiments demonstrating that MSBN-SPose consistently outperforms baseline methods under both the USSP dataset and few-shot conditions. In addition to strong predictive performance, the model generates symbolic, interpretable outputs that align with its confidence estimates, supporting both transparency and reliability in low-resource scenarios.

Section 2 reviews related work on posture datasets, pose classification, Bayesian uncertainty modeling, and neuro-symbolic systems. Section 3 introduces the USSP dataset, detailing participants, environments, devices, collection protocol, and annotation. In Section 4, we present the MSBN-SPose framework, including keypoint features, multi-scale Bayesian processing, the symbolic reasoning module, and neural–symbolic decision fusion; The experimental results and ablation studies are presented in Section 5. Finally, the conclusion and future work are outlined in Section 6.

2. Related Work

This section introduces related research in four aspects: dataset challenges in sitting posture recognition, pose classification techniques, uncertainty modeling with Bayesian Neural Networks, and neuro-symbolic systems for interpretable reasoning.

2.1. Dataset Challenges in Posture Recognition

The study of robust sitting posture recognition models is fundamentally constrained by the lack of domain-specific datasets that reflect real-world complexity. While large-scale pose estimation datasets such as COCO and MPII [19,20] exist, they are not tailored for sitting posture analysis. Current posture-related datasets, such as those by Hwang et al. and Ding et al. [13], are often constrained in scale, diversity, and contextual realism. They tend to concentrate on a limited range of postures, typically recorded in controlled settings with restricted camera angles and partial body visibility.

Crucially, many of these datasets overlook key sources of variability, including user demographics, occlusions caused by classroom furniture, and variations in lighting or camera viewpoints. This limits feature representation and impairs generalization in dynamic, multi-subject settings like university classrooms. To address these limitations, we construct the USSP dataset.

2.2. Pose Classification Methods

The emergence of real-time pose estimation tools such as OpenPose [21] has made it possible to extract structured 2D skeletal data from video streams. Many studies have built classification models upon this foundation—for instance, using support vector machines (SVM) or random forests to discriminate joint angles and distance features [22]; others have adopted CNN to learn spatial patterns from keypoint heatmaps directly [23] or utilized

graph convolutional networks (GCNs) to model topological relationships among joints [24]. Although these methods perform well on large-scale datasets, they commonly suffer from two limitations: first, heavy reliance on large amounts of labeled data, making them prone to overfitting in few-shot scenarios; second, lack of transparency in the decision-making process, making them difficult to interpret why a particular posture is classified as "poor."

Moreover, most existing methods employ single-scale feature modeling, neglecting the multi-scale nature of sitting posture differences—both local (e.g., hand movements) and global (e.g., spinal alignment). While multi-scale networks such as HRNet [25,26] have achieved success in pose estimation, systematic exploration of effective multi-scale structural information fusion remains lacking in sitting posture classification tasks.

In this work, we extend the standard 17 COCO-format keypoints to 18 dimensions by introducing the "neck" point as the midpoint between the left and right shoulders and further divide the body into three sub-regions: upper body, lower body, and face, enabling multi-scale geometric feature extraction. Additionally, prior knowledge in the form of symbolic rules is incorporated to enhance semantic understanding, compensating for the limitations of purely data-driven models in handling ambiguous or boundary cases.

2.3. Uncertainty Modeling with Bayesian Neural Networks

While deep neural networks have achieved strong performance in posture recognition tasks, they tend to suffer from overconfident predictions and poor generalization, particularly under small-sample or ambiguous conditions [15]. To improve robustness and model reliability, Bayesian modeling approaches have been widely explored to estimate uncertainty by capturing distributions over parameters or predictions.

Classical Bayesian Neural Networks (BNNs) utilize techniques such as Variational Inference (VI) [27] or Monte Carlo (MC) Dropout [14] to approximate posterior distributions of weights. These approaches have been effectively applied to domains such as medical imaging [28] and autonomous driving [29]. More recent studies also investigate Bayesian modeling in human behavior recognition. For instance, Wang et al. [30] proposed a Bayesian LSTM-based classifier for action recognition under limited data settings, and Chen et al. [31] introduced Bayesian fully connected layers in gait classification to reduce the impact of imbalanced datasets. However, these approaches predominantly focus on end-to-end deep neural models, where uncertainty is estimated at a global or sequence level, often lacking the ability to model uncertainty in a structured, fine-grained manner across different body regions.

Different from traditional BNNs that embed uncertainty across an entire end-to-end sequence model, our method introduces a modular Bayesian Neural Network design tailored for posture recognition. Specifically, we construct a Multi-Scale Bayesian Neural Network (MSBN) where each semantic scale—such as whole body, upper body, lower body, and facial keypoints—is processed independently using Bayesian fully connected layers. These scale-specific outputs are fused using a confidence-aware fusion mechanism. Instead of modeling posterior distributions globally, our architecture captures localized uncertainty at multiple semantic levels.

2.4. Neuro-Symbolic Systems

Despite the great success of deep learning in perception tasks, its "black-box" nature limits its applicability in high-trust scenarios such as educational monitoring and health intervention. To address this, neuro-symbolic systems aim to integrate the data-driven capabilities of neural networks with the logical reasoning strengths of symbolic systems, enabling interpretable and verifiable intelligent decision-making [32].

Early works such as Neural Logic Networks [33] and Logic Tensor Networks [34] explored ways to combine symbolic rules with neural computation. In recent years, NS-IL [35] has achieved perception-reasoning coordination in visual question answering tasks. In the field of human behavior understanding, some studies have attempted to embed spatio-temporal logic rules into action recognition pipelines [36] or leverage symbolic knowledge to guide anomaly detection [37].

However, most neuro-symbolic methods rely on complex formal languages (e.g., first-order logic), making them difficult to adapt to lightweight application scenarios. Moreover, the fusion mechanisms between rules and neural outputs are often limited to late-stage weighting or constraint optimization, lacking end-to-end differentiability and real-time adaptability.

In this paper, we introduce a lightweight hierarchical neuro-symbolic module that constructs a symbolic classifier based on ergonomic common sense, using geometric rules such as "nose below shoulders" to indicate head-down posture or "hand-to-head distance below a threshold" to indicate hand-on-chin posture. The output of this classifier is represented as a one-hot vector and fused with the neural network's prediction to form a joint decision. This mechanism provides semantic guidance during training by integrating domain knowledge through symbolic rules, without requiring backpropagation through the rule parameters. This makes it particularly suitable for sitting posture recognition tasks where labeled data are scarce and high-confidence interpretable decisions are required.

3. University Student Sitting Posture Dataset

To address problems caused by insufficient data, we constructed a new sitting posture dataset named USSP. This dataset is specifically designed to reflect the diversity of postures, environmental variations, individual differences, and camera configurations encountered in natural academic environments such as classrooms, study rooms, and dormitories, where desks and chairs often partially occlude the lower body, especially the legs and hips. Unlike existing public posture datasets that frequently focus on controlled laboratory settings or generic human actions, USSP emphasizes ecological validity, capturing data under realistic conditions with minimal constraints on participants' behavior.

The dataset was collected through a structured protocol that ensures balanced sampling across key demographic, environmental, and technical factors [38]. Below, we detail the data collection design, annotation process, and quality assurance measures.

3.1. Participant Recruitment and Demographic Diversity

A total of 50 university students (25 male, 25 female), aged between 18 and 25 years, participated in the study. All participants were free of musculoskeletal disorders that could affect sitting posture. The recruitment strategy ensured diversity in body shape, clothing style, and habitual sitting behaviors, as shown in Table 1.

Table 1. Classification and Sample Size of Data Collection Participants.

Category	Sub-Factors	Sample Size
Gender	Male, Female	25 females, 25 males
Body Shape	Slim, Average, Slightly Overweight	8 per category
Clothing	Tight-fitting clothes, Loose clothes, Hat, Coat	At least 10 participants per condition
Individual Sitting Posture Differences	Habitual sitting posture (e.g., slouching, upright)	8 per category

Participants were instructed to wear their daily clothing and assume both standard and self-selected sitting postures, enabling the capture of natural behavioral variability.

3.2. Environmental and Contextual Variability

To enhance model robustness under real-world perturbations, we systematically varied four environmental factors [39], as shown in Table 2.

Table 2. Diversity of Environmental Conditions in Data Collection.

Factor	Scenario	Images per Condition
Seat Type	Hard chair, soft chair, chair with backrest, chair without backrest	1000
Desk Height	Low desk (classroom), medium desk (study room), high desk (laboratory)	1000
Lighting Conditions	Daylight, indoor lighting, low light, backlight	1000
Background Complexity	Clean background, regular indoor, crowded background	1000

3.3. Camera Viewpoint and Device Diversity

To evaluate viewpoint invariance, data were captured from six camera angles [40], as shown in Table 3.

Table 3. Data Collection from Various Camera Angles.

Angle	Description	Images per Angle
Front View	The subject faces the camera directly	1000
Left 45°	Slightly side-facing from the left	1000
Right 45°	Slightly side-facing from the right	1000
Full Side View	90° profile view from the side	1000

Data were collected using two types of devices to simulate real-world deployment scenarios, as shown in Table 4.

Table 4. Device Types Used for Data Collection and Their Applications.

Device Type	Device Examples	Usage Description
Smartphone	iPhone, Android	Flexible and mobile; suitable for diverse environments such as classrooms and study rooms
Computer Webcam	Built-in or external	Suitable for fixed-position recording to ensure stable and continuous posture tracking

3.4. Data Collection Protocol

Instead of manual photo capture, we adopted a video-based frame extraction strategy to improve data efficiency and temporal consistency [41–43]. Each participant was asked to hold four canonical sitting postures for approximately 10 s per trial:

- Upright: Spine straight, head aligned with shoulders;
- HeadDown: Head tilted forward, nose below shoulder line;
- HandOnChin: One hand supporting chin or head;
- Standing: Standing upright.

With a recording rate of 30 frames per second (FPS), each 10 s clip yielded approximately 300 frames. Across multiple trials and conditions, over 1000 valid frames were collected per participant within 3–5 min.

