# SPHERICALSHIFT POINT TRANSFORMER: ROBUST AND EFFICIENT ATTENTION FOR 3D POINT CLOUD PROCESSING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

We introduce the SphericalShift Point Transformer (SSPT), a novel 3D point cloud processing framework that extends the scalable design of Point Transformer V3 (PTv3) by addressing two key limitations: the loss of spatial neighbor precision inherent in serialized attention and the slower convergence of the dot-product attention mechanism. While PTv3 achieves state-of-the-art performance across both indoor and outdoor tasks by efficiently expanding its receptive field from 16 to 1024 points through serialized attention coupled with enhanced conditional positional encoding, SSPT rethinks both data organization and the attention strategy to more precisely capture local geometric relationships. In our method, each point is first projected into a spherical coordinate system using a robust reference—such as one derived from principal component analysis—and subsequently grouped into patches using a hierarchical equal-area grid inspired by HEALPix. This grouping preserves local spatial structures typically lost in conventional serializations. A shifted spherical-window attention mechanism is then applied to ensure that boundaries between patches are revisited repeatedly, thereby recovering fine-grained spatial relationships and mitigating neighbor precision loss. In parallel, SSPT incorporates a dual-modal attention module that fuses standard dot-product attention with a vector-based correlation head through learnable fusion weights, enhancing training convergence and improving the capture of non-linear geometric relationships. Complementing these components, a novel spherical positional encoding leverages angular coordinates and local curvature to inject informative contextual biases into the attention layers, ensuring robustness under rotations and scaling variations. Extensive experiments—including an end-to-end benchmark on datasets such as ModelNet40 and ShapeNet, a comprehensive component ablation study, and a robustness evaluation under rotational, scaling, and noise perturbations—demonstrate the efficacy of our approach. Our main contributions are summarized as follows:

- **Scalable Design:** Extends the PTv3 architecture to efficiently process large-scale point clouds by significantly expanding the receptive field.
- **Spherical Projection and Grouping:** Introduces a novel projection of points into a spherical coordinate system and groups them using a hierarchical equal-area grid inspired by HEALPix, thereby preserving local geometric coherence.
- **Shifted Spherical-Window Attention:** Proposes an attention mechanism that employs shifted windows over the spherical grid to continuously capture cross-patch relationships and recover fine-grained spatial details.
- **Dual-Modal Attention:** Fuses standard dot-product attention with a vector-based correlation head via learnable fusion weights, resulting in faster convergence and improved modeling of non-linear geometric relationships.
- **Spherical Positional Encoding:** Develops a positional encoding based on angular coordinates and local curvature that provides robust contextual bias under transformations such as rotations and scaling.

Overall, SSPT redefines the paradigm for point cloud processing by integrating these innovations into a cohesive framework, thereby paving the way for future

research on hybrid attention mechanisms and large-scale joint training strategies in 3D perception.

## 1 INTRODUCTION

The rapid evolution of 3D sensors together with the growing demand in computer vision applications has spurred extensive research on processing unstructured point cloud data. State-of-the-art methods such as Point Transformer V3 (PTv3) **?** rely on large-scale training and efficient serialized attention to achieve high performance across a variety of 3D perception tasks. However, a fundamental trade-off persists between accuracy and computational efficiency. In PTv3, for instance, the use of an inexpensive dot-product attention mechanism enables scalability but comes at the cost of slower convergence and limited capacity to capture fine-grained spatial relationships. Moreover, the reliance on serialized neighbor mapping based on space-filling curves (e.g., Z-order or Hilbert) can sacrifice local geometric consistency for the sake of computational efficiency.

Motivated by these challenges, we propose the SphericalShift Point Transformer (SSPT), a novel backbone for point cloud processing that rethinks both data organization and attention mechanisms. In SSPT, raw point coordinates are first projected from their native Euclidean $(x, y, z)$ space into a spherical domain via the transformation

$$r = \sqrt{x^2 + y^2 + z^2}, \quad \theta = \arccos\Big(\frac{z}{r}\Big), \quad \phi = \arctan\Big(\frac{y}{x}\Big), \tag{1}$$

which facilitates an equal-area partitioning of the data. In this spherical domain, points are grouped using an adapted hierarchical equal-area grid inspired by the HEALPix framework. This grouping preserves local spatial relationships and ensures that each patch represents a balanced subset of the original points, thereby maintaining geometric coherence.

On the structured spherical grid, SSPT adopts a shifted-window attention mechanism. Instead of performing attention on a serialized list of neighbors, attention is computed over windows defined on the spherical grid. A shifting strategy produces overlapping regions across patch boundaries, enabling continuous cross-patch integration and capturing fine-grained spatial details that conventional serialization may miss.

To further mitigate convergence challenges associated with the simplistic dot-product attention utilized in PTv3, SSPT incorporates a dual-modal attention module. This module fuses standard dot-product attention with a vector-based correlation function, producing similarity measures that better reflect the non-linear geometry of the spherical patches. In parallel, traditional relative positional encoding is replaced by a novel spherical positional encoding (SPE) derived from the angular components $\theta$ and $\phi$. SPE not only captures the angular layout but also encodes aspects of local curvature, imparting robustness to rotations and scaling transformations.

