Transformer-based Geometric Deep Learning for Skeleton Sequence Classification: An improved Approach

Mohamed Amine MEZGHICH[1], Mariem JENDOUBI[1], Yassine ASSIANI[1] and Slim MHIRI[1]
[1]*GRIFT Research Group, CRISTAL laboratory,*
Ecole Nationale des Sciences de l'Informatique (ENSI),
La Manouba University, 2010, La Manouba, Tunisia
{*amine.mezghich, mariam.jendoubi, yassine.assiani, slim.mhiri*}*@ensi-uma.tn*

Keywords: Geometric Deep Learning, Transformer, Skeleton-Based Action Recognition, Manifold Learning, 3D Human Action Recognition

Abstract: This paper presents an improved Geometric Deep Learning Transformer (GDT) architecture for non-Euclidean data classification, with a focus on skeleton-based action recognition. Our model leverages Transformer-based modules to capture both spatial and temporal dynamics in skeleton sequences accurately. Unlike prior approaches, we integrate a manifold learning layer to enhance the understanding of complex geometric patterns in the data, leading to more accurate classification of human actions. Experimental results show that our approach surpasses several state-of-the-art models in 3D human action recognition, as evaluated on benchmark datasets such as NTU RGB+D and NTU RGB+D 120.

1 INTRODUCTION

Action recognition in computer vision is a pivotal task with applications in sectors like video surveillance, human-computer interaction, healthcare, and virtual reality. The ability to recognize human actions accurately from video data offers substantial benefits, enhancing security systems and enabling more intuitive interactions with technology. Among various approaches, skeleton-based action recognition has gained significant traction due to its resilience against variations in appearance, lighting, and viewpoint.

Early approaches to skeleton-based action recognition relied on hand-crafted features and shallow models, often focusing on direct manipulations of skeleton data such as joint angles or distances. While effective to an extent, these methods typically struggled to capture the complex temporal and spatial dependencies inherent in human movement, limiting their adaptability across different activities.

The advent of deep learning has led to the development of models capable of learning rich, hierarchical features directly from data. For instance, Qiu et al. (Qiu et al., 2022) introduced a spatio-temporal tuples transformer that effectively analyzes temporal dynamics and spatial configurations in skeleton sequences, representing a substantial advancement by modeling the interactions between joints over time effectively. Continuing in this vein, Li et al. (Yuan et al., 2019) leveraged the geometric structure inherent in skeleton data through a geometric deep learning framework, illustrating how non-Euclidean data representations can enhance feature extraction and the overall learning process. Moreover, Shi et al. (Shi et al., 2021) introduced the STAR model, which employs sparse representations to focus on critical joints and frames, effectively capturing long-range dependencies within skeleton sequences—a key factor in recognizing complex actions.

Despite these advances, a gap remains in model's ability to fully integrate and leverage the spatio-temporal dimensions of skeleton data. Addressing this, Plizzari et al. (Plizzari et al., 2021) developed a Spatial Temporal Transformer Network, enhancing the model's capacity to synchronize spatial and temporal data streams for a more comprehensive understanding of actions. Furthermore, Friji et al. (Friji et al., 2020) refined the use of geometric deep learning, emphasizing 2D and 3D action recognition from skeleton sequences, thus broadening the applicability of this technology.

Inspired by these groundbreaking works, we present an improved spatio-temporal transformer model that integrates manifold learning to enhance skeleton-based action recognition. Our model com-

bines spatial and temporal transformer modules optimized specifically for the unique dynamics of human motion and incorporates a manifold learning layer. This layer enables the model to capture complex structural patterns in the data, significantly enhancing action recognition capabilities. The main contributions of our work are:

- Integration of a manifold learning layer to enhance the model's ability to capture complex geometric structures within skeleton data, improving representation and recognition of human actions.
- Development of a unified spatial-temporal transformer framework, specifically optimized to handle the unique characteristics of skeleton sequences for improved accuracy in action recognition.
- Comprehensive evaluation on benchmark datasets (NTU RGB+D, NTU RGB+D 120), demonstrating that our model outperforms several leading approaches in 3D action recognition tasks.

2 Related works

Nowadays, deep learning techniques utilizing Recurrent Neural Networks (RNNs), Convolutional Neural Networks (CNNs) , Graph Convolutional Networks (GCNs), and Transformer-based methods have emerged in the action recognition. There are currently numerous studies and surveys in the literature on action recognition.