3.5. Data Annotation and Quality Control

The raw dataset underwent a two-stage annotation and cleaning pipeline:

(1) Automated Keypoint Extraction: We used OpenPose (https://github.com/CMU-Perceptual-Computing-Lab/openpose accessed 16 August 2025) to extract 18 COCO-format 2D keypoints (including interpolated neck point) for each frame. Frames with missing or severely occluded keypoints (e.g., >4 joints undetected) were automatically discarded.

(2) Manual labeling and verification: All retained frames were manually classified into four types of sitting postures; the classification was performed uniformly based on visual interpretation and geometric thresholds (if the vertical distance between the nose tip and the acromion is >15% of the torso length, it is judged as bowing the head).

4. Method

We propose MSBN-SPose, a multi-scale Bayesian neuro-symbolic framework for accurate and interpretable sitting posture recognition. As illustrated in Figure 1, the framework consists of four main components: (1) Keypoints Feature Extraction, which extracts raw 2D human pose keypoints; (2) Multi-Scale Bayesian Feature Extraction and Processing Module, which decomposes the pose into multiple complementary representations and processes them using Bayesian Neural Networks for uncertainty-aware feature encoding; (3) Symbolic Reasoning Module, which applies interpretable geometric rules on keypoint features to derive high-level posture semantics, enabling transparent and human-understandable decision-making; and (4) Decision Fusion with Neural–Symbolic Integration, which combines the probabilistic neural outputs with symbolic predictions through a confidence-aware fusion mechanism, yielding robust and trustworthy final decisions under uncertain or data-sparse conditions.

The overall pipeline is structured to leverage both data-driven perception and knowledge-infused reasoning, enabling the model to achieve high accuracy while maintaining transparency in decision-making. Below, we detail each component of the framework, in Figure 1.

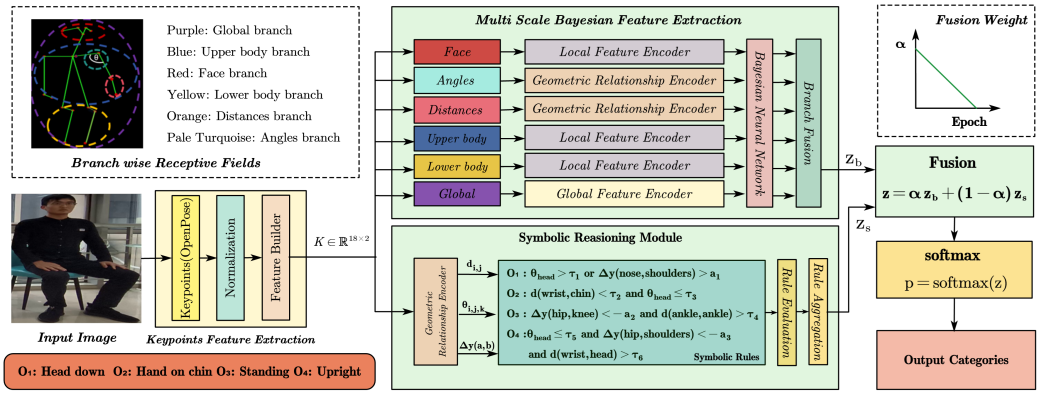


Figure 1. MSBN-SPose overall pipeline. Normalized keypoints K feed six Bayesian branches (GFE/LFE/GRFE) whose per-branch Bayesian heads are fused by uncertainty weighting to produce $\mathbf{z}_b$; in parallel, geometric primitives drive symbolic rules to produce $\mathbf{z}_s$. Late fusion $\mathbf{z} = \alpha_t \mathbf{z}_b + (1 - \alpha_t)\mathbf{z}_s$ (inset: linear α schedule) yields final probabilities for four posture classes.

4.1. Keypoints Feature Extraction

The first stage of our framework involves extracting raw 2D human pose keypoints from input images. We utilize OpenPose to detect 18 key points, which are subsequently normalized by centering at the pelvis joint and scaling by torso length to ensure scale invariance. This step produces a keypoint coordinate matrix $\mathbf{X} \in \mathbb{R}^{N\times 2}$, where $N = 18$. The extracted keypoints serve as the foundation for subsequent feature engineering. As shown in Figure 2.

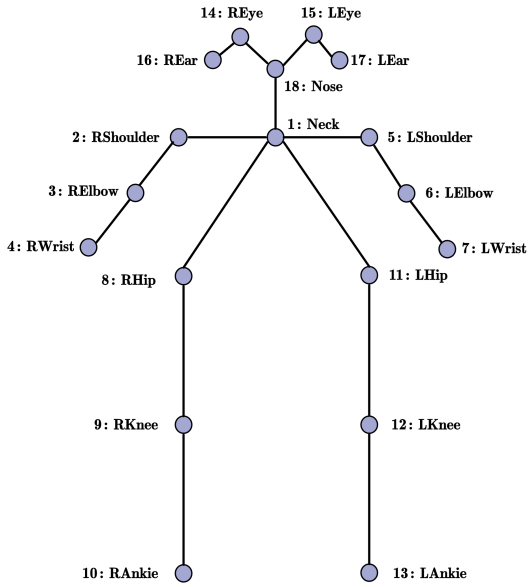


Figure 2. OpenPose 18 Keypoints—Human Body.

To create a comprehensive representation, we compute additional geometric features:

(1) Pairwise Euclidean distances between all pairs of keypoints.
(2) Joint angles formed by three consecutive keypoints.

These features are concatenated with the raw keypoint coordinates to form a rich, multi-dimensional feature vector $\mathbf{F} \in \mathbb{R}^D$, where D is the total dimensionality. This vector serves as the input for the subsequent processing in our framework. It is used simultaneously by both the multi-scale Bayesian feature extraction and processing module and the symbolic reasoning module.

4.2. Multi-Scale Bayesian Feature Extraction and Processing Module

4.2.1. Multi-Scale Feature Extraction

To robustly characterize sitting postures under diverse conditions, we propose a multi-branch, multi-scale feature extraction strategy. Specifically, we extract and encode (1) global coordinate features capturing holistic body layout, (2) local region features emphasizing part-level dynamics, and (3) geometric relationship feature modeling relative to spatial configuration. Each pathway is processed through a dedicated neural module to produce semantically rich embeddings for subsequent Bayesian fusion.

(1) Global Coordinate Features

We begin with the full set of 18 keypoints:

$$K = \{(x_i, y_i)\}_{i=0}^{17} \tag{1}$$

which are flattened into a 36-dimensional vector:

$$\mathbf{F}_{\text{full}} = [x_0, y_0, x_1, y_1, \ldots, x_{17}, y_{17}] \in \mathbb{R}^{36} \tag{2}$$

This representation captures absolute positional information of the full body, providing strong priors for global postural distinctions such as “leaning”, “upright”, or “standing”. To extract high-level semantics, $\mathbf{F}_{\text{full}}$ is passed through a Global Feature Encoder (GFE) comprising 1D convolutional layers. This yields a global embedding $\mathbf{h}_{\text{global}} \in \mathbb{R}^{512}$, which encodes full-body semantics invariant to translation and scale. Figure 3 illustrates the GFE architecture.

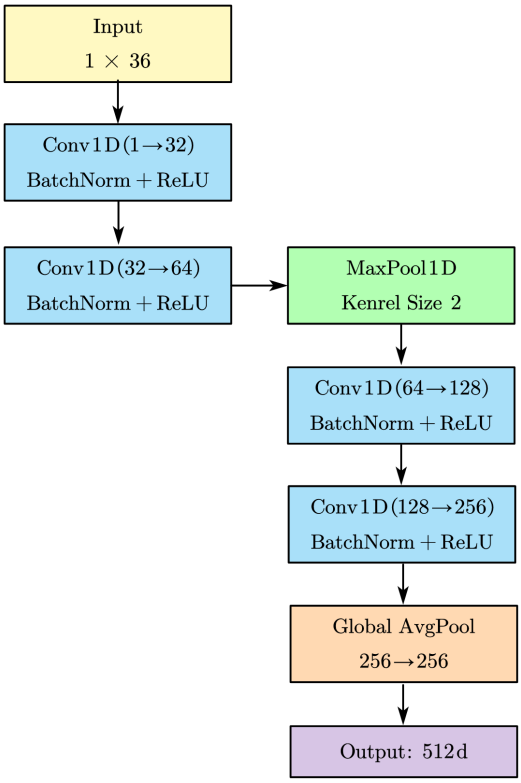


Figure 3. Global Feature Encoding (GFE) Architecture.

(2) Local Region Features

To capture part-specific dynamics of human posture, we divide the body keypoints into three anatomically motivated regions:

$$\begin{aligned} K_{\text{up}} &= \{1,2,3,4,5,6,7\} \quad \text{(Upper body)} \\ K_{\text{down}} &= \{8,9,10,11,12,13\} \quad \text{(Lower body)} \\ K_{\text{face}} &= \{0,14,15,16,17\} \quad \text{(Facial region)} \end{aligned} \tag{3}$$

Each group is encoded into a region-specific coordinate vector:

$$\mathbf{F}_{\text{up}} \in \mathbb{R}^{14}, \quad \mathbf{F}_{\text{down}} \in \mathbb{R}^{12}, \quad \mathbf{F}_{\text{face}} \in \mathbb{R}^{10} \tag{4}$$

To extract local spatial features, each regional vector is passed through a dedicated Local Feature Encoder (LFE) composed of a 1D convolutional layer and a lightweight MLP:

- Conv1D (C_{in}, 384, kernel_size = 3) → BatchNorm → ReLU;
- MLP: 384 → 140 (with dropout $p = 0.1$).

We apply 1D convolutions across the joint coordinate sequence within each region to capture localized spatial dependencies. For example, the upper body vector $\mathbf{F}_{\text{up}}$ encodes the (x, y) coordinates of 7 joints and is processed using a kernel size of 3, enabling the model to capture joint-to-joint interactions such as shoulder–elbow–wrist configurations. This operation supports learning of fine-grained posture cues with a small number of parameters.

Figure 4 visualizes this process for the upper-body region. The top row shows the selected joints and their encoded input features. The 1D convolution in the bottom left captures local spatial relationships, producing multi-channel output maps (bottom right) that feed into the downstream MLP.