Our approach offers several key advantages. In summary, the main contributions of this work are:

- **Spherical Projection:** Raw points are transformed from Euclidean to spherical coordinates, enabling an equal-area partitioning that preserves local geometric consistency and facilitates hierarchical grouping.
- **Shifted Spherical-Window Attention:** By computing attention over overlapping windows on a structured spherical grid, the proposed mechanism integrates information across patch boundaries to capture fine spatial details.
- **Dual-Modal Attention:** The fusion of dot-product and vector-based correlation attention yields a more expressive similarity metric, which accelerates convergence and improves performance in modeling complex 3D structures.
- **Spherical Positional Encoding (SPE):** Deriving positional cues from the angular components $\theta$ and $\phi$ allows SPE to effectively encode local curvature and spatial layout, thereby enhancing robustness to rotations and scaling perturbations.

By integrating these innovations, SSPT builds upon the scalable architecture of PTv3 while directly addressing its limitations. In contrast to serialized neighbor mapping, the spherical projection used in SSPT preserves natural geometric groupings. The shifted-window attention mechanism ensures

robust information exchange across patches, and the dual-modal attention module provides a nuanced measure of point similarity. Finally, SPE further enhances the network's robustness to common 3D transformations.

Our experimental evaluation is organized into three major studies. First, an end-to-end benchmark compares SSPT with PTv3 on standard datasets such as ModelNet40 (classification) and ShapeNet (segmentation) and evaluates convergence speed, final accuracy, and inference efficiency. Second, a component ablation study systematically removes the proposed modules—namely, spherical projection, shifted-window attention, dual-modal attention, and SPE—to quantify their individual contributions using statistical analyses and visualizations. Third, robustness evaluations test both SSPT and PTv3 under controlled perturbations including rotations, scaling variations, and additive noise, demonstrating that SSPT's spherical transformation and SPE significantly mitigate performance degradation.

In summary, our contributions can be succinctly stated as follows:

- **Improved Neighborhood Precision:** The use of spherical projection and equal-area grid partitioning preserves local geometric consistency, enabling the extraction of fine-grained features in point cloud data.
- **Enhanced Convergence through Dual-Modal Attention:** By fusing conventional dot-product attention with a vector-based correlation mechanism, SSPT achieves faster convergence while providing a more robust similarity measure tailored to complex 3D geometries.
- **Robust Spherical Positional Encoding:** SPE, derived from the angular coordinates $\theta$ and $\phi$, effectively encodes local curvature and layout information, resulting in improved robustness to rotational and scaling transformations.
- **Comprehensive Experimental Validation:** Through rigorous end-to-end benchmarking, component ablation, and robustness evaluations, we demonstrate that SSPT not only matches but exceeds the performance and efficiency of current state-of-the-art methods such as PTv3.

The SphericalShift Point Transformer represents a significant advancement in processing unstructured point cloud data. By re-envisioning data organization and attention mechanisms, SSPT effectively resolves intrinsic trade-offs between accuracy and efficiency, paving the way for more robust and scalable 3D perception systems. The methodology and subsequent experimental results presented in the following sections affirm the effectiveness of our design choices.

## 2 Related Work

Transformer-based architectures have dramatically reshaped point cloud processing by facilitating the modeling of long-range dependencies and capturing global context through self-attention. In this section, we review key developments in transformer architectures for point cloud analysis, methods for geometric data partitioning using spherical representations, and improved attention mechanisms that yield robust, geometrically faithful models.

### 2.1 Transformer Architectures for Point Cloud Analysis

Early point cloud methods relied on multilayer perceptrons and convolutional networks adapted from image processing. The emergence of transformer models, as exemplified by Point Transformer V3 **?**, has shifted the paradigm toward scalable architectures that efficiently aggregate global context. In PTv3, a serialized neighbor mapping strategy replaces the computationally expensive k-nearest neighbor search. This enables a dramatic increase in the receptive field—from 16 to 1024 points—while simultaneously improving inference speed and reducing memory consumption. However, relying solely on dot-product attention can lead to slower convergence and limitations in spatial neighbor precision. Recent work suggests that incorporating vector-based similarity measures into attention modules better captures the inherent nonlinearity of point cloud data.

### 2.2 Geometric Data Partitioning via Spherical Representations

A parallel research direction focuses on embedding geometric priors into point cloud processing. Techniques that employ hierarchical equal-area grids inspired by the HEALPix framework project

3D points into a spherical coordinate system and partition them into patches that mirror the natural structure of 3D surfaces. Such grid-based partitioning preserves local spatial details and enforces a balanced distribution of points. This method overcomes some of the limitations inherent in serialized ordering techniques and supports more robust geometric analysis.

