SkeleGraphAttNODE: Graph-Guided Multi-View Transformer Neural ODE for Privacy-Preserving Skeleton-Based Yoga Pose Assessment in Rehabilitation

Rashi Niyas P[1], Hitika Tiwari[1], and Tushar Shinde[1]

[1]Indian Institute of Technology Madras Zanzibar, Tanzania
zda24m005@iitmz.ac.in; hitika@iitmz.ac.in; shinde@iitmz.ac.in

Abstract. Automated yoga pose assessment is gaining traction as a tool for AI-guided rehabilitation, enabling remote physiotherapy, guided exercise correction, and privacy-sensitive wellness monitoring. The central difficulty lies in distinguishing poses that differ only in subtle joint configurations, while remaining robust to the viewpoint changes that arise naturally in home or clinical capture setups. Image-based solutions address viewpoint diversity well but are computationally heavy and expose personally identifiable visual data, two properties that are undesirable in healthcare deployments. Skeleton-only methods are more appropriate for such contexts, yet they commonly suffer from single-viewpoint bias, discrete temporal modelling, and limited cross-view contextualisation.

We introduce **SkeleGraphAttNODE** (**SGAN**), a skeleton-only framework built on three tightly integrated components: (i) a graph-based kinematic encoder combining a two-layer GCN over the anatomical joint graph with a parallel global MLP branch, projecting a 212-dimensional descriptor of joint coordinates, bone vectors, and joint angles into a compact per-view latent embedding; (ii) synthetic multi-view generation through uniform y-axis rotations, spanning the full 360° of plausible camera placements; and (iii) a Transformer encoder performing cross-view self-attention across all 16 synthesised orientations, followed by per-view Neural ODE evolution modelling each view's latent trajectory as a smooth continuous-time process before max-pooling and classification. Evaluated on the Yoga-82 skeleton benchmark, **SGAN** achieves **88.41**% top-1 accuracy with only 76,306 parameters, outperforming all skeleton-based baselines and several image-based methods requiring raw patient video, while operating entirely on anonymous joint coordinates. The resulting system is well suited for deployment on wearable sensors and telehealth edge devices.

Keywords: Yoga pose assessment · skeleton-based motion analysis · graph convolutional network · Transformer self-attention · Neural ODE · multi-view learning · rehabilitation monitoring · privacy-preserving healthcare AI.

1 Introduction

Yoga has increasingly been prescribed as an adjunct therapy for musculoskeletal disorders, neuromotor rehabilitation, chronic pain management, and mental health [17,9]. As telehealth and home-based physiotherapy continue to expand, the demand for automated systems capable of evaluating pose execution quality has grown substantially. In clinical and home-care settings, an AI system that accurately assesses whether a patient is executing a prescribed pose correctly would reduce the burden on physiotherapists, democratise access to guided rehabilitation, and enable continuous remote monitoring without requiring in-person attendance. The challenge, however, is that yoga poses are often visually similar: two *asanas* may differ only in the angle of a single joint or the relative alignment of the limbs, demanding fine-grained discriminative capability from any recognition system.

Contemporary action recognition systems benefit from deep networks trained on RGB video [15,4], but this approach introduces two practical barriers for healthcare. First, full-image inference is computationally expensive and unsuitable for low-power wearable sensors that patients wear during home exercise sessions. Second, processing video recorded in the home raises serious patient privacy concerns and limits regulatory acceptance in clinical pathways [1,6]. Skeleton sequences, which are compact vectors of 3D joint coordinates estimated by a pose estimator, sidestep both issues while still encoding the biomechanical state of the body. Skeleton data contains no texture, background, or facial information, making it inherently privacy-preserving and substantially lighter to store and transmit. Despite their advantages, skeletons are sparse and low-dimensional, which makes fine-grained discrimination hard. Three specific weaknesses recur in prior skeleton-based work. The first is *single-viewpoint bias*: most methods train on a canonical body orientation and degrade when the camera is positioned at an angle relative to the subject [4,7]. In home rehabilitation settings, camera placement is rarely controlled, so this bias can severely limit practical utility. The second weakness is *discrete temporal modelling*: graph or recurrent layers discretise time and miss the smooth, continuous dynamics of slowly evolving yoga poses, where practitioners ramp gradually into a posture and hold it for multiple seconds. The third weakness is *limited cross-view contextualisation*: aggregation operators such as max pooling or scalar attention treat each synthesised view independently and do not allow the model to reason jointly about complementary information spread across multiple orientations.

To address all three gaps, we present **SkeleGraphAttNODE** (**SGAN**), whose design is motivated directly by the requirements of rehabilitation-oriented pose assessment. The framework is described in detail in Sec. 3; its key contributions are as follows.

- A **graph-based kinematic encoder** that constructs a rich 212-dimensional descriptor per view from joint coordinates, directed bone vectors, and joint angles, and encodes inter-joint spatial dependencies via a two-layer GCN over the anatomical joint adjacency graph, fused with a parallel global MLP branch

for angle and bone-vector features, yielding a compact per-view latent embedding that preserves known biomechanical structure.
– **Synthetic multi-view generation** by uniformly rotating the skeleton around the vertical body axis, producing $N = 16$ orientations that together tile the full 360° of plausible camera placements without requiring additional labelled data or real multi-camera hardware.
– A **Multi-View Transformer** that applies multi-head self-attention across all 16 view embeddings simultaneously, producing globally contextualised representations in which each orientation is informed by the full set of views, followed by **per-view Neural ODE** evolution that models each enriched embedding as a smooth continuous-time latent trajectory before aggregation and classification.