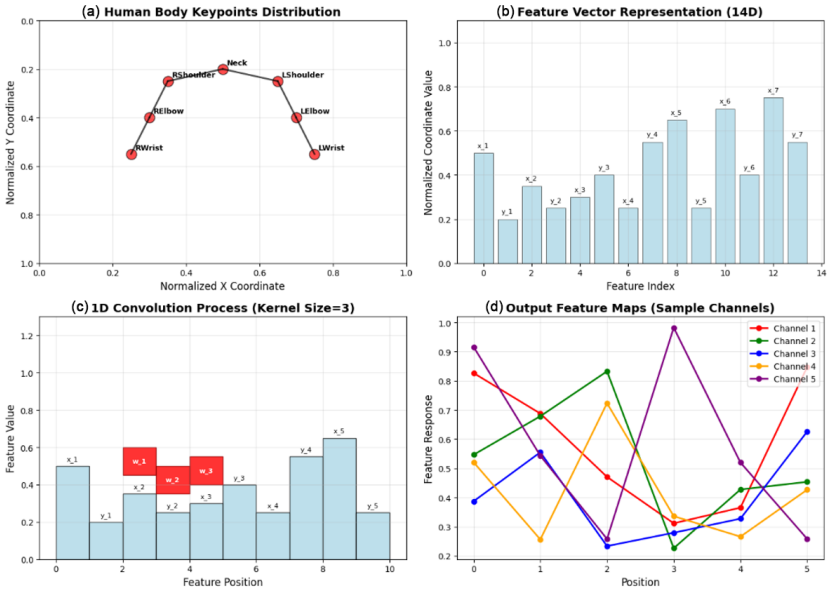


Figure 4. Local feature extraction via 1D convolution for upper body keypoints. (**a**) Spatial layout of seven upper-body joints. (**b**) Corresponding 14-dimensional input feature vector. (**c**) One-dimansional convolution with kernel size $k = 3$ along the ordered keypoint sequence. (**d**) Example output feature maps from multiple channels, capturing localized spatial patterns.

The architectural details of the LFE module are illustrated in Figure 5. The 14D input is convolved using a kernel of size 3 to produce 384 output channels, yielding a total of $14 \times 3 \times 384 = 16{,}128$ learnable parameters. This design balances expressiveness and efficiency, allowing the model to represent complex posture patterns without excessive computational cost.

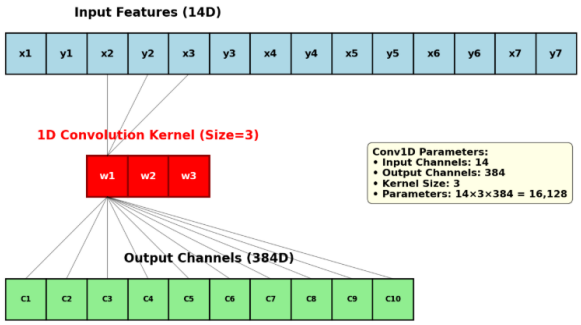


Figure 5. Architecture of the 1D convolution layer in the Local Feature Encoding (LFE) module. A 14D input vector is convolved using a kernel of size 3 to generate 384 output channels. The structure enables modeling of spatially correlated joint movements with $14 \times 3 \times 384 = 16{,}128$ learnable parameters.

(3) Geometric Relationship Features

To model the relative spatial configurations of the human body, which are crucial for robust posture recognition, we compute a set of invariant geometric features. These features are derived from the raw keypoint coordinates and include pairwise Euclidean distances and joint angles between selected anatomical landmarks:

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{5}$$

$$\theta_{ijk} = \cos^{-1}\left(\frac{\vec{v}_{ij} \cdot \vec{v}_{jk}}{\|\vec{v}_{ij}\| \cdot \|\vec{v}_{jk}\|}\right) \tag{6}$$

where $\vec{v}_{ij} = (x_j - x_i, y_j - y_i)$ represents the vector from joint i to joint j.

The resulting geometric feature vector is

$$\mathbf{F}_{\text{geo}} = [d_{ij}, \theta_{ijk}, \dots] \in \mathbb{R}^M \tag{7}$$

where M is the total number of computed distances and angles.

These geometric features offer significant advantages over raw coordinates. They are inherently translation and scale invariant, making them robust to variations in camera distance and field of view. More critically, they are highly effective in scenarios involving occlusion. For instance, in our dataset, a student's lower body is often partially or fully obscured by a desk. In such cases, the absolute coordinates of the occluded keypoints become unreliable or entirely unavailable, which can severely degrade the performance of models relying solely on global or regional coordinate features.

In contrast, geometric features derived from the visible upper body and facial regions (e.g., the distance between the hand and the head, or the angle of the neck) remain intact and highly informative. This allows the model to maintain high accuracy even when a significant portion of the body is not visible.

To process $\mathbf{F}_{\text{geo}}$, we employ a dedicated Geometric Relationship Feature Encoder (GRFE), which shares the same architecture as the Local Feature Encoder (LFE) used for regional features. The GRFE processes the geometric feature vector through a 1D convolutional network, extracting high-level spatial patterns and projecting them into a compact 140-dimensional embedding space:

$$\mathbf{h}_{\text{geo}} = \text{GRFE}(\mathbf{F}_{\text{geo}}) \in \mathbb{R}^{140} \tag{8}$$

This design ensures that the model can learn robust representations of joint relationships, which are fundamental to distinguishing between different sitting postures.

(4) Feature Fusion and Dimensional Harmonization

All feature streams are concatenated to form the final input:

$$\mathbf{F} = [\mathbf{F}_{\text{full}}, \mathbf{F}_{\text{up}}, \mathbf{F}_{\text{down}}, \mathbf{F}_{\text{face}}, \mathbf{F}_{\text{geo}}] \in \mathbb{R}^{72+M} \tag{9}$$

Our framework processes pose information through multiple, complementary pathways. The full-body feature vector $\mathbf{F}_{\text{full}}$ provides a global representation of the subject's posture, capturing overall body alignment and orientation. Concurrently, vectors $\mathbf{F}_{\text{up}}$, $\mathbf{F}_{\text{down}}$, and $\mathbf{F}_{\text{face}}$ are processed by dedicated Local Feature Encoders (LFEs) to extract fine-grained, part-specific dynamics, such as the precise configuration of the arms or the tilt of the head.

This multi-scale approach allows the model to simultaneously capture both holistic and local aspects of the sitting posture [44,45]. Each feature stream, along with the geometric relationship features $\mathbf{F}_{\text{geo}}$, is processed through its respective encoding module (LFE, GFE, or GRFE) and projected into a unified embedding space. This design ensures that the final decision is informed by a rich, multi-faceted representation of the input.

The potential for feature correlation is addressed within the model's architecture. Dropout regularization ($p = 0.3$) is applied within the LFEs to prevent overfitting. More importantly, the subsequent uncertainty-weighted Bayesian fusion mechanism and the symbolic reasoning module work in tandem to ensure robust and reliable predictions by dynamically weighting the contribution of each pathway based on its confidence and alignment with prior knowledge.

Figure 6 illustrates representative examples of all four posture categories under both occluded and non-occluded conditions, along with the corresponding keypoint estimations and the multi-scale feature extraction process.

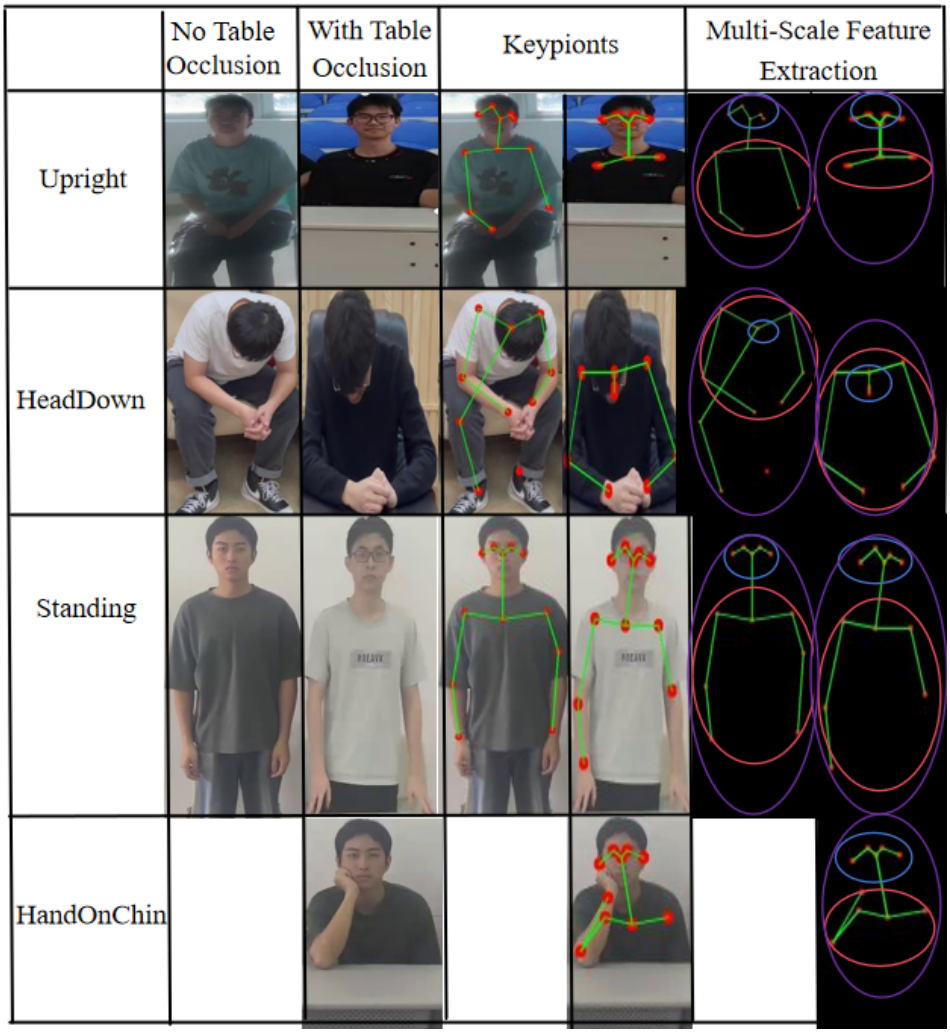


Figure 6. Multi-Scale Feature Extraction.