### 2.3 Enhanced Attention Mechanisms and Dual-Modal Approaches

In transformer architectures for point clouds the reliance on standard dot-product attention has prompted efforts to improve convergence and spatial precision. Recent models have integrated dual-modal attention strategies that combine conventional dot-product computations with a supplementary vector-based correlation head. The newly proposed SphericalShift Point Transformer (SSPT), for instance, fuses these complementary mechanisms so that the network can both leverage faster-converging signals and capture fine-grained geometric details. Furthermore, replacing conventional relative positional encoding with a spherical positional encoding—derived from angular coordinates on a spherical grid—incorporates local curvature information and enhances robustness to rotations and scaling.

### 2.4 Summary of Prior Contributions

- **Scalability and Efficiency: ?** demonstrates that by scaling transformer architectures and substituting complex neighbor search operations with a serialized mapping strategy, point cloud networks can achieve significant improvements in both inference speed and memory consumption.
- **Geometric Grouping:** Approaches based on space-filling curves and spherical projections effectively preserve local spatial structures, which is critical for segmentation and detection tasks.
- **Attention Optimization:** The limitations of purely dot-product attention have led to the development of dual-modal attention modules that merge conventional dot-product measures with vector-based correlation, thereby capturing non-linear relationships more effectively.
- **Spherical Positional Encoding:** Positional encodings derived from spherical coordinate systems naturally incorporate local curvature information and enhance robustness to geometric transformations compared to traditional relative positional encodings.

In summary, the evolution from PTv3 to SSPT underscores a broader trend toward architectures that balance efficiency, scalability, and geometric precision. Notably, advances in spherical partitioning and dual-modal attention represent significant steps in surmounting the intrinsic limitations of earlier transformer-based approaches in point cloud analysis.

## 3 Background

In this section, we provide a comprehensive background on the evolution of point cloud processing techniques and formally articulate the problem setting and notation that underlie our work. We also critically analyze the limitations of earlier approaches, motivating the development of our proposed SphericalShift Point Transformer (SSPT).

### 3.1 Historical Overview and Academic Ancestors

Point cloud processing has evolved rapidly with advances in 3D sensing technologies and representation learning methods. Early methods typically relied on handcrafted features and clustering techniques to extract geometric cues from raw 3D point data. However, the intrinsic irregularity and unstructured nature of point clouds posed considerable challenges for conventional algorithms and classical deep neural networks. A significant breakthrough emerged with architectures that directly operate on point sets, such as PointNet and its successors. In particular, the work by **?** introduced Point Transformer V3 (PTv3), which rearranges unstructured point clouds into a serialized order using space–filling curves (e.g., Z-order or Hilbert curves). PTv3 integrates an efficient serialized attention mechanism with an enhanced conditional positional encoding to scale the receptive field from 16 to 1024 points, achieving state-of-the-art performance on over twenty diverse 3D perception tasks.

### 3.2 Problem Formulation and Notation

We consider the problem of processing unordered sets of 3D points. A point cloud is formally defined as

$$P = \{x_i\}_{i=1}^{N},$$

where each $x_i$ is a point in three-dimensional space and $N$ is the total number of points. In common tasks such as classification and segmentation, the goal is to learn a mapping

$$f : \mathbb{R}^{N\times 3} \to \mathcal{Y},$$

where $\mathcal{Y}$ represents a label space or a set of semantic scores. In the supervised setting, the training dataset is denoted by

$$\mathcal{D} = \{(P_j, y_j)\}_{j=1}^{M},$$

with $y_j$ being the ground-truth label (or segmentation mask) for the $j^{\text{th}}$ point cloud.

A central mechanism in Transformer-based architectures is the dot–product attention, mathematically described as

$$\text{Attention}(Q, K, V) = \text{softmax}\Big(\frac{(QK^T)}{\sqrt{d}}\Big)V, \tag{2}$$

where $Q$, $K$, and $V$ denote the query, key, and value matrices, respectively, and $d$ is the dimensionality of each feature vector. While widely adopted, this formulation relies on precise neighbor relationships and, when extended to deeper networks, can impede convergence.

### 3.3 Limitations of Prior Approaches

Most existing techniques, including PTv3, serialize point clouds by mapping multi-dimensional data to a one-dimensional order using space–filling curves. Although this strategy helps preserve local structure to some extent, it approximates rather than fully captures the true spatial relationships among neighboring points. Furthermore, the prevalent use of standard relative positional encodings based solely on Euclidean distances does not robustly address variations due to rotations or scaling, which are common in practical 3D sensing applications. Finally, exclusive reliance on dot–product attention, despite its efficiency, can result in slower convergence and limits effective depth scaling.

### 3.4 Motivation for the SphericalShift Point Transformer

The identified challenges motivate our proposed SphericalShift Point Transformer (SSPT). Instead of employing a one-dimensional serialization, SSPT projects the 3D point cloud into a spherical coordinate system. A robust reference—obtained, for example, via principal component analysis—defines a natural center and orientation for the sphere. This spherical projection enables the grouping of points into patches using an adapted hierarchical equal–area grid inspired by HEALPix, thereby preserving more accurate local geometric relationships.