2.1 CNN-Based Methods

In (Du et al., 2015a), autors proposed an $end-to-end$ hierarchical architecture for skeleton based action recognition with CNN. Li et al. (Li et al., 2017) proposed to encode the spatio-temporal information of skeleton sequences into color texture images, referred to as Joint Distance Maps (JDMs), and Convolutional Neural Networks (ConvNets) are employed to exploit the discriminative features from the JDMs for human action and interaction recognition. (Yang et al., 2019) proposed a Double-feature Double-motion Network (DD-Net) to make the CNN based recognition models run faster. Additionally, (Tang et al., 2022), proposed a self-supervised learning framework under the unsupervised domain adaptation setting, which segments and permutes the time segments or body parts to reduce domain-shift and improve the generalization ability of the model.

2.2 RNN-Based Methods

In (Du et al., 2015b), Du et al. proposed an end-to-end hierarchical RNN for skeleton based action recognition. (Liu et al., 2016) introduced new gating mechanism within LSTM to learn the reliability of the sequential input data and accordingly adjust its effect on updating the long-term context information stored in the memory cell. (Usmani et al., 2023) used a deep recurrent neural network to perform human action recognition through skeletal joint tracking.

2.3 Graph-Based Methods

(Shi et al., 2019) represented the skeleton data as a directed acyclic graph (DAG) based on the kinematic dependency between the joints and bones in the natural human body. (Liu et al., 2020) proposed a new multi-scale aggregation scheme that tackles the biased weighting problem by removing redundant dependencies between further and closer neighborhoods, thus disentangling their features under multi-scale aggregation. (Zhang et al., 2023) proposed a graph-aware transformer (GAT), which can make full use of the velocity information and learn discriminative spatial–temporal motion features from the sequence of the skeleton graphs in a data-driven way.

2.4 Manifold-Based Methods

(Li et al., 2019) introduced a new spatio-temporal manifold network (STMN) that leverages data manifold structures to regularize deep action feature learning. (Zhang et al., 2020) proposed an end-to-end deep manifold-to-manifold transforming network (DMT-Net), which can make SPD matrices flow from one Riemannian manifold to another one for facilitating the action recognition task. (Chen et al., 2022) presented a novel hyperbolic manifold aware network without introducing a dynamic graph. Instead, it leverages Riemannian geometry attributes of a hyperbolic manifold. (Lin et al., 2023) proposed a novel framework Bayesian Contrastive Learning with Manifold Regularization (BCLR).

3 Data Modeling on Sphere

In this section, we introduce the concept of Spherical Modelling and the Inverse Exponential Map technique.

3.1 Spherical Modelling

Let $X \in \mathcal{R}^{n\times k}$ be a body skeleton, where n indicates the number of body joints and k denotes the dimension of X. To remove scale, we propose to model skeletons as elements on a $(n \times k - 1)$ dimension Riemannian manifold, more specifically, the unit sphere S embedded in $R^{n\times k}$. To do so, we divide every skeleton X by its Frobenius norm given by Eq.1:

$$\|X\|_F = \left(\sum_{i,j=1}^{n} |x_{ij}|^2 \right)^{1/2} \tag{1}$$

With this process, we consequently get skeletons' representations as well as their temporal evolution, called trajectories on the unit sphere S embedded in $\mathcal{R}^{n\times k}$. Accordingly, each motion sequence of a moving skeleton is represented with a trajectory on the unit sphere S embedded in $\mathcal{R}^{n\times k}$ as shown in Fig.1

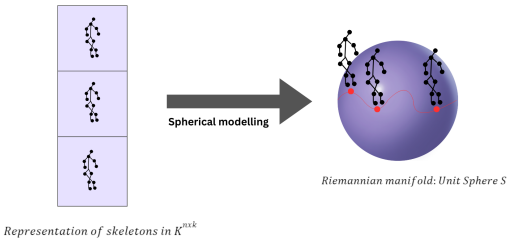

Figure 1: Spherical modelling of the skeleton sequence data.

3.2 Inverse Exponential Map

Based on this work (Friji et al., 2020), the unit sphere S included in $\mathcal{R}^{n\times k}$ has the structure of a Riemannian manifold, the manifold can be assimilated, locally around each point x, to an Euclidean space known as the tangent space $T_X(S)$.