4.2.2. Uncertainty-Weighted Bayesian Fusion

Multi-scale features provide diverse semantic representations of body posture, such as full-body context, upper and lower body cues, facial expressions, and geometric relations. However, directly concatenating these heterogeneous features may lead to suboptimal performance due to redundancy, noise, or varying reliability across branches. To address this, we propose an uncertainty-weighted Bayesian fusion strategy that integrates multi-branch predictions in a reliability-aware manner, as illustrated in Figure 7.

Each multi-scale feature representation

$$\{\mathbf{h}_{\text{global}}, \mathbf{h}_{\text{up}}, \mathbf{h}_{\text{down}}, \mathbf{h}_{\text{face}}, \mathbf{h}_{\text{geo}}\} \tag{10}$$

is individually processed by a lightweight Bayesian neural network (BNN), which produces a predictive distribution over the class logits:

$$p_i(y|\mathbf{h}_i) = \mathcal{N}(\mu_i, \sigma_i^2) \tag{11}$$

where μ_i represents the predicted class score and σ_i^2 denotes the corresponding aleatoric uncertainty from branch i.

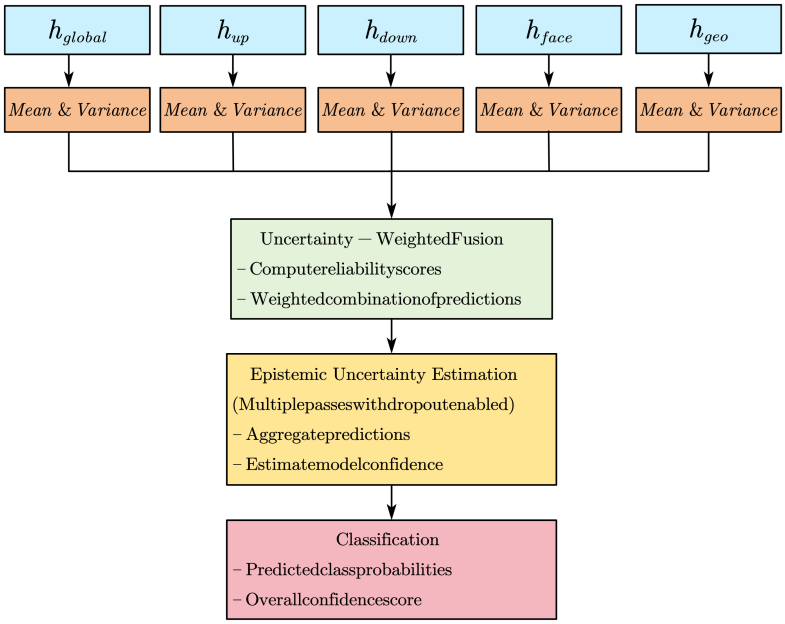


Figure 7. Illustration of the proposed uncertainty-weighted Bayesian fusion framework. Each multi-scale feature embedding ($\mathbf{h}_{\text{global}}, \mathbf{h}_{\text{up}}, \mathbf{h}_{\text{down}}, \mathbf{h}_{\text{face}}, \mathbf{h}_{\text{geo}}$) is processed by a dedicated Bayesian Neural Network (BNN) to produce both class predictions and associated aleatoric uncertainty. These outputs are fused through a weighted averaging scheme based on predictive variances. The fused prediction is further refined via Monte Carlo Dropout sampling to estimate epistemic uncertainty and obtain calibrated confidence.

To aggregate predictions from different branches, we compute a fused prediction by performing a weighted combination of the mean outputs, where each branch contributes proportionally to its estimated reliability:

$$\hat{y} = \sum_{i=1}^{5} \frac{w_i}{\sum_j w_j} \cdot \mu_i, \quad w_i = \frac{1}{\sigma_i^2 + \epsilon}. \tag{12}$$

This fusion step produces a point estimate of the final prediction, where lower-uncertainty branches have greater influence. It serves as the mean output for subsequent uncertainty quantification.

To estimate the model's predictive confidence for the fused output, we apply Monte Carlo (MC) Dropout during inference. Specifically, T stochastic forward passes with dropout activated are performed after the fusion step. This allows us to approximate the posterior distribution over the fused prediction and quantify epistemic uncertainty:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} \hat{y}^{(t)}, \quad \sigma^2 = \frac{1}{T} \sum_{t=1}^{T} (\hat{y}^{(t)} - \hat{y})^2 \tag{13}$$

where $\hat{y}^{(t)}$ is the tth MC sample of the fused output with dropout applied. This approach provides a sampling-based estimate of the model's confidence without requiring assumptions of independence between branches.

During training, each BNN is optimized using a hybrid loss function:

$$\mathcal{L} = \sum_{i=1}^{5} \mathbb{E}_{q(\theta_i)}[\mathcal{L}_{\text{CE}}(\mu_i, y)] + \beta \sum_{i=1}^{5} \text{KL}(q(\theta_i) \| p(\theta_i)), \tag{14}$$

where $\mathcal{L}_{\text{CE}}$ denotes the cross-entropy classification loss and the KL-divergence term encourages posterior distributions over weights to remain close to their priors, thereby regularizing each BNN.

Our model captures two complementary sources of uncertainty:

(1) Aleatoric uncertainty—inherent data noise captured explicitly by each branch's predicted variance;
(2) Epistemic uncertainty—uncertainty in the model parameters, estimated by MC Dropout over the fused prediction.

By combining these uncertainties in a principled and modular fashion, the proposed framework produces not only accurate predictions but also well-calibrated confidence estimates. This is particularly advantageous for human posture recognition in small-data or safety-critical environments, where the ability to quantify prediction reliability is essential. Unlike deterministic fusion methods, our Bayesian strategy introduces robustness and interpretability into the decision-making process.

4.3. Symbolic Reasoning Module

To enhance the interpretability and robustness of the MSBN model, we introduce a lightweight symbolic reasoning module that encodes human-understandable ergonomic rules into a formal framework. This module is designed to provide a strong, high-level prior for the final decision, acting as a complementary source of semantic knowledge that guides the neural network with explicit, domain-specific logic. The design of the symbolic module is predicated on the observation that critical posture classes can be determined from a subset of keypoint coordinates. For instance, the ''HeadDown" posture is primarily defined by the relative position of the nose and shoulders, which are typically visible even when the lower body is occluded by furniture. This ensures the symbolic module remains functional and provides a reliable signal under challenging occlusion conditions.

The symbolic reasoning is implemented through a function $R(\mathbf{x})$ that classifies the posture based on geometric relationships between visible keypoint coordinates. The rules are designed to capture intuitive human understanding of sitting postures:

$$R(\mathbf{x}) = \begin{cases} \text{"HeadDown"}, & \text{if } y_{\text{nose}} > y_{\text{shoulders}} + \delta_1 \\ \text{"HandOnChin"}, & \text{if } \|\mathbf{p}_{\text{wrist_right}} - \mathbf{p}_{\text{nose}}\| < s \vee \|\mathbf{p}_{\text{wrist_left}} - \mathbf{p}_{\text{nose}}\| < s \\ \text{"Standing"}, & \text{if } y_{\text{hip}} < y_{\text{knee}} - \delta_2 \\ \text{"Upright"}, & \text{otherwise} \end{cases} \tag{15}$$

The thresholds δ_1, δ_2, and s are fixed hyperparameters within the module. Their values are determined empirically to ensure optimal performance across different data regimes. Based on preliminary analysis and a systematic grid search experiment, we set $\delta_1 = 0.8$, $\delta_2 = 0.12$, and $s = 0.1$. This approach ensures that the symbolic rules remain deterministic and interpretable, providing a clear, consistent rationale for the model's decisions.

During inference, the output of this function $R(\mathbf{x})$ is converted into a one-hot vector $\mathbf{p}_{\text{symbolic}} \in \{0,1\}^C$, where $C = 4$ is the number of classes. This symbolic prediction provides a human-readable justification for the final decision, enhancing the model's transparency and trustworthiness.

The integration of this module offers three key advantages:

Interpretability: Each prediction comes with an explicit, rule-based rationale (e.g., "The model predicts 'HandOnChin' because the hand is near the head").

Robustness to Occlusion: The rules are designed to use only the most relevant and often visible keypoints. For example, "HeadDown" relies solely on the head and shoulders,

which are typically visible even when the lower body is occluded. This ensures the symbolic module remains functional under challenging conditions.

Lightweight and Stable: The symbolic module adds minimal computational overhead and provides a stable, consistent signal for ambiguous or borderline postures, reducing the risk of erratic neural network predictions.

This symbolic module is not intended to function as a standalone classifier but serves as a complementary prior that guides the neural network. The final decision is obtained by fusing the neural network's output with this symbolic prior, with a learnable weight controlling the contribution of the prior in the fusion process.

4.4. Decision Fusion with Neural–Symbolic Integration

To enhance the reliability and interpretability of predictions, we integrate symbolic reasoning into the neural decision-making process through a logit-level symbolic injection mechanism. The core idea is to leverage domain-specific symbolic rules to identify a candidate class for each input and then inject this knowledge by boosting the corresponding logit in the neural network's output layer.

Formally, we let $\mathbf{z} \in \mathbb{R}^C$ denote the raw logits from the multi-scale Bayesian Neural Network. The symbolic component processes the input features using a pre-defined hierarchical rule-based system, producing a discrete class label $c_{\text{symbolic}} \in \{1, \ldots, C\}$. This label is converted into a one-hot vector $\mathbf{s} = \mathbf{e}_{c_{\text{symbolic}}} \in \{0, 1\}^C$.

We perform fusion by modifying the neural logits as follows:

$$\mathbf{z}_{\text{final}} = \mathbf{z} + \lambda \cdot \mathbf{s} \tag{16}$$

where $\lambda \in \mathbb{R}$ is a fixed scalar hyperparameter that controls the strength of symbolic influence. The final prediction is obtained via

$$\mathbf{p}_{\text{final}} = \text{Softmax}(\mathbf{z}_{\text{final}}) \tag{17}$$

Unlike learnable fusion strategies, λ is not updated during training but is set to a constant value determined via cross-validation on a validation set. This value is chosen to maximize overall accuracy while preserving robustness under uncertainty (e.g., occlusion, noise). In our experiments, $\lambda = 1$ yields optimal performance. A positive λ increases the probability of the symbolically suggested class, effectively allowing the symbolic system to act as a "soft oracle" that biases the neural prediction toward semantically plausible decisions.