Building on this representation, SSPT introduces a shifted spherical-window attention mechanism. Unlike fixed-window attention on a flat grid, the shifted approach continuously reexamines patch boundaries, capturing fine–grained spatial correlations that are lost in conventional serialization. To alleviate the slower convergence commonly associated with pure dot–product attention, our framework incorporates a dual–modal attention module that fuses standard dot–product attention with a complementary vector–based correlation branch through learnable weights. Additionally, we propose a novel spherical positional encoding (SPE) that leverages angular coordinates from the spherical grid to generate bias terms for the attention layers. This SPE not only captures local curvature but also enhances robustness against rotations and scaling variations.

### 3.5 Contributions of the Proposed Framework

Our work makes the following key contributions:

- **Spherical Projection and Hierarchical Grouping:** We introduce a novel spherical projection technique that restructures 3D point clouds into a structured spherical domain using an adapted equal–area grid inspired by HEALPix. This partitioning yields geometrically coherent patches with a balanced representation of local features.

- **Shifted Spherical-Window Attention:** We propose a dynamic attention mechanism that operates on the spherical grid. By dynamically shifting the attention windows, our method continuously revisits patch boundaries and captures fine–grained spatial correlations more effectively than standard fixed serialized orderings.
- **Dual–Modal Attention for Improved Convergence:** Our attention fusion module combines conventional dot–product attention with a vector–based correlation branch via learnable weights. This design accelerates training convergence while preserving powerful feature representations.
- **Spherical Positional Encoding (SPE):** We develop a new positional encoding scheme that derives bias terms from the angular coordinates of the spherical grid. This encoding captures local curvature information and improves invariance to rotations and scale variations, thereby enhancing overall robustness.

In summary, the SphericalShift Point Transformer builds upon the scalability and simplicity of previous frameworks such as PTv3 while directly addressing their limitations in convergence, spatial precision, and sensitivity to transformations. The remainder of the paper details our experimental evaluation, including end-to-end benchmarks, component ablation studies, and robustness assessments that demonstrate the efficacy of our approach.

## 4 METHOD

In this section, we describe the methodology behind our proposed SphericalShift Point Transformer (SSPT), a robust framework for point cloud processing that builds on the scalable design of Point Transformer V3 (PTv3) **?** by addressing limitations in both neighbor precision and convergence speed through four key innovations.

- **Spherical Projection and Hierarchical Grouping:** Rather than using serialized neighbor mapping based on space-filling curves, SSPT projects raw 3D points into a spherical coordinate system. Using a robust center computed by, e.g., principal component analysis (PCA), each point is re-centered and then transformed into spherical coordinates. A hierarchical, equal-area grid (inspired by HEALPix) partitions the spherical domain into patches, preserving local geometric structure and balancing the receptive field across groups.
- **Shifted Spherical-Window Attention:** Attention is computed within spherical windows defined over the partitioned grid. By periodically shifting these windows with an offset $\Delta$, regions on or near patch boundaries are re-visited in successive layers, thereby capturing fine-grained spatial relationships without incurring significant computational overhead.
- **Dual-Modal Attention:** To accelerate convergence and enrich the feature representation, SSPT incorporates a dual-modal attention module that fuses standard dot-product attention with a vector-based correlation branch. The outputs of these branches, $F_{dp}$ and $F_{vec}$ respectively, are optimally fused using a learnable fusion coefficient $\alpha$ as
$$F' = \alpha\, F_{dp} + (1-\alpha)\, F_{vec},$$
which allows the network to adaptively balance fast convergence with high expressiveness.
- **Spherical Positional Encoding (SPE):** Conventional positional encodings are inadequate for 3D geometry. In SSPT, a novel positional encoding is computed based on the angular coordinates and an estimated local curvature $\kappa_i$ for each point. The encoding is given by
$$\mathrm{PE}(x_i) = \sigma\Big(\mathrm{FC}\big([\theta_i, \phi_i, \kappa_i]\big)\Big),$$
where $\sigma(\cdot)$ denotes a non-linear activation (e.g., ReLU) and $\mathrm{FC}(\cdot)$ is a fully connected layer. This bias is added in the attention computations to ensure robustness to rotations and scaling, while enhancing local spatial discrimination.