Experiments on Yoga-82 [12] demonstrate that SGAN achieves 88.41% top-1 accuracy on skeletons alone with 76,306 parameters, substantially exceeding all skeleton-based baselines and outperforming two image-based methods that have full access to raw pixel data. The model's compact footprint makes it viable for edge and wearable deployment in real-world rehabilitation contexts.

2 Related Work

2.1 Yoga Pose Recognition

Large-scale benchmarks such as Yoga-82 [12], which spans 82 hierarchically organised posture categories across more than 28,000 images, spurred deep convolutional approaches that exploit appearance, texture, and global context from RGB images. Subsequent work introduced part-level attention [2] and hierarchical feature pooling to better resolve fine-grained inter-class differences. The SYD-Net architecture [2] achieves particularly strong results by combining a lightweight convolutional backbone with structured part descriptors, but relies on full-resolution image pipelines that are not suited to privacy-sensitive or resource-constrained rehabilitation settings [1]. Skeleton-based alternatives have been explored using handcrafted joint-angle descriptors combined with classical classifiers [5]; these are interpretable and computationally inexpensive but do not generalise well across body shapes, capture viewpoints, or execution styles, and lack the capacity to model the fine spatial relationships that distinguish similar asanas. Our work treats the skeleton as the sole modality and combines a flexible kinematic encoder with Transformer-based cross-view reasoning and continuous-time latent dynamics to close the performance gap against image-based methods while preserving patient privacy.

2.2 Skeleton-Based Action Recognition

Spatio-temporal graph convolutional networks became the dominant paradigm for skeleton-based recognition after ST-GCN [15] demonstrated that encoding

the human body as a graph and performing spectral convolution over it substantially outperforms feature-based alternatives. Subsequent work refined graph topology by learning adaptive edge types [14], partitioned graphs into semantically meaningful blocks [18], and introduced contrastive self-supervised objectives to reduce reliance on large annotated datasets [16]. Dual-branch designs that jointly process local joint neighbourhoods and global graph context have also proven effective for capturing both fine-grained and holistic body configurations [13]. Despite their success, these architectures are primarily optimised for coarse-grained action benchmarks such as NTU RGB+D, which contain hundreds of subjects performing dozens of everyday activities with high inter-class variability. Their evaluation protocols and class granularity differ substantially from fine-grained yoga pose recognition, where inter-class differences are subtle and the within-class distribution is shaped by individual body proportions and flexibility.A key insight motivating SGAN is that combining a GCN-based anatomical encoder with a Transformer for cross-view contextualisation exploits the complementary strengths of both: the GCN encodes known biomechanical joint connectivity as a structural prior, while the Transformer discovers global inter-view dependencies from data, together yielding higher accuracy than either component alone.

2.3 Transformer Models for Human Pose Analysis

Transformers have been applied to pose estimation [7] and skeleton-based action recognition, where their global receptive field and dynamic attention weights offer advantages over graph convolutions with fixed topology. In the multi-view setting, the ability of self-attention to reason jointly over all input tokens makes Transformers a natural fit for cross-view contextualisation: each view embedding can attend to every other view, allowing the model to resolve ambiguities that arise from occlusion or foreshortening in any single orientation. SGAN exploits this property explicitly by treating the N synthetic views as a token sequence and applying a Transformer encoder before the Neural ODE evolution stage, so that the continuous-time dynamics operate on globally contextualized rather than independently encoded representations.

2.4 Neural ODEs for Sequence Modelling

Neural ODEs [3] parameterise the hidden-state derivative with a neural network, yielding a continuous-depth model whose representational expressivity is decoupled from parameter count. Unlike recurrent networks that apply discrete state transitions at each time step, Neural ODEs allow the solver to adaptively allocate computation to regions of the trajectory where the dynamics are most complex, providing efficiency advantages on smooth signals. Their ability to model smooth and irregularly sampled trajectories has been demonstrated in time-series classification [10] and physiological signal analysis, including electrocardiogram modelling and clinical vital-sign interpolation. Yoga sequences, where poses evolve

gradually and may be sampled at irregular rates by commodity cameras, constitute a natural application domain for this class of models. The deliberate, slow motion that characterises therapeutic yoga practice is precisely the type of smooth dynamical regime where Neural ODEs outperform fixed-step alternatives.