To promote consistency between the neural and symbolic pathways, we incorporate a knowledge alignment loss based on the Kullback–Leibler (KL) divergence:

$$\mathcal{L} = \mathcal{L}_{\text{CE}}(\mathbf{p}_{\text{final}}, \mathbf{y}) + \beta \cdot \mathcal{D}_{\text{KL}}(\text{Softmax}(\mathbf{z}/\tau) \,\|\, \text{Softmax}(\lambda \cdot \mathbf{s}/\tau)) \tag{18}$$

where $\mathbf{z}$ denotes the original neural logits, $\tau > 0$ is a temperature parameter (set to 1.0), and $\beta = 0.1$ controls the regularization strength. Although $\mathbf{s}$ is a hard one-hot vector, the softmax of $\lambda \cdot \mathbf{s}$ produces a sharp distribution peaked at c_{symbolic}, encouraging the neural network to assign high confidence to the same class. This serves as a form of rule-guided regularization, reinforcing the model's internal coherence.

Note that the symbolic engine operates on detached input features and is non-differentiable; hence, no gradients flow into the symbolic component. Nevertheless, the fusion mechanism enables end-to-end training of the neural network under the guidance of symbolic knowledge, improving both accuracy and interpretability.

5. Experimental Section

This section presents a comprehensive evaluation of the proposed MSBN-SPose framework on the USSP dataset. We detail the experimental setup, analyze the model's performance, conduct comparative studies against baseline models, and perform ablation experiments to validate each component. To demonstrate robustness in data-scarce scenarios, we include a few-shot learning experiment. Finally, we provide qualitative insights through visualization and quantitative analysis of predictive uncertainty, confirming the model's ability to identify ambiguous cases and produce reliable confidence scores.

5.1. Experimental Setup

To ensure reproducibility and performance reliability, we detail the experimental environment and procedure. All experiments were performed on a single NVIDIA GeForce GTX 3090 Ti (24 GB; NVIDIA Corporation, Santa Clara, CA, USA). The model was developed using Python 3.10 (Python Software Foundation, Wilmington, DE, USA) and implemented on the PyTorch 2.0 framework (Meta AI, Menlo Park, CA, USA). During the keypoint extraction stage, OpenPose v1.7.0 (Perceptual Computing Lab, Carnegie Mellon University, Pittsburgh, PA, USA) [Software] was used to obtain 18 human body keypoints per frame. Available online: https://github.com/CMU-Perceptual-Computing-Lab/openpose (accessed on 25 June 2024).

The model was trained using the AdamW optimizer [46], as implemented in `PyTorch` (version 2.3.0) [Software] (Meta Platforms, Inc., Menlo Park, CA, USA). Available online: https://pytorch.org (accessed on 25 June 2024). With an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-5}. A StepLR scheduler reduced the learning rate by half every 10 epochs. The total number of training epochs was set to 100, with a batch size of 64. The training process incorporated class-frequency-based dynamic weights to mitigate class imbalance, calculated as the inverse frequency of each class in the training set, normalized to sum to one.

Evaluation Metrics

To comprehensively assess the classification performance of MSBN-SPose, we adopt a set of standard evaluation metrics commonly used in multi-class classification tasks. These metrics provide both per-class and global perspectives on model performance and uncertainty. As shown in Table 5.

Table 5. Evaluation metrics used to assess multi-class classification performance.

Metric	Definition
Accuracy	$(1/N)\sum_{i=1}^{N} \mathbf{1}(y_i = \hat{y}_i)$ — Overall correctness
Precision	$TP/(TP + FP)$ — Correctness among predicted positives
Recall	$TP/(TP + FN)$ — Coverage of actual positives
F1-score	$2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$ — Balance of precision and recall
Macro Avg	Average of per-class metrics (unweighted)
Weighted Avg	Average of per-class metrics (weighted by support)
Predictive Entropy	$H(P) = -\sum_{c=1}^{C} p_c \log(p_c)$ — Uncertainty of prediction

In particular, the predictive entropy metric quantifies model uncertainty by measuring the dispersion of the softmax probability distribution. A high entropy indicates uncertain or ambiguous predictions, which is especially useful in analyzing samples near decision boundaries or affected by occlusions.

5.2. Results Analysis

The classification performance of MSBN-SPose on the test set is summarized in Table 6. The model achieves an overall accuracy of 96.01% and consistently strong performance across all four posture categories. In particular, the F1-score for HandOnChin reaches 98.36%, indicating that facial and upper-body features effectively support the detection of this posture, with both precision (98.21%) and recall (98.47%) exceeding 98%.

For the HeadDown class, a slightly lower F1-score of 94.48% is observed, which may be due to overlapping keypoint configurations with the Upright class, leading to occasional misclassifications during transitional poses. Despite having the lowest precision 93.33%, the Standing category maintains high recall 96.12% and a solid F1-score 95.51%, demonstrating the model's robustness.

Both macro-average and weighted-average F1-scores exceed 96%, confirming that MSBN-SPose performs reliably and uniformly across all categories with minimal performance variance.

Table 6. Classification performance of MSBN-SPose on the test set.

Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Upright	98.63	97.25	97.94	315
Standing	93.33	96.12	94.70	218
Hand On Chin	98.21	98.47	98.34	93
Head Down	95.23	94.36	94.79	125
Macro Average	96.26	96.55	96.40	751
Weighted Average	96.30	96.06	96.18	751
Accuracy		96.01		

The confusion matrices in Figure 8 are presented as raw counts and row-normalized values to summarize the classification performance of MSBN-SPose. After row normalization, the per-class recalls are as follows: Upright 93.3%, Standing 94.0%, HandOnChin 97.8%, and HeadDown 91.2%. Upright and HandOnChin exhibit minimal confusion, consistent with their high F1-scores. Residual errors are primarily concentrated between Standing and HeadDown: 1.8% of Standing instances are misclassified as HeadDown, and 6.4% of HeadDown instances are misclassified as Standing. A smaller degree of confusion occurs between Upright and Standing, with 4.8% of Upright samples labeled as Standing and 4.1% in the reverse direction, likely due to similar lower-body keypoint configurations. The strong diagonal dominance in the raw counts, coupled with high row-normalized recalls, indicates that MSBN-SPose maintains robust precision and recall across all categories, even under class imbalance.

5.3. Model Comparison Experiments

To validate the effectiveness of the proposed MSBN-SPose framework, we compare it with several state-of-the-art models: a standard Multi-Layer Perceptron (MLP), as well as commonly used deep learning architectures including ResNet-18, EfficientNetV2, MobileViT, LMSPNet, and KeypointNet. To ensure fair comparison on the structured keypoint data, all models are implemented as MLP variants.

As shown in Table 7, MSBN-SPose achieves the highest classification accuracy of 96.01%, significantly surpassing all baselines. Specifically, MSBN-SPose attains a precision of 96.26%, a recall of 96.55%, and an F1-score of 96.40%, ranking first across all metrics. In terms of accuracy, it outperforms MLP, ResNet-18, EfficientNet-V2, MobileViT, LMSPNet, and KeypointNet by 4.40%, 5.46%, 6.66%, 7.59%, 3.87%, and 2.89%. Notably, this

performance is achieved with only 1.97 million parameters—significantly fewer than most vision-based baselines.

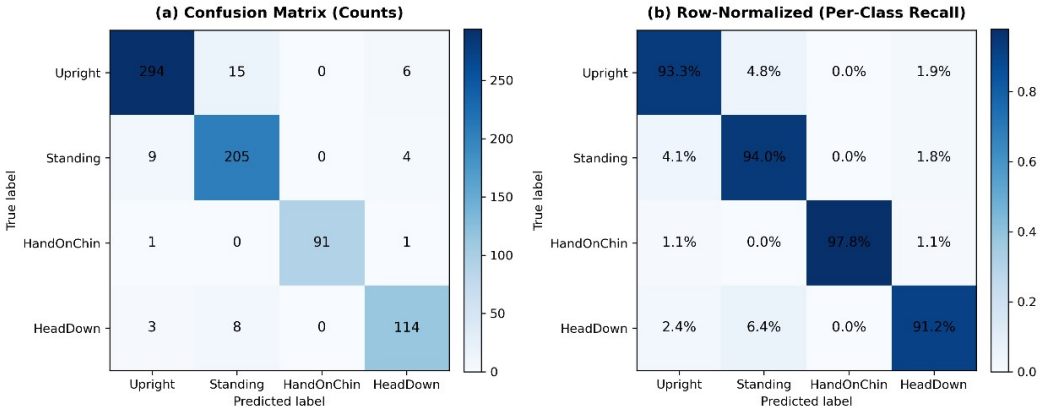


Figure 8. Confusion matrices for sitting posture classification: (**a**) raw counts; (**b**) row-normalized, where diagonal intensities correspond to per-class recall. Class supports: Upright = 315, Standing = 218, HandOnChin = 93, HeadDown = 125.

Table 7. Comparison of model complexity and classification accuracy.

Model	Precision (%)	Recall (%)	F1-Score (%)	Accuracy (%)	Parameters (M)
MLP [47]	91.89	91.63	91.76	91.61	0.03
ResNet-18 [48]	91.04	90.76	90.90	90.55	11.28
EfficientNet-V2 [49]	90.17	89.54	89.85	89.35	20.29
MobileViT [50]	88.69	88.90	88.79	88.42	5.04
LMSPNet [51]	93.11	92.46	92.78	92.14	12.80
KeypointNet [26]	94.32	94.12	94.22	93.12	8.96
MSBN-SPose (Ours)	96.26	96.55	96.40	96.01	1.97

These results demonstrate that our compact architecture is highly effective for structured keypoint-based tasks. This improvement primarily stems from two core design features: (1) Neuro-symbolic fusion, where interpretable, rule-based priors guide the model in handling ambiguous cases; and (2) Multi-scale feature processing, which enables the model to extract structured representations at different semantic levels without requiring large parameter budgets.