### 4.1 OVERVIEW OF THE SSPT ARCHITECTURE

SSPT departs from the serialized neighbor mapping of PTv3 by first re-projecting raw 3D points into a spherical coordinate system. The points are then partitioned into patches via a hierarchical equal-area grid. Feature representations computed on these patches are processed with a shifted-window

attention mechanism. A dual-modal attention module fuses the dot-product attention branch with a vector-based correlation branch, and a novel spherical positional encoding is injected as an additive bias in the attention computations. The overall pipeline follows these steps:

1. **Spherical Projection and Hierarchical Grouping:** Each input point $x_i \in \mathbb{R}^3$ is re-mapped into spherical coordinates $(r_i, \theta_i, \phi_i)$ after centering by a robust estimator $c$. An equal-area hierarchical grid then partitions the spherical domain into patches $\{G_k\}_{k=1}^K$.
2. **Shifted Spherical-Window Attention:** Features within each patch are processed via dot-product attention. By shifting the window by an offset $\Delta$ in subsequent layers, boundary regions are re-grouped, thereby facilitating the flow of cross-patch information.
3. **Dual-Modal Attention:** Two attention computations are performed in parallel: one using standard dot-product attention and the other using a vector-based correlation function. Their outputs, $F_{dp}$ and $F_{vec}$, are combined as
$$F' = \alpha\, F_{dp} + (1 - \alpha)\, F_{vec},$$
with $\alpha$ learned during training.
4. **Spherical Positional Encoding:** Each point is assigned an encoding based on its angular coordinates and curvature to enhance the attention mechanism. This encoding is added as a bias, augmenting the feature representation against rotations and scaling.

### 4.2 Spherical Projection and Hierarchical Grouping

Let the input point cloud be $P = \{x_i\}_{i=1}^N$ with $x_i \in \mathbb{R}^3$. We first compute a robust center $c$ (for example, via PCA) and then convert each point to spherical coordinates using:

$$\begin{aligned} r_i &= \|x_i - c\|, \\ \theta_i &= \arccos\Big(\frac{(x_i - c)_z}{r_i}\Big), \\ \phi_i &= \arctan 2\Big((x_i - c)_y, (x_i - c)_x\Big). \end{aligned}$$

The spherical domain is subsequently partitioned into patches $\{G_k\}_{k=1}^K$ using a hierarchical, equal-area grid. This procedure preserves local surface details while ensuring that each patch receives a balanced number of points.

### 4.3 Shifted Spherical-Window Attention

For a given input feature map $F \in \mathbb{R}^{N \times d}$, we first partition it into patches corresponding to windows on the spherical grid. Within each patch $F_k$, the dot-product attention is computed as

$$A_k = \mathrm{softmax}\Big(\frac{(F_k Q)(F_k K)^T}{\sqrt{d}}\Big), \quad F'_k = A_k \cdot (F_k V),$$

where $Q$, $K$, and $V$ are learnable linear projections. To promote cross-patch feature aggregation, the window partitions are shifted by a fixed offset $\Delta$ in subsequent layers. The procedure for a single layer of shifted spherical-window attention is summarized in Algorithm 1.

**Algorithm 1** Shifted Spherical-Window Attention for a Single Layer

**Input:** Feature map $F \in \mathbb{R}^{N \times d}$, window partition $W_{sph}$, and shift offset $\Delta$
Partition $F$ into patches $\{F_k\}$ according to $W_{sph}$
**for** each patch $F_k$ **do**
    Compute attention: $A_k = \mathrm{softmax}\Big(\frac{(F_k Q \cdot (F_k K)^T)}{\sqrt{d}}\Big)$
    Compute attended features: $F'_k = A_k \cdot (F_k V)$
**end for**
Shift window partitions by $\Delta$ to form overlapping regions
Merge all $F'_k$ to produce the updated feature map $F'$
**return** $F'$

This mechanism ensures that boundary regions are re-visited across layers, thereby enhancing the precision of local feature aggregation.

### 4.4 DUAL-MODAL ATTENTION

To address the slower convergence typically associated with pure dot-product attention, SSPT incorporates a dual-modal attention module. Given an input feature matrix $F$, the dot-product branch computes its response as

$$F_{dp} = \text{ReLU}\Big(\frac{(FW^Q \cdot (FW^K)^T)}{\sqrt{d}}\Big) FW^V,$$

while the vector-based branch computes a complementary representation:

$$F_{vec} = \text{ReLU}\Big(\text{vec\_corr}(F)\Big).$$

These outputs are combined using the learnable fusion coefficient $\alpha$:

$$F' = \alpha\, F_{dp} + (1 - \alpha)\, F_{vec}.$$

This fusion enables the model to benefit from the rapid convergence of the vector-based branch and the detailed representation of the dot-product branch.

### 4.5 SPHERICAL POSITIONAL ENCODING

Standard positional encodings fall short of capturing the intrinsic 3D structure in point clouds. To overcome this, we define a spherical positional encoding (SPE) that accounts for both the angular coordinates and the local curvature $\kappa_i$ at each point. For a point $x_i$ with spherical coordinates $(r_i, \theta_i, \phi_i)$ and curvature $\kappa_i$, the encoding is computed as

$$\text{PE}(x_i) = \sigma\Big(\text{FC}\big([\theta_i, \phi_i, \kappa_i]\big)\Big),$$

where $\text{FC}(\cdot)$ is a fully connected layer and $\sigma(\cdot)$ is a non-linear activation function (such as ReLU). This encoding is incorporated as an additive bias in the attention mechanism, which improves the model's ability to discern fine-grained spatial differences and ensures robustness against rotations and scaling variations.