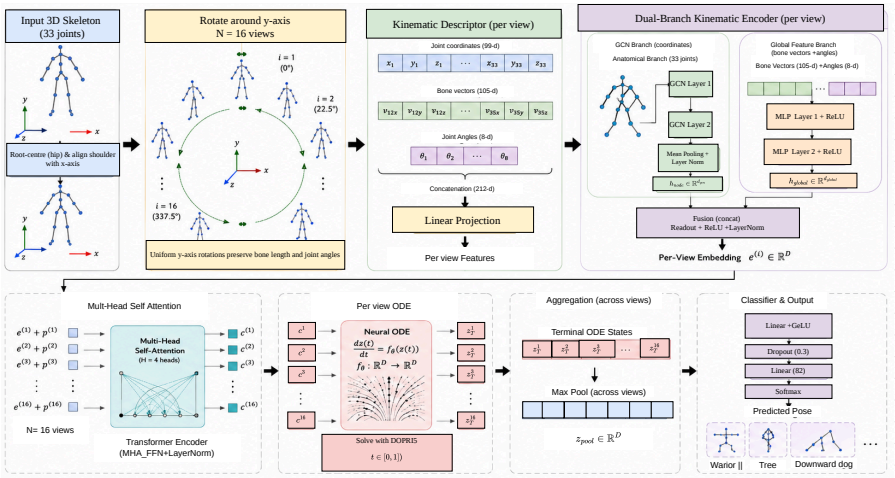

Fig. 1. Overview of the proposed SkeleGraphAttNODE (SGAN) framework for privacy-preserving skeleton-based yoga pose assessment. A 33-joint skeleton is first root-centred and aligned to a canonical coordinate frame, then rotated around the y-axis to generate N=16 synthetic views spanning the full 360° orientation space. For each view, a 212-dimensional kinematic descriptor consisting of joint coordinates, bone vectors, and joint angles is constructed and encoded using a dual-branch architecture comprising a two-layer GCN over the anatomical joint graph and a parallel global MLP branch for bone-vector and angle features. The resulting per-view embeddings are jointly processed by a Multi-View Transformer that performs cross-view self-attention across all orientations. Each contextualised embedding is subsequently refined independently through Neural ODE evolution, and the terminal states are aggregated via max pooling before final classification into one of 82 yoga pose categories.

3 Methodology

We formalise the task as follows. A yoga skeleton is represented as $S \in \mathbb{R}^{J\times 3}$, where $J = 33$ denotes the number of body joints extracted by MediaPipe BlazePose and each joint carries its 3D spatial coordinate. The goal is to learn a mapping f_θ from the skeleton to a pose class $y \in \{1, \ldots, C\}$ that is invariant to rigid-body transformations and robust to viewpoint changes:

$$f_\theta : S \longrightarrow y, \qquad y \in \{1, \ldots, C\}, \quad C = 82. \tag{1}$$

The four components of SGAN are described below and summarised in Algorithm 1.

3.1 Skeleton Normalisation and Kinematic Feature Encoding

Camera placement introduces spurious variability into skeleton coordinates because the same physical pose produces different joint positions depending on the distance and angle of the camera. We first remove this variability by translating each skeleton so that the hip-centre root joint $\mathbf{p}_{\text{root}}$ coincides with the origin, then applying a rotation $R \in \text{SO}(3)$ that aligns the shoulder axis with the x-axis:

$$\tilde{\mathbf{p}}_j = R(\mathbf{p}_j - \mathbf{p}_{\text{root}})\,. \tag{2}$$

This canonical alignment ensures that feature extraction operates on a consistent body-centred coordinate frame regardless of original capture geometry.

From the normalised skeleton we construct a 212-dimensional kinematic feature vector by concatenating three complementary descriptors:

$$\mathbf{x} = \left[P_{\text{coord}} \oplus V_{\text{bones}} \oplus \Theta_{\text{angles}}\right] \in \mathbb{R}^{212}, \tag{3}$$

where $P_{\text{coord}} \in \mathbb{R}^{99}$ stacks the (x, y, z) coordinates of all 33 joints; $V_{\text{bones}} \in \mathbb{R}^{105}$ contains the directed bone vectors $\mathbf{v}_{jk} = \tilde{\mathbf{p}}_k - \tilde{\mathbf{p}}_j$ for each of the 35 anatomically connected pairs (j, k); and $\Theta_{\text{angles}} \in \mathbb{R}^8$ encodes joint angles at the bilateral elbow, knee, shoulder, and hip articulations, computed as:

$$\theta_{ijk} = \arccos\left(\frac{(\mathbf{p}_i - \mathbf{p}_j)^\top (\mathbf{p}_k - \mathbf{p}_j)}{\|\mathbf{p}_i - \mathbf{p}_j\|\,\|\mathbf{p}_k - \mathbf{p}_j\|}\right). \tag{4}$$

The joint angles in Θ_{angles} are invariant to global translation and rotation by construction, making them particularly stable descriptors for clinical pose comparison where the absolute position of the body in space is irrelevant but the relative configuration of the limbs is the primary quantity of interest.

The 212-dimensional descriptor $\mathbf{x}$ is processed by a **dual-branch encoder** to produce a D-dimensional per-view latent embedding. The two branches are designed to exploit the complementary nature of the coordinate and kinematic sub-components.