We then describe the tangent space shown in Fig.2 and the inverse exponential map layer used to map data from the Riemannian manifold which is the unit sphere embedded in $\mathcal{R}^{n\times k}$ to a tangent space.

A differentiable d-dimensional manifold X is a topological space where each point x has a neighborhood, which is homeomorphic to a d-dimensional Euclidean space, a.k.a the tangent space and denoted by $T_X(S)$. In other words, at each point x on the manifold X, it is possible to associate a linear space $T_X(S)$. The space $T_X(S)$ is a local Euclidean representation of the manifold X around x. This space is called the tangent space of the manifold X at the point x. Considering that the tangent space is linear and hence equipped with the inner product, the Riemannian metric on S is defined by Eq.2:

$$< X_1, X_2 >= \text{trace}\,(X_1, X_2)\,, X_1, X_2 \in T_X(S) \tag{2}$$

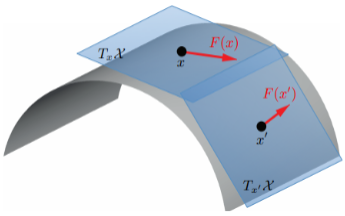

Figure 2: Examples of two tangent spaces: $T_x(X)$ at a point x of the manifold X and $T'_x(X)$ at a point x′ of the manifold X.

The inverse exponential map illustrated in Fig.3, also known as the logarithm map and uniquely defined around a small neighborhood of a point x on the manifold X, is given by Eq.3:

$$\exp_{X_1}^{-1}(X_j) = \frac{\theta}{\sin\theta}\,(X_j - \cos(\theta)X_1) \tag{3}$$

With $\theta = \cos^{-1}\left(\text{trace}\left(X_1\,(X_j)^T\right)\right.$. Here X_1 and X_j represent skeletons on the unit S embedded in $\mathcal{R}^{n\times k}$.

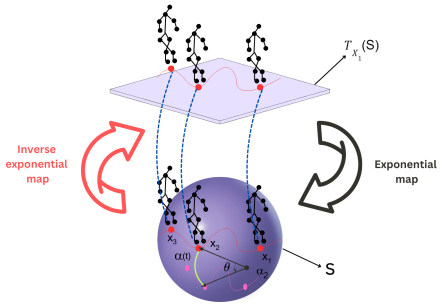

Figure 3: Unit Sphere S embedded in $\mathcal{R}^{n\times k}$, the trajectories α1 and α2 of two sequences of skeletons, the geodesic $\alpha(t)$ connecting arbitrary points on α1 and α2, the tangent space $T_{X_1}(S)$ at the skeleton X1 and skeletons X2 and X3 mapped on $T_{X_1}(S)$

4 Proposed approach

The overall architecture of the proposed methods is illustrated in Figure 4, Figure 5 and Figure 6 Àll models share similar global structural components, with variations specific to each architecture: the Spatial Transformer Attention Network, the Temporal Transformer Attention Network, and the Spatial Temporal Transformer Attention Network.

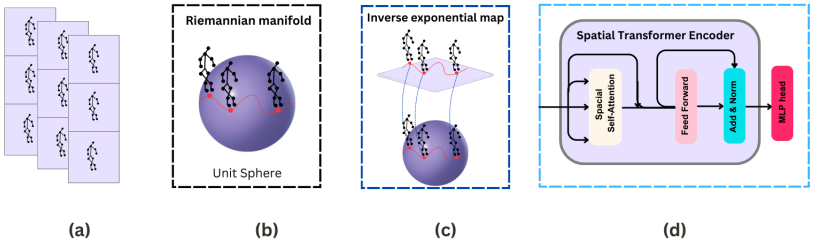

Figure 4: Architecture of the Spatial Transformer Network

To facilitate the design of these architectures, we establish a general framework where in each input is an element of the unit sphere S embedded in $\mathbb{R}^{n\times k}$. Our proposed Spatial-Temporal Transformer (ST-TR) architecture leverages the Transformer self-attention mechanism to operate on both spatial and temporal dimensions. This architecture consists of two modules: Spatial Self-Attention (SSA) and Temporal Self-Attention (TSA), each of which extracts correlations in one of the two dimensions. To address the non-Euclidean structure of the input data, we employ spherical modeling and an inverse exponential map layer. Following spherical modeling of the skeleton sequence data, we utilize Eq. 3 to map each skeleton X_j from the sphere S to the tangent space $T_{X_1}(S)$ at the reference skeleton X_1. As depicted in Fig. 2, we choose a reference skeleton X_1 and map all other skeletons to the tangent space of X_1. Since the tangent space is an Euclidean space, the input data can be fed into any of the three proposed architectures.