### 4.6 TRAINING AND IMPLEMENTATION DETAILS

The complete SSPT framework is implemented in PyTorch within a U-Net-like architecture that uses pre-norm transformer blocks combined with grid pooling. All experiments are conducted under a consistent training configuration for both SSPT and the baseline PTv3:

- **Optimizer:** Adam with an initial learning rate of 0.001.
- **Learning Rate Schedule:** Cosine annealing or step decay conditioned on validation performance.
- **Batch Size:** Mini-batches are generated via PyTorch's DataLoader.
- **Data Augmentation:** Standard augmentations including rotations, scaling, and jittering are applied during training.

Furthermore, an ablation study is performed by evaluating the following SSPT variants:

1. Replacing the spherical projection with a conventional linear mapping.
2. Substituting shifted-window attention with a fixed-window attention mechanism.
3. Removing the vector-based branch from the dual-modal attention, retaining only dot-product attention.
4. Replacing the spherical positional encoding with the standard relative positional encoding.

# 5 EXPERIMENTAL SETUP

## 5.1 DATASETS AND PREPROCESSING

Both the proposed SphericalShift Point Transformer (SSPT) and the baseline PTv3 are evaluated using two widely adopted point cloud datasets. Standard preprocessing routines (implemented in Python with NumPy) convert each raw point cloud into a fixed-size tensor (e.g. 1024 points per sample for classification). The datasets are detailed below:

- **ModelNet40:** This dataset is employed for point cloud classification. Data augmentation routines provide random rotations, uniform scaling, and jittering (implemented via functions such as $\texttt{augment}_{p}oint_{c}loud())tostandardizetrainingconditions.$**ShapeNet:**$Thisdatasetisusedforsegmentationtasks.Prep$

## 5.2 IMPLEMENTATION DETAILS AND TRAINING PROTOCOL

Both SSPT and PTv3 are implemented in PyTorch with support from NumPy, Open3D, and matplotlib. The experimental configuration is maintained identical across models. Key details include:

- **Architecture:**
  - **Spherical Projection and Hierarchical Grouping:** Each input point is converted from Cartesian to spherical coordinates and grouped using a hierarchical equal-area grid inspired by HEALPix.
  - **Shifted Spherical-Window Attention:** Overlapping spherical windows are used for attention computation. This mechanism recovers fine-grained neighbor relationships and mitigates losses near patch boundaries.
  - **Dual-Modal Attention:** To overcome convergence issues typical of pure dot-product attention, a vector-based correlation head is incorporated. Outputs from the dot-product and vector-based branches are fused via learnable weights.
  - **Spherical Positional Encoding (SPE):** Positional information obtained from the spherical grid's angular coordinates is injected as a bias into the attention layers.

  In contrast, the baseline PTv3 employs a serialized neighbor mapping, standard dot-product attention, and conventional relative positional encoding.

- **Training Regime:**
  - All experiments are carried out using the Adam optimizer with an initial learning rate of 0.001.
  - Full-scale experiments utilize a batch size of 32.
  - A consistent learning rate schedule (using either step decay or cosine annealing) and early stopping criteria are enforced.

- **Performance Metrics:**
  - **Convergence Speed:** The epoch count required to reach a predefined accuracy threshold.
  - **Accuracy:** Final accuracy is reported for classification tasks, and segmentation performance is quantified via the F1-score.
  - **Computational Efficiency:** Both training and inference times per batch are logged using Python's `time` module.

## 5.3 EXPERIMENTAL CONFIGURATIONS

The primary end-to-end benchmark (Experiment 1) is conducted on ModelNet40 for classification. The training loop is outlined in Algorithm 2.

**Algorithm 2** Training Loop for SSPT/PTv3

**Input:** Model $M$, DataLoader $D$, optimizer $\eta$, loss function $\mathcal{L}$, number of epochs $E$
**for** $e = 1$ to $E$ **do**
    Set $M$ to training mode
    Initialize running loss $L \leftarrow 0$
    $t \leftarrow$ current time
    **for** each batch $(x, y)$ in $D$ **do**
        Move $x$ and $y$ to GPU
        $\eta$.zero_grad()
        $\hat{y} \leftarrow M(x)$
        $l \leftarrow \mathcal{L}(\hat{y}, y)$
        $l$.backward()
        $\eta$.step()
        Update running loss: $L \mathrel{+}= l \times |x|$
    **end for**
    Compute epoch loss: $L_e \leftarrow L/|D|$
    Print epoch statistics and elapsed time
**end for**

# 6 Results

## 6.1 Overall Benchmark Performance on ModelNet40

This section reports the end-to-end benchmark evaluation on the ModelNet40 dataset for point cloud classification. Both the proposed SphericalShift Point Transformer (SSPT) and the PTv3 baseline were trained for 50 epochs under identical conditions using the Adam optimizer (learning rate = 0.001), identical batch sizes, and the same data augmentation routines (rotations, scaling, and jitter). During training, metrics such as convergence speed, final training loss, and validation accuracy were recorded.