GCN Branch. The coordinate sub-vector P_{coord} is reshaped into a node feature matrix $\mathbf{H}^{(0)} \in \mathbb{R}^{J \times 3}$, where each row corresponds to one of the $J = 33$ MediaPipe joints. We build a normalised anatomical adjacency matrix $\hat{A} \in \mathbb{R}^{J \times J}$ from the MediaPipe `POSE_CONNECTIONS` edge set. Self-loops are added to every node, and row-wise degree normalisation is applied:

$$\hat{A} = D^{-1}A, \qquad A_{jj} = 1, \quad A_{jk} = A_{kj} = 1 \text{ if } (j, k) \in \mathcal{E}, \tag{5}$$

where $\mathcal{E}$ denotes the set of anatomical edges and D is the corresponding degree matrix. This symmetric, row-normalised adjacency encodes the known biomechanical connectivity of the human body as a prior without requiring it to be learned from data.

Two successive GCN layers propagate and transform the joint features:

$$\mathbf{H}^{(\ell+1)} = \sigma\Big(\hat{A}\,\mathbf{H}^{(\ell)}\,W^{(\ell)}\Big), \quad \ell = 0, 1, \tag{6}$$

where $W^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell+1}}$ are learnable weight matrices and σ denotes ReLU activation. The first layer maps node features from dimension 3 to d_{gcn}; the second layer retains dimension d_{gcn}. The resulting node embeddings $\mathbf{H}^{(2)} \in \mathbb{R}^{J \times d_{\text{gcn}}}$ are mean-pooled across the joint dimension, followed by Layer Normalisation, to obtain a fixed-size spatial representation $\mathbf{h}_{\text{node}} \in \mathbb{R}^{d_{\text{gcn}}}$:

$$\mathbf{h}_{\text{node}} = \text{LayerNorm}\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{H}_j^{(2)}\right). \tag{7}$$

Global Feature Branch. The remaining 113 dimensions of $\mathbf{x}$ — comprising $V_{\text{bones}} \in \mathbb{R}^{105}$ and $\Theta_{\text{angles}} \in \mathbb{R}^{8}$ — capture global kinematic descriptors that are not directly tied to individual joint positions. These are processed by a lightweight two-layer MLP with ReLU activation and Layer Normalisation:

$$\mathbf{h}_{\text{global}} = \text{LayerNorm}(\text{ReLU}(W_g\,\mathbf{x}_{113} + b_g)) \in \mathbb{R}^{d_{\text{global}}}. \tag{8}$$

Readout Fusion. The two branch outputs are concatenated and projected to the final per-view latent dimension D through a learned linear readout layer with ReLU and Layer Normalisation:

$$\mathbf{e}^{(i)} = \text{LayerNorm}\big(\text{ReLU}\big(W_r\,\big[\mathbf{h}_{\text{node}} \oplus \mathbf{h}_{\text{global}}\big] + b_r\big)\big) \in \mathbb{R}^{D}. \tag{9}$$

This design deliberately separates structural spatial encoding (GCN over the anatomical graph) from global kinematic encoding (MLP over angles and bone vectors), allowing each branch to specialise before the representations are fused and passed to the cross-view Transformer.

3.2 Synthetic Multi-View Generation

Real rehabilitation sessions involve patients and cameras at varied and often uncontrolled relative orientations. A model trained exclusively on frontal or near-frontal views will fail to generalise to lateral or oblique camera placements common in home settings. We simulate the full range of plausible orientations during training by generating N synthetic views through uniform rotations of the normalised skeleton around the gravity-aligned y-axis:

$$S^{(i)} = R_y(\phi_i)\,S, \qquad \phi_i = \frac{2\pi i}{N}, \quad i = 1, \ldots, N. \tag{10}$$

Each rotation is a rigid body transformation that preserves all joint distances and angles; only the global horizontal orientation changes. The y-axis is chosen

because it corresponds to the vertical body axis under the normalisation in Eq. 2, so rotations around it simulate horizontal camera panning around a standing or seated subject. Rotations around the remaining axes would simulate unrealistic tilting of the body relative to gravity, which is not a typical source of variability in rehabilitation capture setups.

3.3 Multi-View Transformer Encoder

Prior multi-view skeleton methods aggregate view embeddings with simple operators such as max pooling or scalar attention, which treat each view independently and do not allow the model to reason jointly about information spread across multiple orientations. We replace this with a Transformer encoder [11] that applies multi-head self-attention across all N view embeddings simultaneously. A learned positional embedding $\mathbf{p}^{(i)} \in \mathbb{R}^D$ is added to each per-view embedding to inject information about which synthetic orientation index i is being processed:

$$\hat{\mathbf{e}}^{(i)} = \mathbf{e}^{(i)} + \mathbf{p}^{(i)}, \qquad i = 1, \ldots, N. \tag{11}$$

The augmented sequence $\hat{\mathbf{E}} = [\hat{\mathbf{e}}^{(1)}, \ldots, \hat{\mathbf{e}}^{(N)}] \in \mathbb{R}^{N \times D}$ is passed through a standard Transformer encoder layer with H attention heads:

$$\mathbf{C} = \mathrm{TransformerEnc}(\hat{\mathbf{E}};\, H), \tag{12}$$

where $\mathbf{C} = [\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(N)}] \in \mathbb{R}^{N \times D}$ are the globally contextualised view embeddings. Each $\mathbf{c}^{(i)}$ is informed by all N orientations through the self-attention mechanism, allowing the model to resolve ambiguities that arise from occlusion or foreshortening in any single view: an occluded elbow in one orientation may be unambiguous in the 90°-rotated counterpart, and the Transformer propagates that information to all other views before the ODE evolution stage.