4.1 Spatial Transformer Attention Network

In the architecture of a Spatial Transformer Encoder, the core component is the Spatial Self-Attention mechanism, which captures intricate spatial relationships within the data. This mechanism computes interactions between different spatial regions, enhancing feature representation for subsequent processing layers.

The self-attention mechanism operates as follows:

Generate query $\mathbf{q}_i^t \in \mathbb{R}^{d_q}$, key $\mathbf{k}_i^t \in \mathbb{R}^{d_k}$, and value $\mathbf{v}_i^t \in \mathbb{R}^{d_v}$ vectors from input features using distinct linear transformations. Calculate attention scores using Equation (4):

$$\alpha_{ij}^t = \mathbf{q}_i^t \cdot \mathbf{k}_j^{t^T}, \forall t \in T, \quad \mathbf{z}_i^t = \sum_j \text{softmax}_j\left(\frac{\alpha_{ij}^t}{\sqrt{d_k}}\right)\mathbf{v}_j^t \tag{4}$$

Normalize the scores using a softmax function, scaled by the square root of the key dimension (d_k):

$$\mathbf{z}_{ti} = \sum_j \text{softmax}_j\left(\frac{o_{tij}}{\sqrt{d_k}}\right)\mathbf{v}_{tj}$$

After self-attention, the Add & Norm layer refines the feature stabilization. This layer consists of a residual connection followed by layer normalization. The output of the Add & Norm layer is then passed through a feed-forward network (FF), typically comprising two linear transformations with a non-linearity in between. This additional processing further transforms the features.

Finally, an MLP head, consisting of one or more linear layers, is utilized to map the encoded features to the desired output space. This mapping facilitates classification or other tasks. All of these steps are illustrated in Figure 4.

4.2 Temproal Transformer Attention Network

The Temporal Transformer Encoder depicted in the diagram is designed to handle sequential data by focusing on the temporal relationships between data points. At the core of this architecture is the "Temporal Self-Attention" module, where the dynamics of each joint are analyzed independently across all frames. This module computes correlations by comparing features of the same body joint across different times, enhancing the ability to distinguish temporal moments through position encoding. The mathematical formulation involves equations where the attention weights $\alpha_{vi,vj}$ are computed by taking the dot product of query $\mathbf{q}_{vi}$ and key $\mathbf{k}_{vj}$, scaled by the dimensionality of the key:

$$\alpha_{ij}^v = \mathbf{q}_i^v \cdot \mathbf{k}_j^v \quad \forall v \in V, \quad \mathbf{z}_i^v = \sum_j \text{softmax}_j\left(\frac{\alpha_{ij}^v}{\sqrt{d_\mathbf{k}}}\right)\mathbf{v}_j^v \tag{5}$$

This result is then used in a softmax function to normalize the weights, which are multiplied by the value $\mathbf{v}_{vj}$ to produce the output embedding $\mathbf{z}_{vi}$:

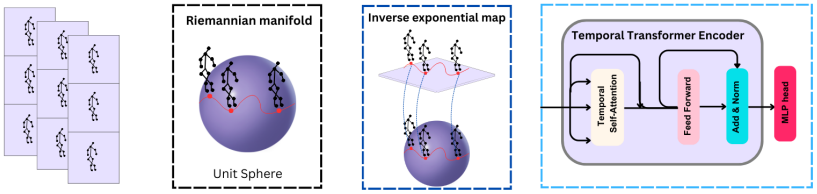

Figure 5: Architecture of the Temporal Transformer Network

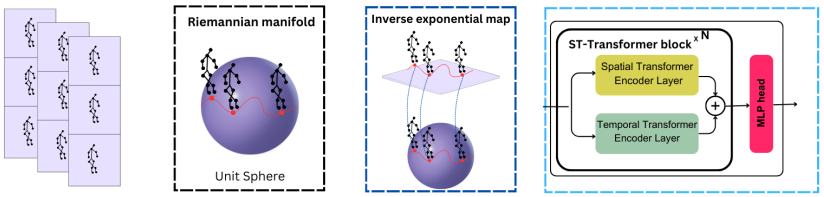

Figure 6: Overview of the Spatial-Temporal Transformer Network

$$\mathbf{z}_{vi} = \sum_j \text{softmax}_j(\alpha_{vi,vj})\mathbf{v}_{vj} \tag{6}$$