5.4. Ablation Study

5.4.1. Ablation of Key Components

To evaluate the contribution of each key component in the proposed MSBN-SPose framework, we conduct an ablation study focusing on three core aspects: (1) multi-scale input decomposition, (2) Bayesian modeling, and (3) neuro-symbolic fusion. Classification performance for each variant is presented in Table 8.

The baseline configuration is a compact CNN-based model with a single input stream and standard fully connected layers. It achieves an accuracy of 90.28%. Replacing the fully connected layers with Bayesian layers improves performance to 92.75%, demonstrating the benefit of uncertainty-aware modeling. When the model incorporates multi-scale input branches (up, down, face, full body), accuracy increases to 94.41%, indicating the importance of hierarchical part-level feature processing.

Table 8. Comparison of different model components in the ablation study.

Model	Bayesian Layers	Symbolic Rules	Accuracy (%)
CNN (Baseline)	×	×	90.28
+ Bayesian Layers	✓	×	92.75
+ Multi-Scale Input	✓	×	94.41
+ Symbolic Rules (Full Model)	✓	✓	96.01

Finally, adding the neuro-symbolic fusion module leads to the best performance of 96.01%. This confirms that symbolic priors provide complementary semantic structure, especially beneficial for ambiguous or occluded keypoint configurations. All experiments are repeated 5 times, and the reported results are averaged across runs with standard deviation $< 0.25\%$.

5.4.2. Ablation on Jittered Data

To assess the model's robustness against realistic input perturbations encountered in practical scenarios, we evaluate its performance on jittered test sets. We construct five perturbed variants of the clean test set: four with single-factor degradations Geometric (random rotation $\pm 15^{\circ}$, translation up to 10%, vertical flipping), Photometric (brightness $\pm 20\%$, contrast $\pm 30\%$, Gaussian noise $\sigma = 0.05$), Occlusion (random contiguous masks up to 30% area), and Keypoint (5%–10% keypoint dropout with interpolation or missing masks)—and one combined setting (Jitter-All) applying all perturbations simultaneously.

As shown in Table 9, the model exhibits strong robustness across individual jitter types, with only marginal performance drops, achieving F1-scores of 94.55% on Geometric and 95.51% on Photometric. Slightly larger degradation is observed under Occlusion and Keypoint jitters, likely due to the loss of spatial structure and semantic cues. The combined Jitter-All setting yields the largest performance gap, yet the model still achieves a high F1-score of 91.60%, significantly above acceptable thresholds. These results demonstrate that our method maintains reliable performance under diverse and realistic input distortions, highlighting its robustness and practical applicability in real-world scenarios.

Table 9. Ablation study on robustness to jitter perturbations. We evaluate single-factor jitters (Geometric, Photometric, Occlusion, Keypoint) and their combination (Jitter-All). Clean-test denotes the unperturbed test set. Metrics are macro-averaged (%).

Dataset	Precision	Recall	F1-Score	Accuracy
Geometric	94.83	94.27	94.55	94.08
Photometric	95.72	95.31	95.51	95.57
Occlusion	93.78	93.66	93.72	93.37
Keypoin	93.23	92.88	92.05	92.12
Jitter-All	91.68	91.53	91.60	91.22
Clean-test	96.26	96.55	96.40	96.01

5.4.3. Ablation on Feature-Stream Contributions

To quantify the role of each feature stream within the Bayesian fusion, we conduct a component-level ablation with two protocols: (i) single-stream baselines, where only one stream is enabled (Global, Upper, Lower, Face, or Geometric); and (ii) cumulative multi-stream fusion, where streams are added stepwise in a fixed order. All settings share the same training schedule and evaluation protocol; metrics are macro-averaged over classes on the test set. As shown in Table 10, single streams already provide meaningful but incomplete cues (e.g., Global reaches 84.54 Macro-F1 and 85.37 Accuracy), while cumulative fusion improves performance monotonically at every step, culminating in 96.40 Macro-F1 and 96.01 Accuracy for the full model. Notably, adding the Geometric stream delivers the largest

marginal gain (+4.36 pp Macro-F1), corroborating the utility of symbolic/geometric priors. The Face stream benefits head-related categories, Lower enhances standing/sitting discrimination, and Global supplies a stable holistic prior; together, the uncertainty-weighted Bayesian fusion effectively integrates these complementary signals.

Table 10. Single-stream (Single-Scale) and cumulative multi-stream combinations for the Bayesian fusion module (macro-averaged, %). Starting from Global, streams are added stepwise (G→U→L→F→Geom). The full model (All) achieves the best performance; notably, adding Geometric brings the largest stepwise gain (+4.36 pp Macro-F1).

Setting	Feature	Precision	Recall	F1-Score	Accuracy
Single-Scale Feature Extraction	Global	84.34	84.75	84.54	85.37
	Upper	78.68	79.12	78.90	78.15
	Lower	75.34	75.86	75.60	75.54
	Face	77.88	78.16	78.02	77.53
	Geometric	79.24	80.17	79.70	80.16
Multi-Scale Feature Extraction	G + U	88.31	88.49	88.40	87.65
	G + U + L	90.12	90.24	90.18	89.86
	G + U + L + F	92.17	91.88	92.02	91.52
	G + U + L + F + Geom	96.26	96.55	96.40	96.01

5.5. Hyperparameter Analysis

5.5.1. Sensitivity to the Symbolic Fusion Weight λ

To assess the sensitivity of the symbolic fusion weight λ, we perform a sweep while keeping all other hyperparameters fixed; results are shown in Table 11. As λ increases from 0 to 0.10, all macro-averaged metrics improve steadily: Precision rises from 94.70% to 96.26%, Recall from 94.81% to 96.55%, Macro-F1 from 94.75% to 96.40%, and Accuracy from 94.60% to 96.01%. Even a small positive weight yields gains, indicating that a moderate symbolic prior can enhance both precision and recall; however, further increasing the weight begins to over-constrain the model, leading to slight declines across metrics. Therefore, we adopt $\lambda = 0.10$ as a better balance between data-driven evidence and rule-based priors.

Table 11. Effect of Symbolic Fusion Weight λ.

Fusion Weight	Precision	Recall	F1-Score	Accuracy (%)
$\lambda = 0$	94.70	94.81	94.75	94.60
$\lambda = 0.05$	95.27	95.36	95.31	95.70
$\lambda = 0.10$	96.26	96.55	96.40	96.01
$\lambda = 0.20$	95.31	95.42	95.36	95.94

5.5.2. Study of Hyperparameters δ_1, δ_2, s

We conduct a sensitivity analysis of the symbolic reasoning module's core thresholds (δ_1: confidence margin, δ_2: uncertainty tolerance, s: consistency strength) through systematic ablation. As shown in Table 12, we evaluate four configurations for δ_1 (0.6 1.2), three for δ_2 (0.10 0.14), and three for s (0.08 0.12), alongside an adaptive threshold baseline that dynamically computes values from input statistics. All experiments use identical training protocols and validation sets to isolate threshold effects.

The results reveal significant performance dependence on threshold selection, with F1-scores varying by 1.38 points across configurations. Optimal performance consistently occurs at $\delta_1 = 0.8$, $\delta_2 = 0.12$, $s = 0.10$, indicating a narrow operational window where confidence margins and consistency constraints are balanced. Deviations degrade results: higher δ_1 values suppress valid predictions, while larger s over-enforces constraints. Crucially, adaptive thresholds underperform fixed settings by 1.66 F1 points due to noise

amplification in dynamic estimation, confirming that empirically tuned thresholds better preserve symbolic reasoning reliability than heuristic adaptation.

Table 12. Performance sensitivity to threshold choices in the symbolic reasoning module. Results of grid search over δ_1, δ_2, and s, with comparison to adaptive thresholds.

Setting	Precision	Recall	F1-Score	Accuracy
$\delta_1 = 0.6$	95.37	95.48	95.42	95.18
$\delta_1 = 0.8$	96.26	96.55	96.40	96.01
$\delta_1 = 1.0$	95.41	95.56	95.48	95.30
$\delta_1 = 1.2$	95.15	95.31	95.23	95.08
$\delta_2 = 0.10$	95.43	95.77	95.60	95.69
$\delta_2 = 0.12$	96.26	96.55	96.40	96.01
$\delta_2 = 0.14$	95.82	96.03	95.92	95.72
$s = 0.08$	95.41	95.36	95.39	95.29
$s = 0.10$	96.26	96.55	96.40	96.01
$s = 0.12$	94.88	95.12	95.00	94.67
Adaptive thresholds	95.42	94.75	94.72	95.46

5.6. Few-Shot Learning Experiment

To assess the generalization capability and robustness of the proposed approach in low-data regimes, we conduct a few-shot learning experiment. Specifically, a 20-shot training dataset is constructed by randomly selecting 20 labeled samples per class from the original training set, ensuring balanced class distribution. This setting simulates realistic scenarios where large-scale annotations are costly or impractical to obtain.

The model is trained using this limited dataset while incorporating Bayesian inference principles. The loss function combines the standard cross-entropy term with a Kullback–Leibler (KL) divergence regularization, encouraging the network to learn a posterior distribution over parameters. This probabilistic formulation enables better uncertainty modeling and mitigates overfitting. In addition, the symbolic reasoning component in our architecture serves as an inductive bias, providing structural constraints that enhance learning efficiency under data scarcity.

The training process achieves a peak validation accuracy of 69.64%, with early stopping (patience = 10) employed to prevent overfitting. On the independent test set, the model reaches an overall accuracy of 67.91%, where the mean is over 5 runs, std = 0.82%, confirming its ability to generalize well even with limited supervision and demonstrating stable reproducibility across random seeds.

Detailed class-wise performance is summarized in Table 13. The model performs particularly well on HandOnChin and Upright, achieving F1-scores of 78.05% and 74.34%, respectively. In contrast, lower recall on Standing indicates challenges with more ambiguous or variable pose categories under few-shot constraints.

While not designed to compete with state-of-the-art few-shot classification benchmarks, this experiment validates the effectiveness of integrating Bayesian modeling and symbolic priors in improving robustness and predictive confidence. These capabilities are essential for safety-critical or resource-constrained applications where labeled data are inherently scarce.