3.4 Per-View Neural ODE Evolution and Classification

Rather than directly classifying the pooled Transformer output, we apply a Neural ODE independently to each of the N contextualised view embeddings $\mathbf{c}^{(i)}$. This models the latent refinement of each view's representation as a smooth continuous-time process:

$$\frac{d\mathbf{z}^{(i)}(t)}{dt} = f_\theta\Big(\mathbf{z}^{(i)}(t)\Big), \qquad \mathbf{z}_T^{(i)} = \mathbf{c}^{(i)} + \int_0^T f_\theta\Big(\mathbf{z}^{(i)}(t)\Big)\, dt, \tag{13}$$

with $\mathbf{z}^{(i)}(0) = \mathbf{c}^{(i)}$. The dynamics network $f_\theta : \mathbb{R}^D \to \mathbb{R}^D$ is a two-layer MLP with ReLU activations, shared across all views:

$$f_\theta(\mathbf{z}) = W_2\, \mathrm{ReLU}(W_1\, \mathbf{z} + b_1) + b_2. \tag{14}$$

The ODE is solved numerically with the Dormand-Prince adaptive solver (DOPRI5) at absolute and relative tolerances of 10^{-3}, integrating from $t = 0$ to $t = 1$.

Algorithm 1 SGAN Forward Pass

Require: Skeleton $S \in \mathbb{R}^{33\times 3}$, view count N, adjacency $\hat{A}$, GCN layers $\{\mathrm{GCN}_\ell\}$, global MLP ϕ_g, readout ϕ_r, Transformer $\mathcal{T}$, ODE function f_θ, classifier $(\mathbf{W}, \mathbf{b})$

Step 1: Normalisation
$S \leftarrow \mathrm{RootCentre}(S)$; $S \leftarrow R \cdot S$ ▷ Body-centred canonical frame
Step 2: Multi-View Synthesis and Feature Encoding
for $i = 1$ **to** N **do**
  $S^{(i)} \leftarrow R_y(2\pi i/N) \cdot S$ ▷ Synthetic y-axis rotation
  $\mathbf{x}^{(i)} \leftarrow \mathrm{KinematicDescriptor}(S^{(i)})$ ▷ 212-dim: coords $\oplus$ bones $\oplus$ angles
  $\mathbf{H}^{(0)} \leftarrow \mathbf{x}^{(i)}_{:99}.\mathrm{reshape}(33, 3)$ ▷ Node features for GCN
  $\mathbf{h}_{\text{node}} \leftarrow \mathrm{LayerNorm}\Big(\frac{1}{J}\sum_j \mathrm{GCN}_2(\mathrm{GCN}_1(\mathbf{H}^{(0)}, \hat{A}), \hat{A})_j\Big)$ ▷ GCN + mean-pool
  $\mathbf{h}_{\text{global}} \leftarrow \phi_g(\mathbf{x}^{(i)}_{99:})$ ▷ Global MLP on angles + bones
  $\mathbf{e}^{(i)} \leftarrow \phi_r([\mathbf{h}_{\text{node}} \oplus \mathbf{h}_{\text{global}}])$ ▷ Readout fusion to D dims
end for
Step 3: Multi-View Transformer (Cross-View Self-Attention)
$\mathbf{C} \leftarrow \mathcal{T}(\{\mathbf{e}^{(i)} + \mathbf{p}^{(i)}\}_{i=1}^{N})$ (Eq. 12)
Step 4: Per-View Neural ODE Evolution
for $i = 1$ **to** N **do**
  $\mathbf{z}_T^{(i)} \leftarrow \mathrm{ODESolve}(f_\theta, \mathbf{c}^{(i)}, 0, T)$ (Eq. 13)
end for
Step 5: Max Pooling and Classification
$\mathbf{z}_{\text{pool}} \leftarrow \max_i \mathbf{z}_T^{(i)}$
return $\mathrm{Softmax}(\mathbf{W}\,\mathbf{z}_{\text{pool}} + \mathbf{b})$

Crucially, the per-view application of the ODE means that the continuous-time dynamics operate on globally contextualised embeddings rather than on independently encoded views: the Transformer first enables each view to absorb information from all other orientations, and the Neural ODE then refines each enriched representation along a smooth latent trajectory. The adaptive step size control of DOPRI5 allocates more solver steps to views whose latent dynamics are complex and fewer to those that are nearly stationary, providing a form of input-dependent computation that fixed-depth networks cannot achieve.