Following the self-attention, the "Feed Forward" block processes the attention-modified data, which is then passed through an "Add Norm" layer. This layer helps maintain training stability by combining the original input with the output of the feed-forward layer, followed by layer normalization. Finally, the "MLP Head" at the end of the encoder transforms the encoded features into a suitable form for classification tasks, as illustrated in Figure 5.

4.3 Spatial Temproal Transformer Network

The Spatial-Temporal Transformer Network in Figure 6 is constructed using multiple Spatial-Temporal Transformer blocks (ST-blocks), followed by an MLP head for classification.

Each ST-block contains two pipelines: a spatial Transformer encoder and a temporal Transformer encoder. Each encoder includes several key components such as multi-head self-attention (MHSA), a skip connection (shown as 'Add Norm' in Figure 5 and 6), and a feed-forward network. The spatial Transformer encoder employs sparse attention to capture the topological correlations of connected joints within each frame. The temporal Transformer encoder uses segmented linear attention to grasp the correlations of joints along the temporal dimension. The combined outputs of both encoders are then fed into the MLP head for classification.

Positional encoding is also applied before each ST-block to provide context for the sequence ordering of the input.

5 Experimentals results

We first introduce the datasets used in our experiments (Section 5.1). Next, we present implementation details (Section 5.2) and compare our work to previous studies (Section 5.3). Finally, we provide qualitative results and discuss limitations.

5.1 Datasets

In this section, we introduce the datasets used for our experiments, including the NTU RGB+D dataset, where we tested our proposed architecture.

NTU-RGB+D: (Shahroudy et al., 2016) is currently the largest dataset with 3D joints annotations for human action recognition tasks. This dataset contains $56,000$ action clips in 60 action classes. These clips are all performed by 40 volunteers captured in a constrained lab environment, with three camera views recorded simultaneously. The provided annotations give 3D joint locations (X,Y,Z) in the camera coordinate system, detected by the Kinect depth sensors. There are 25 joints for each subject in the skeleton sequences. Each clip is guaranteed to have at most 2 subjects.

NTU RGB+D 120: is a large-scale dataset for RGB+D human action recognition, which is collected

from 106 distinct subjects and contains more than 114 thousand video samples and 8 million frames. This dataset contains 120 different action classes including daily, mutual, and health-related activities. For evaluating action recognition models on this dataset, two standard evaluation protocols are used: Cross-Subject and Cross-Setup protocols.

5.2 Evaluation Protocols

5.2.1 NTU-RGB+D

we present our criteria for two types of action classification evaluation, in order to obtain standard evaluations for all the reported results on this benchmark.

Cross-subject Protocol

For the cross-subject evaluation protocol, we split the 40 subjects into training and testing sets, each composed of 20 subjects. The training and testing groups are made up of 40,320 and 16,560 samples, respectively. In our work, we use for training the subjects whose IDs are among the following list of values: 1, 2, 4, 5, 8, 9, 13, 14, 15,16, 17, 18, 19, 25, 27, 28, 31, 34, 35, 38. The 20 remaining subjects are reserved for testing.

Cross-view Protocol

In cross-view protocol, we select the samples from cameras 2 and 3 for training and the samples from camera 1 for testing. The training set consists then of the front and two side views of the actions,whilst the testing set incorporates left and right 45 degree views of the action performances. For this assessment, the training and testing sets have 37,920 and 18,960 samples, respectively

5.2.2 NTU RGB+D 120

We present our criteria for two types of action classification evaluation to ensure standardized assessments across all reported results on this benchmark: Cross-Subject and Cross-Setup.

Cross-subject Protocol

The Cross-Subject protocol in the NTU RGB+D 120 dataset divides the data based on individuals. The training set includes samples from 53 subjects, while the testing set includes samples from a different 53 subjects. This separation ensures no overlap between subjects in the training and testing sets. By doing so, the Cross-Subject protocol evaluates the generalization capability of the model, testing its ability to recognize and classify actions performed by new, unseen subjects.