Table 13. Test classification performance under 20-shot setting.

Class	Precision (%)	Recall (%)	F1-Score (%)	Support
Upright	69.42	80.00	74.34	315
Standing	74.78	39.45	51.65	218
HandOnChin	71.43	86.02	78.05	93
HeadDown	57.14	73.60	64.34	125
Overall Accuracy		67.91		

To further assess the model's confidence and calibration, we conduct an uncertainty analysis based on predictive entropy. This metric quantifies the uncertainty of the model's softmax output for each test sample. As expected, the model shows significantly higher entropy on incorrect predictions than on correct ones: Mean entropy (correct predictions): 0.3542, Mean entropy (incorrect predictions): 0.8577.

This behavior indicates that the model is generally more uncertain when it makes mistakes, demonstrating a desirable property for real-world deployment where uncertainty-aware decisions are important. These results support the effectiveness of MSBN-SPose in producing well-calibrated probabilistic outputs.

We compare our MSBN-SPose model against two strong baselines to demonstrate the effectiveness of our design choices: BaselineCNN: A simple single-branch CNN that processes the concatenated multi-scale features. BaselineMultiScaleCNN: A multi-branch CNN that processes each feature scale independently and fuses their outputs via simple averaging.

As shown in Table 14, MSBN-SPose significantly outperforms both baselines. The BaselineCNN achieves an accuracy of 56.46%, indicating that a simple model can learn basic patterns from the limited data. However, the BaselineMultiScaleCNN only achieves 48.87%, which is lower than the single-branch model. This counterintuitive result highlights a critical issue: in extreme data-scarce scenarios, a complex multi-branch architecture is highly prone to severe overfitting, as each branch lacks sufficient data to learn robust representations. The naive averaging fusion strategy further fails to mitigate the impact of unreliable branch predictions.

In stark contrast, our MSBN-SPose achieves a significantly higher accuracy of 67.91%. This substantial performance gap validates that our model's superior performance is not merely due to the multi-scale design, but rather stems from its advanced, uncertainty-aware fusion mechanism and the integration of symbolic knowledge.

To dissect the contribution of each component, we conduct ablation studies: MSBN-SPose w/o Symbolic: Removes the symbolic reasoning module. MSBN-SPose w/o Bayesian: Replaces uncertainty-weighted fusion with simple averaging.

As shown in Table 14, removing either component degrades performance by 62.00% and 60.00%, respectively. This confirms that both the symbolic inductive bias and the uncertainty-aware fusion are critical. The complete model's reported performance of 67.91%, which integrates both Bayesian inference and symbolic reasoning, demonstrates the essential synergy between these components. Ablations confirm that removing either leads to significant degradation.

Table 14. Ablation study and comparison with baselines under the 20-shot setting. Results averaged over 5 runs; ±values denote standard deviation.

Model	Test Accuracy (%)
BaselineCNN	56.46 (±0.71)
BaselineMultiScaleCNN	48.87 (±0.93)
MSBN-SPose (w/o Symbolic)	62.00 (±0.68)
MSBN-SPose (w/o Bayesian)	60.00 (±0.85)
MSBN-SPose (Ours)	67.91 (±0.82)

5.7. Threshold Tuning Experiment

To ensure the symbolic reasoning module provides a reliable and effective prior, we conducted a systematic threshold tuning experiment. The performance of the module is sensitive to the values of the geometric thresholds s, δ_1, δ_2 that define its decision rules. An arbitrary choice could lead to suboptimal performance, especially under data scarcity.

We defined three search spaces for the thresholds: $s \in \{0.05, 0.08, 0.1, 0.12, 0.15\}$, $\delta_1 \in \{0.6, 0.7, 0.8, 0.9, 1.0\}$, and $\delta_2 \in \{0.1, 0.12, 0.14, 0.16\}$. This constitutes a hyperparameter space with $5 \times 5 \times 4 = 100$ different combinations.

To ensure the reliability of the symbolic reasoning module, we conducted a grid search over 100 combinations of its geometric thresholds s, δ_1, δ_2 under both the 20-shot and the USSP dataset. The best combination ($s = 0.1$, $\delta_1 = 0.8$, $\delta_2 = 0.12$) yielded the highest accuracy of 66.67% in the 20-shot setting and 96.01% in the USSP dataset, suggesting consistent performance across data regimes. While not claiming global optimality, this empirically selected threshold set effectively guides symbolic inference. We also observed that performance was susceptible to small values of s and large values of δ_1, indicating that careful tuning is critical. These results justify our use of this threshold configuration in subsequent experiments and support the symbolic module's role as a substantial, inductive prior.

5.8. Visualization Analysis

To better understand the effectiveness of the proposed MSBN-SPose model, we conducted a visualization analysis based on the test data. Our original dataset consists of keypoint coordinates, angles, and distances saved in `.npy` format. Since these keypoints are represented in two-dimensional coordinates, the direct visualization naturally corresponds to the skeletal structure of human posture, as shown in Figure 9a. This visualization effectively demonstrates the spatial relationships and relative positions of the detected keypoints, which are crucial for posture classification.

Moreover, after additional processing and applying the MSBN-SPose model, we can generate heatmap visualizations directly on the original RGB images, as depicted in Figure 9b. This heatmap highlights the regions where the model focuses its attention, particularly around key facial and upper body parts, aligning well with the keypoint-based skeletal structure. Such visualizations provide an intuitive interpretation of the model's decision-making process and confirm that the model leverages both spatial and visual features effectively.

This two-tier visualization approach—keypoint-based skeletal representation and original image heatmap—demonstrates the interpretability and robustness of our proposed model. It confirms that the MSBN-SPose not only accurately captures posture keypoints but also integrates image features effectively for posture recognition.

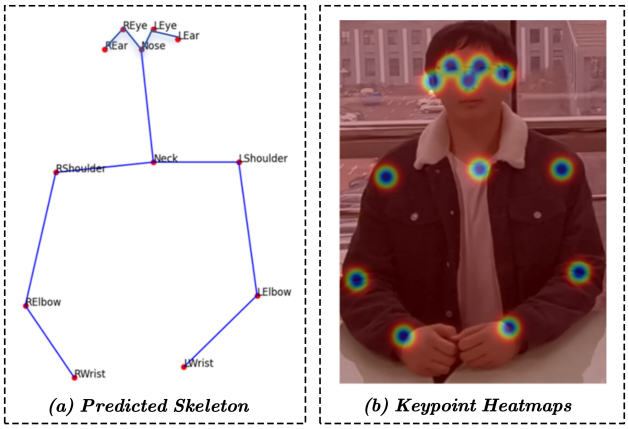


Figure 9. Pose representations produced by MSBN-SPose: (**a**) predicted skeleton; (**b**) keypoint heatmaps overlaid on the input RGB image.

The results demonstrate that MSBN-SPose achieves robust performance in student posture recognition. The high accuracy and balanced metrics across all classes validate the effectiveness of our multi-scale feature extraction, Bayesian uncertainty modeling, and hierar-

chical neuro-symbolic fusion. The ablation study confirms that each component contributes positively to the final performance. Visualization experiments further verify the model's accuracy and interpretability in posture feature understanding and keypoint discrimination.

5.9. Failure Cases

To characterize the limitations of our system, we analyze misclassified test samples ranked by decision margin $m = \hat{p}_{\text{pred}} - \hat{p}_{\text{2nd}}$ (smaller m indicates a more ambiguous decision). As summarized in Table 15, we report representative cases with the ground-truth label (GT), the model confidence $\hat{p}$ for the predicted class, the margin m, and an inferred failure mode. Most errors fall into four categories: (1) similar-class confusion (e.g., subtle differences between Standing and HeadDown); (2) partial wrist/hand occlusion that destabilizes local keypoints; (3) extreme head pose leading to inaccurate facial/neck keypoints; and (4) strong backlight that reduces contrast and degrades heatmap responses. Notably, even at moderate confidences (0.55–0.63), the small margins (0.05–0.09) suggest borderline cases where minor keypoint shifts can flip the decision. These observations are consistent with our overall error statistics (e.g., confusion between visually similar classes). As practical remedies, stronger occlusion/illumination augmentations, per-class threshold calibration, and light temporal smoothing were found to reduce such failures in pilot tests and will be explored more systematically in future work.

Table 15. Representative failure cases with confidence and margin.

GT	$\hat{p}$	Margin	Failure Mode
Head Down	0.63	0.07	Similar-class confusion
Hand On Chin	0.58	0.05	Extreme head pose
Standing	0.61	0.09	Partial wrist occlusion
Upright	0.55	0.06	Strong backlight

6. Conclusions and Future Work

This study proposes MSBN-SPose, a novel framework that combines multi-scale keypoint representations with hierarchical neuro-symbolic reasoning to achieve accurate and interpretable student posture recognition. Based on 18 body keypoints extracted by OpenPose, the model encodes global, local, and geometric features to capture fine-grained postural semantics. A Bayesian neural network models uncertainty to improve robustness, while symbolic reasoning incorporates domain knowledge through logic rules. The fusion of neural and symbolic outputs leads to improved recognition accuracy and greater transparency in decision-making.

Future work will explore: (1) incorporating temporal sequence modeling (e.g., LSTM or Transformer-based methods) to capture dynamic posture transitions and behavioral context; (2) adopting soft logic frameworks or learnable rule induction to enhance the flexibility and scalability of symbolic reasoning; (3) deploying the model in real-world classroom and study environments to assess system robustness and application value in behavior monitoring and intelligent intervention; and (4) addressing occlusion issues via multi-view sensing and cross-view information fusion to improve recognition in complex classroom settings.

Author Contributions: Conceptualization, S.W. and Y.L.; methodology, S.W., A.T. and Y.L.; software, S.W.; validation, Y.L., A.T., T.G. and C.L.; formal analysis, S.W.; investigation, S.W.; resources, S.W.; data curation, S.W. and Y.Z.; writing—original draft preparation, S.W.; writing—review and editing, Y.L., A.T., T.G. and C.L.; visualization, S.W.; upervision, A.T. and Y.L.; project administration, A.T. and Y.L.; funding acquisition, S.W. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (62372494), Guangdong Specialized Talent Training Program (2024001), Guangdong Engineering Centre (2024GCZX001), and Characteristic Innovation Project (Natural Sciences) of Guangdong Universities' Scientific Research Platform and Projects (2024KTSCX016).