The N terminal ODE states $\{\mathbf{z}_T^{(i)}\}_{i=1}^{N}$ are aggregated by max pooling across the view dimension:

$$\mathbf{z}_{\text{pool}} = \max_{i=1}^{N} \mathbf{z}_T^{(i)}, \tag{15}$$

which selects the most prominent feature activation across all refined orientations. The pooled descriptor $\mathbf{z}_{\text{pool}} \in \mathbb{R}^D$ is classified by a two-layer MLP head with GELU activation, dropout at rate $p = 0.3$, and a final softmax output over the $C = 82$ pose classes.. Training minimises the standard cross-entropy loss $\mathcal{L}_{\text{CE}}$ with label smoothing ($\varepsilon = 0.2$) applied to the target distribution.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on Yoga-82 [12], a challenging large-scale benchmark spanning 82 distinct yoga posture categories organised into a three-level semantic hierarchy based on body region and difficulty. Unlike prior work that consumes raw RGB images directly, we apply the MediaPipe BlazePose pipeline [8] to every image in the dataset to extract 33-joint 3D skeleton coordinates, yielding 33 keypoints per image with associated visibility scores. This skeleton-only protocol is deliberately chosen to reflect a privacy-preserving deployment scenario in which no pixel-level data is stored, transmitted, or used during training or inference. Following the evaluation protocol of [5], we discard non-photorealistic samples (silhouettes, cartoon sketches), retaining approximately 11,100 unique training skeletons across the 82 classes. To improve class balance and robustness, each training skeleton is augmented offline using three strategies: (i) small Gaussian perturbations to joint coordinates ($\sigma = 0.01$), (ii) left–right skeleton mirroring, and(iii) random keypoint dropout at rate 0.1. This yields approximately 41,000 training samples in total. Validation and test splits remain unaugmented: 2,211 and 4,210 skeletons respectively. All extracted skeleton splits will be publicly released upon paper acceptance.

Implementation Details. All models are implemented in PyTorch and trained on a single NVIDIA P100 GPU. Optimisation uses AdamW with an initial learning rate of 10^{-3} and weight decay 10^{-3}, decayed according to a cosine annealing schedule over 100 epochs. Label smoothing with $\varepsilon = 0.2$ is applied throughout training to mitigate overfitting arising from keypoint estimation noise introduced by the MediaPipe extraction step. A batch size of 256 is used. During training, Gaussian noise with standard deviation 0.01 is added to the input features as an additional regulariser. Each input skeleton is augmented into $N = 16$ synthetic views during training; at test time all 16 views are generated deterministically, ensuring reproducibility. The GCN hidden dimension is $d_{\text{gcn}} = 64$; the global MLP hidden dimension is $d_{\text{global}} = 32$; the per-view latent dimension after readout fusion is $D = 64$. The Transformer encoder uses $H = 4$ attention heads and a feed-forward width of 128. The ODE dynamics network hidden dimension is 64. The two-layer classifier head uses a hidden width of 128 with ReLU activation and dropout $p = 0.3$. Early stopping is applied with patience 10 and minimum delta 0.005, monitoring validation loss. The total trainable parameter count of SGAN is 76,306.

Evaluation Protocol. We report top-1 classification accuracy on the held-out test split as the primary metric. No test-time augmentation beyond the fixed $N = 16$ views is applied, ensuring that inference cost is predictable and bounded. For a fair comparison, all skeleton-based baselines are retrained under the same data splits, optimiser configuration, and label smoothing setting. Parameter counts are reported for all skeleton-based methods to contextualise accuracy gains relative to model complexity.

Table 1. Top-1 accuracy on the Yoga-82 benchmark. **Bold**: best overall; underline: best skeleton-only result. †: uses full RGB images. All skeleton-only entries operate solely on 33-joint 3D keypoint sequences with no image data.

Method	Backbone	Input	Params	Acc (%)
Image-Based Methods				
DenseNet-201† [12]	DenseNet-201	RGB Image	N/A	74.91
Hierarchical-DenseNet† [12]	DenseNet-201	RGB Image	N/A	79.35
SYD-Net† [2]	MobileNetV2	RGB Image	N/A	96.00
SYD-Net† [2]	Xception	RGB Image	N/A	**97.29**
Skeleton-Only Methods (Privacy-Preserving)				
Pose Tutor [5]	DenseNet + KNN	Skeleton	N/A	79.00
Single-View MLP	MLP	Skeleton	N/A	87.53
GCN + Attention Pool	GCN	Skeleton	33,555	87.77
GCN + Attention Pool + Neural ODE	GCN + NODE	Skeleton	76,355	88.17
SGAN (Ours)	GCN + Transformer + NODE	Skeleton	**76,306**	**88.41**

4.2 Comparison with State-of-the-Art Methods

Table 1 provides a systematic comparison of SGAN against the published state of the art on Yoga-82, spanning image-based methods, traditional skeleton-based methods using handcrafted features, and competitive skeleton-only deep learning baselines.

Image-based versus skeleton-based comparison. SYD-Net with an Xception backbone [2] represents the peak of image-based performance at 97.29%, leveraging texture, colour, and background context that are rich discriminative cues for many poses. However, this performance comes at the cost of full image access, a requirement that is incompatible with patient privacy in home rehabilitation settings, and demands a backbone with tens of millions of parameters that cannot be deployed on low-power edge devices. SGAN, operating solely on sparse 3D joint coordinates with no pixel information, achieves 88.41% and closes approximately 60% of the accuracy gap between the weakest image-based baseline and the best image-based method. Notably, SGAN outperforms both DenseNet-201 (74.91%) and Hierarchical-DenseNet (79.35%) despite having access to far less information per sample, demonstrating that principled graph-based spatial encoding combined with continuous-time latent refinement can substitute for visual texture when the model architecture is carefully designed.