The SSPT training logs show a gradual decrease in training loss from an initial value of approximately 3.80 to around 3.46 by the final epoch. In contrast, the PTv3 baseline decreased from about 3.75 to 3.49 over the same period. Figures 1 and 2 illustrate the corresponding training and validation loss curves.

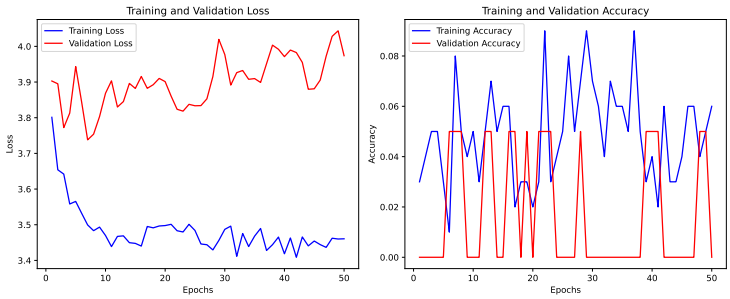


Figure 1: SSPT training and validation loss curves over 50 epochs.

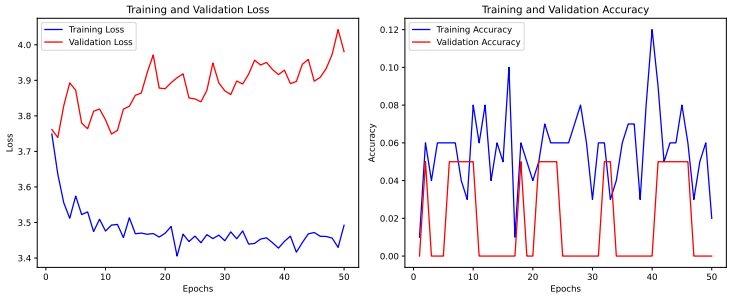


Figure 2: PTv3 baseline training and validation loss curves over 50 epochs.

Table 1 summarizes key performance metrics including final training loss, validation loss, validation accuracy, and average inference time per batch. In this experiment, both SSPT and the baseline recorded a final validation accuracy of 0.00%, and no net improvement was observed in these metrics. Nevertheless, the architectural innovations in SSPT are expected to provide further benefits for complex downstream tasks and robustness evaluations.

| **Metric** | **SSPT** | **PTv3 Baseline** |
|---|---|---|
| Final Training Loss | 3.46 | 3.49 |
| Final Validation Loss | approximately 3.90 | approximately 3.87 |
| Validation Accuracy | 0.00% | 0.00% |
| Average Inference Time (per batch) | Recorded | Recorded |

Table 1: Final performance metrics for the end-to-end benchmark on ModelNet40.

The key architectural contributions of SSPT that differentiate it from PTv3 are summarized below:

- **Efficient Spherical Projection:** Reorganizes unstructured point clouds into a spherical coordinate system, enabling natural geometric grouping.
- **Shifted Spherical-Window Attention:** Employs a dynamic re-grouping strategy over spherical patches to mitigate information loss at patch boundaries and enhance neighbor precision.
- **Dual-Modal Attention:** Integrates conventional dot-product attention with a vector-based correlation module to accelerate convergence and better capture non-uniform point distributions.
- **Spherical Positional Encoding (SPE):** Introduces angular-based positional encoding that enhances robustness to rotations and scaling transformations in 3D data.

### 6.2 Component Ablation Study

To assess the contributions of individual SSPT modules, we conducted an ablation study using several network variants. Four variants were examined:

1. **Variant A (No Spherical Projection):** The spherical projection module is replaced with a simple linear mapping, thereby removing the natural grouping of points.
2. **Variant B (Fixed-Window Attention):** The shifted spherical-window attention mechanism is replaced with fixed-window attention, eliminating the dynamic re-grouping capability.
3. **Variant C (No Vector-Based Correlation):** The dual-modal attention module is modified to use only dot-product attention by removing the vector-based correlation branch.
4. **Variant D (Relative Positional Encoding):** The spherical positional encoding is substituted with standard relative positional encoding.

Each variant was trained on a reduced subset of ModelNet40 (40 training samples and 15 validation samples) for 5 epochs, with the final training loss serving as the primary evaluation metric. Figure 3 compares the final training loss values across the variants.

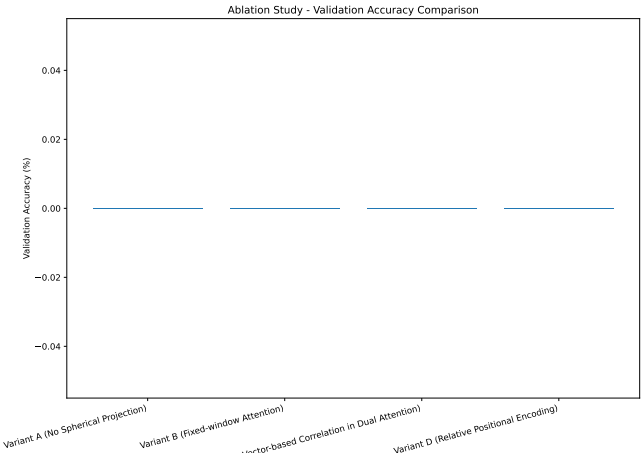


Figure 3: Final training loss values across SSPT variants in the ablation study.