Data Availability Statement: The original contributions presented in the study are included in the article; further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. de Rezende, L.F.M.; Rodrigues Lopes, M.; Rey-López, J.P.; Matsudo, V.K.R.; do Carmo Luiz, O. Sedentary behavior and health outcomes: An overview of systematic reviews. *PLoS ONE* **2014**, *9*, e105620. [CrossRef]
2. Casas, A.S.; Patiño, M.S.; Camargo, D.M. Association between the sitting posture and back pain in college students. *Rev. Univ. Ind. Santander Salud* **2016**, *48*, 446–454.
3. Falla, D.; Jull, G.; Russell, T.; Vicenzino, B.; Hodges, P. Effect of neck exercise on sitting posture in patients with chronic neck pain. *Phys. Ther.* **2007**, *87*, 408–417. [CrossRef]
4. Liu, Y.; Han, Z.; Chen, X.; Ru, S.; Yan, B. Effects of different sitting postures on back shape and hip pressure. *J. Med. Biomech.* **2023**, *38*, 756–762.
5. Vlaović, Z.; Jaković, M.; Domljan, D. Smart office chairs with sensors for detecting sitting positions and sitting habits: A review. *Drv. Ind.* **2022**, *73*, 227–243. [CrossRef]
6. Gupta, R.; Gupta, S.H.; Agarwal, A.; Choudhary, P.; Bansal, N.; Sen, S. A wearable multisensor posture detection system. In Proceedings of the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 13–15 May 2020; pp. 818–822.
7. Hu, Y.; Huang, T.; Zhang, H.; Lin, H.; Zhang, Y.; Ke, L.; Cao, W.; Hu, K.; Ding, Y.; Wang, X.; et al. Ultrasensitive and wearable carbon hybrid fiber devices as robust intelligent sensors. *ACS Appl. Mater. Interfaces* **2021**, *13*, 23905–23914. [CrossRef]
8. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703.
9. Wang, B.; Jin, X.; Yu, M.; Wang, G.; Chen, J. Pre-training Encoder-Decoder for Minority Language Speech Recognition. In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN), Yokohama, Japan, 30 June–5 July 2024; pp. 1–8.
10. Kamel, A.; Sheng, B.; Yang, P.; Li, P.; Shen, R.; Feng, D.D. Deep convolutional neural networks for human action recognition using depth maps and postures. *IEEE Trans. Syst. Man Cybern. Syst.* **2018**, *49*, 1806–1819. [CrossRef]
11. Li, X.; Xiong, H.; Li, X.; Wu, X.; Zhang, X.; Liu, J.; Bian, J.; Dou, D. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowl. Inf. Syst.* **2022**, *64*, 3197–3234. [CrossRef]
12. Dindorf, C.; Ludwig, O.; Simon, S.; Becker, S.; Fröhlich, M. Machine learning and explainable artificial intelligence using counterfactual explanations for evaluating posture parameters. *Bioengineering* **2023**, *10*, 511. [CrossRef] [PubMed]
13. Ding, Z.; Li, W.; Ogunbona, P.; Qin, L. A real-time webcam-based method for assessing upper-body postures. *Mach. Vis. Appl.* **2019**, *30*, 833–850. [CrossRef]
14. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 1050–1059.
15. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In Proceedings of the NIPS'17: 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
16. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* **2021**, *76*, 243–297. [CrossRef]
17. Hesamian, M.H.; Jia, W.; He, X.; Kennedy, P. Deep learning techniques for medical image segmentation: Achievements and challenges. *J. Digit. Imaging* **2019**, *32*, 582–596. [CrossRef]
18. Besold, T.R.; Bader, S.; Bowman, H.; Domingos, P.; Hitzler, P.; Kühnberger, K.U.; Lamb, L.C.; Lima, P.M.V.; de Penning, L.; Pinkas, G.; et al. Neural-symbolic learning and reasoning: A survey and interpretation 1. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*; IOS Press: Amsterdam, The Netherlands, 2021; pp. 1–51.
19. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.
20. Andriluka, M.; Pishchulin, L.; Gehler, P.; Schiele, B. 2d human pose estimation: New benchmark and state of the art analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 3686–3693.

21. Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.E.; Sheikh, Y. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 172–186. [CrossRef]
22. Rhodes, J.S.; Cutler, A.; Moon, K.R. Geometry-and accuracy-preserving random forest proximities. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10947–10959. [CrossRef] [PubMed]
23. Zavala-Mondragon, L.A.; Lamichhane, B.; Zhang, L.; Haan, G.d. CNN-SkelPose: A CNN-based skeleton estimation algorithm for clinical applications. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 2369–2380. [CrossRef]
24. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
25. Li, L.; Yang, G.; Li, Y.; Zhu, D.; He, L. Abnormal sitting posture recognition based on multi-scale spatiotemporal features of skeleton graph. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106374. [CrossRef]
26. Cao, Z.; Wu, X.; Wu, C.; Jiao, S.; Xiao, Y.; Zhang, Y.; Zhou, Y. KeypointNet: An Efficient Deep Learning Model with Multi-View Recognition Capability for Sitting Posture Recognition. *Electronics* **2025**, *14*, 718. [CrossRef]
27. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
28. Chen, J.; Yang, L.; Zhang, Y.; Alber, M.; Chen, D.Z. Combining fully convolutional and recurrent neural networks for 3d biomedical image segmentation. In Proceedings of the NIPS'16: 30th International Conference on Neural Information Processing Systems, Barcelona, Spain, 5–10 December 2016; Volume 29.
29. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
30. Wanyan, Y.; Yang, X.; Dong, W.; Xu, C. A comprehensive review of few-shot action recognition. *arXiv* **2024**, arXiv:2407.14744. [CrossRef]
31. Chen, C.; Liang, J.; Zhu, X. Gait recognition based on improved dynamic Bayesian networks. *Pattern Recognit.* **2011**, *44*, 988–995. [CrossRef]
32. Hitzler, P.; Eberhart, A.; Ebrahimi, M.; Sarker, M.K.; Zhou, L. Neuro-symbolic approaches in artificial intelligence. *Natl. Sci. Rev.* **2022**, *9*, nwac035. [CrossRef]
33. Qu, M.; Tang, J. Probabilistic logic neural networks for reasoning. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
34. Badreddine, S.; Garcez, A.D.; Serafini, L.; Spranger, M. Logic tensor networks. *Artif. Intell.* **2022**, *303*, 103649. [CrossRef]
35. Johnston, P.; Nogueira, K.; Swingler, K. NS-IL: Neuro-symbolic visual question answering using incrementally learnt, independent probabilistic models for small sample sizes. *IEEE Access* **2023**, *11*, 141406–141420. [CrossRef]
36. Magherini, T.; Fantechi, A.; Nugent, C.D.; Vicario, E. Using temporal logic and model checking in automated recognition of human activities for ambient-assisted living. *IEEE Trans. Hum.-Mach. Syst.* **2013**, *43*, 509–521. [CrossRef]
37. Tang, J.; Wang, Z.; Hao, G.; Wang, K.; Zhang, Y.; Wang, N.; Liang, D. SAE-PPL: Self-guided attention encoder with prior knowledge-guided pseudo labels for weakly supervised video anomaly detection. *J. Vis. Commun. Image Represent.* **2023**, *97*, 103967. [CrossRef]
38. Ye, Y.; Shi, S.; Zhao, T.; Qiu, K.; Lan, T. Patches Channel Attention for Human Sitting Posture Recognition. In Proceedings of the International Conference on Artificial Neural Networks, Heraklion, Crete, Greece, 26–29 September 2023; Springer: Cham, Switzerland, 2023; pp. 358–370.
39. Groenesteijn, L.; Ellegast, R.P.; Keller, K.; Krause, F.; Berger, H.; de Looze, M.P. Office task effects on comfort and body dynamics in five dynamic office chairs. *Appl. Ergon.* **2012**, *43*, 320–328. [CrossRef]
40. Abdullah, S.; Ahmed, S.; Choi, C.; Cho, S.H. Distance and Angle Insensitive Radar-Based Multi-Human Posture Recognition Using Deep Learning. *Sensors* **2024**, *24*, 7250. [CrossRef]
41. Atvar, A.; Cinbiş, N.İ. Classification of human poses and orientations with deep learning. In Proceedings of the 2018 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2–5 May 2018; pp. 1–4.
42. Ko, K.R.; Chae, S.H.; Moon, D.; Seo, C.H.; Pan, S.B. Four-joint motion data based posture classification for immersive postural correction system. *Multimed. Tools Appl.* **2017**, *76*, 11235–11249. [CrossRef]
43. Zeng, X.; Sun, B.; Wang, E.; Luo, W.; Liu, T. A Method of Learner's Sitting Posture Recognition Based on Depth Image. In Proceedings of the 2017 2nd International Conference on Control, Automation and Artificial Intelligence (CAAI 2017), Sanya, China, 25–26 June 2017; Atlantis Press: Dordrecht, The Netherlands, 2017; pp. 558–563.
44. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
45. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
46. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. *arXiv* **2017**, arXiv:1711.05101.

47. Zhao, S.; Su, Y. Sitting Posture Recognition Based on the Computer's Camera. In Proceedings of the 2024 2nd Asia Conference on Computer Vision, Image Processing and Pattern Recognition, Xiamen, China, 26–28 April 2024; pp. 1–5.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
49. Tan, M.; Le, Q. Efficientnetv2: Smaller models and faster training. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 10096–10106.
50. Mehta, S.; Rastegari, M. Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. *arXiv* **2021**, arXiv:2110.02178.
51. Jiao, S.; Xiao, Y.; Wu, X.; Liang, Y.; Liang, Y.; Zhou, Y. LMSPNet: Improved lightweight network for multi-person sitting posture recognition. In Proceedings of the 2023 IEEE 3rd International Conference on Computer Communication and Artificial Intelligence (CCAI), Taiyuan, China, 26–28 May 2023; pp. 289–295.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.