Cross-Setup Protocol

The Cross-Setup protocol in the NTU RGB+D 120 dataset divides the data based on camera configurations. The training set includes samples from half of the camera setups, while the testing set includes samples from the other half. This ensures that the camera positions and angles differ between the training and testing sets, evaluating the model's robustness to various camera setups.

5.3 Implmentation Details

Our model architecture consists of four transformer layers with a dimension of 64, employing eight attention heads and a multi-layer perceptron (MLP) with a dimension of 1024. Dropout regularization with a rate of 0.1 is applied to both the attention mechanism and the embedding layer to enhance generalization and prevent overfitting. Notably, our data preprocessing involves resizing the number of frames from 300 to 32, which optimizes computational efficiency without compromising performance. This configuration, tailored to our dataset with 60 classes and input dimensions (32, 2, 32, 25, 3), provides a robust framework for achieving accurate results in our experiments.

5.4 Results

In this section, we evaluate the model's performance based on accuracy across two major datasets: NTU RGB+D and NTU RGB+D 120. The results are reported for two variations of our model: the Normal Model (without the manifold layer) and our Improved Model (with the manifold layer), using three configurations Spatial, Temporal, and Temporal-Spatial.
Our Improved Model incorporates a manifold learning layer, designed to capture the underlying geometric structure of the skeleton data, which enhances the model's ability to understand complex motion patterns. This additional layer differentiates our method from the Normal Model and contributes to consistently higher accuracy across various configurations.

5.4.1 NTU-RGB+D

For the NTU RGB+D dataset, we use two evaluation protocols, cross-view and cross-subject. The results are presented in Table 1 and 2.

Cross View Data

Table 1 presents accuracy scores for both model variations under the cross-view protocol, where the model is trained on samples from two camera views and tested on an unseen third view.
The Improved Model consistently achieves higher accuracy across all configurations. In the Spatial configuration, the addition of the manifold layer raises accuracy from 74.942% to 75.666%, demonstrating that

Table 1: Results on NTU-RGB+D dataset (Cross View Data)

Model	Spatial	Temporal	Temporal-Spatial
Normal Model	74,942%	80.345%	81.432%
Manifold Model	75.666%	80.715%	81.666%

Table 2: Results on NTU-RGB+D dataset (Cross Subject Data)

Model	Spatial	Temporal	Temporal-Spatial
Normal Model	69.855%	74.204%	75.823%
Manifold Model	70.413%	74.641%	76.248%

the manifold layer enhances the model's spatial understanding by effectively capturing spatial relationships in the skeleton structure. The Temporal-Spatial configuration also benefits, with accuracy improving from 81.432% to 81.666%, which indicates the advantage of combining spatial and temporal modeling with manifold learning in cross-view settings.

Cross Subject Data

Table 2 shows accuracy for both models under the cross-subject protocol, which divides data by subjects to evaluate generalization to unseen individuals.
In the cross-subject protocol, the Improved Model again outperforms the Normal Model across all configurations. The manifold layer improves accuracy in the Temporal-Spatial configuration from 75.823% to 76.248%, indicating that the model benefits from the geometric understanding provided by the manifold layer when dealing with unseen subjects. The highest gain is in the Temporal-Spatial configuration, reinforcing the layer's value in complex motion understanding, where both spatial and temporal dependencies must be modeled.

The comparison presented in Table 3 underscores the effectiveness of our improved manifold-based model in achieving better performance on the NTU RGB+D dataset. The 1-layer and 2-layer PLSTM models introduced by (Shahroudy et al., 2016) achieve accuracies of 62.05% and 62.93% for cross-subject and 69.40% and 70.27% for cross-view protocols, respectively. While the Euclidean CNN-LSTM model by (Friji et al., 2020) demonstrates a foundational capability in capturing spatial relationships with cross-subject accuracy at 56.61% and cross-view accuracy at 62.32%, the non-Euclidean CNN-LSTM model enhances performance by addressing the non-Euclidean nature of skeleton data, reaching a cross-subject accuracy of 61.45% and a cross-view accuracy of 71.03%. In comparison, our improved manifold model achieves 70.41% for cross-subject and 74.64% for cross-view accuracies, highlighting its ability to capture the underlying geometric structures in skeleton data and enabling more robust recognition across varied subjects and views.