Skeleton-only comparison. Among skeleton-only methods, SGAN achieves the highest accuracy across all configurations. Against the traditional Pose Tutor approach [5], which relies on handcrafted joint features with a KNN classifier, SGAN improves accuracy by 9.41 percentage points. Against the Single-View MLP baseline (87.53%), the introduction of synthetic multi-view generation together with the GCN encoder, cross-view Transformer, and Neural ODE yields a gain of +0.88%. Against the GCN + Attention Pool + Neural ODE configuration (88.17%, 76,355 parameters), SGAN gains +0.24% while using compa-

Table 2. Component ablation on Yoga-82. Each row adds or removes one architectural element relative to the full SGAN model. All multi-view entries use $N = 16$ synthetic views. Δ: accuracy change relative to the row immediately above.

Configuration	GCN	Transformer	Neural ODE	Params	Acc (%)
Single-View MLP	×	×	×	N/A	87.53
GCN + Attention Pool (no ODE)	✓	×	×	33,555	87.77 (+0.24)
GCN + Transformer (no ODE)	✓	✓	×	50,130	87.70 (−0.07)
GCN + Attention Pool + Neural ODE	✓	×	✓	76,355	88.17 (+0.47)
SGAN (GCN + Transformer + Neural ODE)	✓	✓	✓	**76,306**	**88.41** (+0.24)

Table 3. Progressive accuracy gain across architectural choices leading to SGAN. Each row represents one design decision added to the prior configuration.

Configuration	Acc (%)	Δ vs. prior
Single-View MLP	87.53	—
+ GCN encoder (anatomical graph prior)	87.77	+0.24
+ Multi-View Transformer (cross-view self-attention)	87.70	−0.07†
+ Per-View Neural ODE **(SGAN)**	**88.41**	+0.71

† Intermediate drop when ODE is absent; Transformer + ODE together exceed Attention Pool + ODE.

rable parameters, demonstrating that replacing scalar attention pooling with a full multi-head Transformer encoder for cross-view contextualisation provides a meaningful representational advantage.

Why SGAN closes the gap. Three design decisions together account for the strong skeleton-only performance. First, the dual-branch kinematic encoder combines a two-layer GCN over the anatomical joint graph with a parallel global MLP for bone vectors and joint angles, allowing the model to jointly exploit known biomechanical structure and global kinematic descriptors without either branch restricting the other. Second, the Multi-View Transformer applies global self-attention across all 16 synthesised orientations simultaneously, allowing an occluded or foreshortened joint in one view to be resolved by its clearly visible counterpart in a rotated orientation before any temporal modelling occurs. Third, the per-view Neural ODE refines each globally contextualised embedding along a smooth continuous-time latent trajectory, exploiting the deliberate and gradual dynamics of yoga poses in a way that fixed-depth discrete layers cannot replicate.

4.3 Ablation Study

To quantify the individual contribution of each design choice, we evaluate a structured set of configurations on the Yoga-82 test set. Table 2 presents the full component grid and Table 3 analyses progressive accuracy gains across architectural choices leading to SGAN.

Impact of the GCN Encoder. Replacing the flat Single-View MLP with a two-layer GCN operating over the MediaPipe anatomical adjacency graph (GCN + Attention Pool, no ODE) improves accuracy by +0.24%, from 87.53% to 87.77% (Table 2). This gain confirms that encoding known biomechanical connectivity as a structural prior provides a meaningful inductive bias for fine-grained pose discrimination, even before any cross-view reasoning or continuous-time refinement is applied. The GCN forces each joint's representation to aggregate information from its anatomically adjacent neighbours — elbow from shoulder and wrist, knee from hip and ankle — producing spatially coherent node embeddings that a flat linear encoder cannot achieve.

Impact of the Multi-View Transformer. Adding the Transformer encoder without the Neural ODE (GCN + Transformer, no ODE) yields 87.70%, a marginal decrease of −0.07% relative to GCN + Attention Pool alone. This intermediate drop mirrors a pattern observed in prior multi-view work: cross-view self-attention enriches each view's representation by propagating information globally, but without a subsequent refinement stage the added context cannot be fully exploited for classification. Crucially, when the Transformer is paired with per-view Neural ODE evolution in the full SGAN model, the combination reaches 88.41%, demonstrating that the Transformer and ODE are complementary components. The Transformer resolves cross-view ambiguities by propagating information globally across all 16 orientations, and the Neural ODE then refines each enriched embedding along a smooth continuous-time latent trajectory in a way that benefits from the richer initialisation provided by global self-attention.

Impact of Per-View Neural ODE Evolution. The Neural ODE contributes the largest single gain when introduced on top of the full Transformer pipeline: +0.71% from 87.70% to 88.41% (Table 3). Comparing the two ODE-equipped configurations directly — GCN + Attention Pool + Neural ODE (88.17%) versus SGAN (88.41%) — reveals that the ODE benefits substantially from operating on globally contextualised embeddings produced by the Transformer rather than on independently pooled embeddings. The per-view application of the ODE is crucial: each view's latent trajectory is initialised from a representation that already incorporates cross-view context, so the continuous-time dynamics refine a richer starting point. The Dormand-Prince adaptive solver (DOPRI5) allocates more computation to views whose latent dynamics are complex and fewer evaluations to views that are nearly stationary, providing a form of input-dependent efficiency that is well matched to the diverse angular perspectives synthesised during multi-view generation.