5.4.2 NTU RGB+D 120

The NTU RGB+D 120 dataset, which is larger and more diverse, provides a more challenging testbed. We use the cross-subject and cross-setup protocols here as well and the results are presented in Table 4 and 5.

Cross Subject Data

Table 4 illustrates the model's performance on NTU RGB+D 120 under the cross-subject protocol, where data is split by subjects.
On this larger dataset, the Improved Model shows even greater advantages. The manifold layer enhances accuracy in the Temporal-Spatial configuration from 74.298% to 75.934%, showcasing its role in effectively capturing more complex, diverse motions in skeleton data. This consistent improvement across configurations highlights the manifold layer's ability to generalize to unseen subjects by leveraging geometric insights into skeleton data.

Cross Setup Data

Table 5 presents accuracy under the cross-setup protocol, where the training and testing samples are split based on camera configurations.
The Improved Model again surpasses the Normal Model in all configurations. The manifold layer boosts accuracy in the Temporal-Spatial configuration from 79.379% to 80.951%. This increase indicates the manifold layer's effectiveness in capturing structural nuances, making the model robust to variations in camera setups. The manifold layer's benefits in the cross-setup protocol demonstrate its potential for real-world applications where different camera views are common.

Across both the NTU RGB+D and NTU RGB+D 120 datasets, and under both evaluation protocols (cross-subject and cross-setup), the Improved Model with the manifold layer consistently outperforms the Normal Model in all three perspectives: spatial, temporal, and temporal-spatial. The most significant gains are observed in the Temporal-Spatial configurations, highlighting the manifold layer's ability to capture complex interactions in both space and time,

Table 3: Results on NTU RGB+D dataset using two evaluation protocols: cross-subject and cross-view

Model	Cross Subject Accuracy	Cross View Accuracy
1 Layer PLSTM (Shahroudy et al., 2016)	62.05%	69.40%
2 Layer PLSTM (Shahroudy et al., 2016)	62.93%	70.27%
Euclidean CNN-LSTM (Friji et al., 2020)	56.61%	62.32%
Non Euclidean CNN-LSTM (Friji et al., 2020)	61.45%	71.03%
Manifold Model	70.413%	74.641%

Table 4: Results on NTU RGB+D 120 dataset (Cross Subject Data)

Model	Spatial	Temporal	Temporal-Spatial
Normal Model	69.332%	72.715%	74.298%
Manifold Model	71.512%	74.611%	75.934%

especially in settings where understanding nuanced body movements is essential.
By leveraging a non-Euclidean geometric approach, the manifold layer provides a deeper and more robust representation of skeleton data, accommodating the intricate structural patterns within human actions. This advanced layer allows the model to go beyond conventional spatio-temporal processing by embedding action sequences on a Riemannian manifold, which aligns closely with the underlying geometry of human motion. While spatio-temporal models are not novel on their own, our approach stands out by integrating manifold learning, which sets a new benchmark for accurately recognizing complex, multi-dimensional movements in skeleton-based action recognition. This innovation provides a stronger foundation for interpreting and classifying diverse human actions in real-world applications.

6 Conclusion

This paper has explored the evolution of skeleton-based action recognition, a field that plays a crucial role in various applications ranging from video surveillance to healthcare. Initially reliant on handcrafted features and shallow learning models, the domain has witnessed significant transformation with the integration of deep learning techniques that efficiently capture the complex spatial and temporal dynamics of human movement.

Recent advancements have leveraged sophisticated models like the spatio-temporal tuples transformer, geometric deep learning frameworks, and the Spatial Temporal Transformer Network, each enhancing the understanding and analysis of skeleton data. These developments underscore a shift towards models that not only recognize basic movements but also interpret intricate human interactions with high accuracy and adaptability.

Our contribution, a novel spatio-temporal transformer-based model with an added manifold learning layer, represents a synthesis of these advancements. It innovatively combines spatial and temporal analysis with the ability to discern complex structural patterns, setting a new standard for action recognition technology. This model promises enhanced accuracy in recognizing a broader array of human actions and could serve as a stepping stone for future research aiming to fully exploit the potential of skeleton-based frameworks in real-world applications.

Looking forward, the potential for further advancements in skeleton-based action recognition is substantial, with research efforts increasingly focused on refining and optimizing existing models. The ongoing exploration of advanced learning architectures and innovative techniques, such as manifold learning, suggests a trajectory towards more sophisticated action recognition systems.

7 Acknowledgements

This work is supported in part by the Smart It Partner company and the GRIFT research group, the CRISTAL laboratory, Tunisia.

REFERENCES

Chen, J., Zhao, C., Wang, Q., and Meng, H. (2022). Hmanet: Hyperbolic manifold aware network for skeleton-based action recognition. *IEEE Transactions on Cognitive and Developmental Systems*.

Table 5: Results on NTU RGB+D 120 dataset (Cross Setup Data)

Model	Spatial	Temporal	Temporal-Spatial
Normal Model	73.688%	78.417%	79.379%
Manifold Model	74.399%	79.885%	80.951%

Du, Y., Fu, Y., and Wang, L. (2015a). Skeleton based action recognition with convolutional neural network. In *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, pages 579–583. IEEE.

Du, Y., Wang, W., and Wang, L. (2015b). Hierarchical recurrent neural network for skeleton based action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118.

Friji, R., Drira, H., and Chaieb, F. (2020). Geometric deep learning on skeleton sequences for 2d/3d action recognition. In *VISIGRAPP (5: VISAPP)*, pages 196–204.

Li, C., Hou, Y., Wang, P., and Li, W. (2017). Joint distance maps based action recognition with convolutional neural networks. *IEEE Signal Processing Letters*, 24(5):624–628.

Li, C., Zhang, B., Chen, C., Ye, Q., Han, J., Guo, G., and Ji, R. (2019). Deep manifold structure transfer for action recognition. *IEEE Transactions on Image Processing*, 28(9):4646–4658.

Lin, L., Zhang, J., and Liu, J. (2023). Bayesian contrastive learning with manifold regularization for self-supervised skeleton based action recognition. In *2023 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE.

Liu, J., Shahroudy, A., Xu, D., and Wang, G. (2016). Spatio-temporal lstm with trust gates for 3d human action recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 816–833. Springer.

Liu, Z., Zhang, H., Chen, Z., Wang, Z., and Ouyang, W. (2020). Disentangling and unifying graph convolutions for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 143–152.

Plizzari, C., Cannici, M., and Matteucci, M. (2021). Spatial temporal transformer network for skeleton-based action recognition. In *Pattern recognition. ICPR international workshops and challenges: virtual event, January 10–15, 2021, Proceedings, Part III*, pages 694–701. Springer.

Qiu, H., Hou, B., Ren, B., and Zhang, X. (2022). Spatio-temporal tuples transformer for skeleton-based action recognition. *arXiv preprint arXiv:2201.02849*.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019.

Shi, F., Lee, C., Qiu, L., Zhao, Y., Shen, T., Muralidhar, S., Han, T., Zhu, S.-C., and Narayanan, V. (2021). Star: Sparse transformer-based action recognition. *arXiv preprint arXiv:2107.07089*.

Shi, L., Zhang, Y., Cheng, J., and Lu, H. (2019). Skeleton-based action recognition with directed graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7912–7921.

Tang, Y., Liu, X., Yu, X., Zhang, D., Lu, J., and Zhou, J. (2022). Learning from temporal spatial cubism for cross-dataset skeleton-based action recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–24.

Usmani, A., Siddiqui, N., and Islam, S. (2023). Skeleton joint trajectories based human activity recognition using deep rnn. *Multimedia Tools and Applications*, 82(30):46845–46869.

Yang, F., Wu, Y., Sakti, S., and Nakamura, S. (2019). Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the 1st ACM International Conference on Multimedia in Asia*, pages 1–6.

Yuan, X., He, P., Zhu, Q., and Li, X. (2019). Adversarial examples: Attacks and defenses for deep learning. *IEEE transactions on neural networks and learning systems*, 30(9):2805–2824.

Zhang, J., Xie, W., Wang, C., Tu, R., and Tu, Z. (2023). Graph-aware transformer for skeleton-based action recognition. *The Visual Computer*, 39(10):4501–4512.

Zhang, T., Zheng, W., Cui, Z., Zong, Y., Li, C., Zhou, X., and Yang, J. (2020). Deep manifold-to-manifold transforming network for skeleton-based action recognition. *IEEE Transactions on Multimedia*, 22(11):2926–2937.