Parameter Efficiency. SGAN achieves the highest accuracy among all evaluated configurations with 76,306 parameters. The GCN + Transformer variant without the ODE uses only 50,130 parameters yet achieves 87.70%, demonstrating that the anatomical graph encoder combined with cross-view self-attention

already provides a compact and competitive baseline. Adding the per-view Neural ODE increases the parameter count to 76,306 — an overhead of 26,176 parameters — while delivering a +0.71% accuracy gain, a favourable trade-off for a healthcare-oriented system where both accuracy and deployability on edge devices are priorities.

5 Conclusion

We introduced **SkeleGraphAttNODE** (**SGAN**), a skeleton-only framework for fine-grained yoga pose assessment designed for privacy-preserving and resource-constrained healthcare deployment. The architecture is built around three tightly integrated components: a dual-branch kinematic encoder that combines a two-layer Graph Convolutional Network operating over the anatomical joint adjacency graph with a parallel global MLP for bone vectors and joint angles, projecting each synthetic view into a compact per-view latent embedding that preserves known biomechanical structure; a Multi-View Transformer that applies cross-view self-attention across 16 synthetically generated orientations simultaneously, producing globally contextualised per-view representations; and per-view Neural ODE evolution that refines each contextualised embedding as a smooth continuous-time latent trajectory before max-pooling and classification. Evaluated on the Yoga-82 benchmark, SGAN achieves 88.41% top-1 accuracy with 76,306 parameters, establishing a new state of the art among skeleton-only methods, surpassing two image-based baselines that have full access to raw pixel data, and closing approximately 60% of the remaining gap to the best image-based system at a fraction of its computational cost. A structured ablation confirms that each design decision contributes a measurable accuracy gain: the GCN encoder provides an initial +0.24% gain through biomechanical structural priors, and the Transformer combined with per-view Neural ODE together deliver a further +0.71% by enabling global cross-view contextualisation followed by continuous-time latent refinement.

References

1. Bastico, M., Belmonte-Hernández, A., García, F.Á.: Continuous person identification and tracking in healthcare by integrating accelerometer data and deep learning filled 3d skeletons. IEEE Sensors Journal **22**(15), 15402–15409 (2022)
2. Bera, A., Nasipuri, M., Krejcar, O., Bhattacharjee, D.: Fine-grained sports, yoga, and dance postures recognition: A benchmark analysis. IEEE Transactions on Instrumentation and Measurement **72**, 1–13 (2023)
3. Chen, R.T., Rubanova, Y., Bettencourt, J., Duvenaud, D.K.: Neural ordinary differential equations. Advances in neural information processing systems **31** (2018)
4. Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., Lu, H.: Skeleton-based action recognition with shift graph convolutional network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 183–192 (2020)

5. Dittakavi, B., Bavikadi, D., Desai, S.V., Chakraborty, S., Reddy, N., Balasubramanian, V.N., Callepalli, B., Sharma, A.: Pose tutor: an explainable system for pose correction in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3540–3549 (2022)
6. Li, S., Zhang, J., Fu, X., Zhang, X., Shang, J., Gupta, R.: Matching skeleton-based activity representations with heterogeneous signals for har. In: Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems. pp. 574–587 (2025)
7. Liu, Z., Zhang, H., Chen, Z., Wang, Z., Ouyang, W.: Disentangling and unifying graph convolutions for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 143–152 (2020)
8. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019)
9. Meghana, J., Chethan, H., KS, S.K., SP, S.P.: Comprehensive analysis of pose estimation and machine learning classifiers for precise yoga pose detection and classification. Procedia Computer Science **258**, 3345–3356 (2025)
10. Rubanova, Y., Chen, R.T., Duvenaud, D.: Latent ordinary differential equations for irregularly-sampled time series. Advances in Neural Information Processing Systems **32** (2019)
11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
12. Verma, M., Kumawat, S., Nakashima, Y., Raman, S.: Yoga-82: a new dataset for fine-grained classification of human poses. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 1038–1039 (2020)
13. Wu, Z., Sun, P., Chen, X., Tang, K., Xu, T., Zou, L., Wang, X., Tan, M., Cheng, F., Weise, T.: Selfgcn: Graph convolution network with self-attention for skeleton-based action recognition. IEEE Transactions on Image Processing **33**, 4391–4403 (2024)
14. Xie, J., Meng, Y., Zhao, Y., Nguyen, A., Yang, X., Zheng, Y.: Dynamic semantic-based spatial graph convolution network for skeleton-based human action recognition. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 6225–6233 (2024)
15. 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. vol. 32 (2018)
16. Zhou, H., Liu, Q., Wang, Y.: Learning discriminative representations for skeleton based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10608–10617 (2023)
17. Zhou, L., Zhang, W., Zhang, B., Li, X., Zhu, J.: A strong benchmark for yoga action recognition based on lightweight pose estimation model. Multimedia Systems **31**(1), 66 (2025)
18. Zhou, Y., Yan, X., Cheng, Z.Q., Yan, Y., Dai, Q., Hua, X.S.: Blockgcn: Redefine topology awareness for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2049–2058 (2024)