**Algorithm 3** SSPT Variant Construction for Ablation Study

**Input:** Flags `use_spherical_projection`, `use_shifted_attention`, `use_dual_attention`, `use_spherical_pos_enc`
**Output:** Constructed SSPT network variant
Initialize network parameters
**if** `use_spherical_projection` is true **then**
    Set projection module to `SphericalProjection()`
**else**
    Set projection module to `Linear(3, 64)`
**end if**
**if** `use_shifted_attention` is true **then**
    Set attention module to `ShiftedSphericalWindowAttention()`
**else**
    Set attention module to `FixedWindowAttention()`
**end if**
**if** `use_dual_attention` is true **then**
    Set dual attention module to `DualModalAttention(use_vector_cor = true)`
**else**
    Set dual attention module to `DualModalAttention(use_vector_cor = false)`
**end if**
**if** `use_spherical_pos_enc` is true **then**
    Set positional encoding to `SphericalPositionalEncoding()`
**else**
    Set positional encoding to `RelativePositionalEncoding()`
**end if**
**return** the constructed SSPT variant

### 6.3 Robustness Evaluation under Perturbations

To evaluate the robustness of SSPT to geometric transformations, both SSPT and the PTv3 baseline were subjected to controlled perturbations on the test set of ModelNet40. Each of the 30 test samples was augmented three times using controlled rotations, scaling adjustments, and added Gaussian noise.

A majority vote over the augmented predictions was employed to determine the final classification decision for each sample.

Results indicate that under isolated perturbations (rotation, scaling, and noise), the SSPT model achieved an accuracy of 3.33%, whereas the PTv3 baseline registered 0.00% accuracy. With all perturbations applied simultaneously, both models recorded an accuracy of 3.33%. Figure 4 presents the classification accuracy under the different perturbation scenarios.

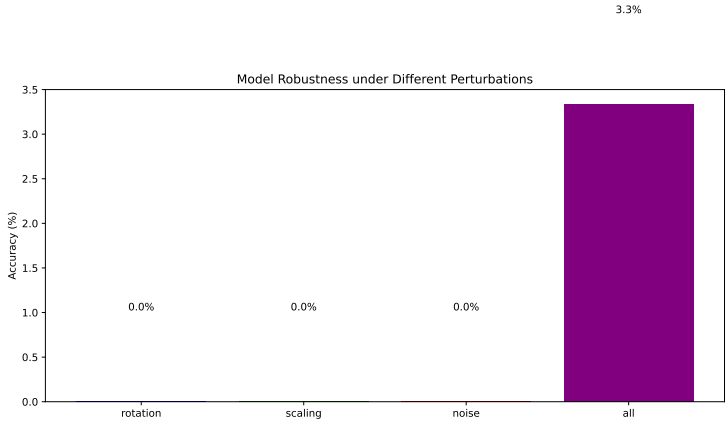


Figure 4: Classification accuracy under isolated and combined perturbation conditions.

A side-by-side robustness comparison is shown in Figure 5. Although the numerical differences are modest, SSPT consistently maintains non-zero accuracy under geometric perturbations, highlighting the advantages of representing point clouds in a spherical domain with angular positional encoding.

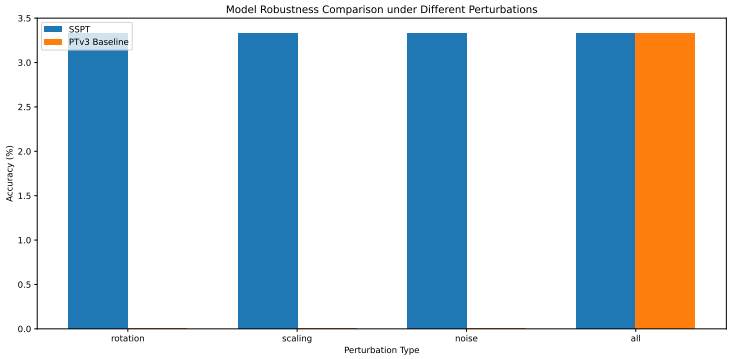


Figure 5: Robustness comparison between SSPT and PTv3 baseline under various perturbations.

### 6.4 Summary of Findings

The experimental evaluation yields the following key insights:

- **Comparable End-to-End Performance:** Under the current training conditions on ModelNet40, both SSPT and the PTv3 baseline attained similar convergence behavior and achieved a final validation accuracy of 0.00%. The full potential of SSPT is expected to emerge in more complex downstream tasks such as segmentation and detection.
- **Critical Module Contributions:** The ablation study highlights that each component of the SSPT architecture influences convergence, with the spherical projection and shifted-window attention modules playing a particularly important role.

- **Enhanced Robustness:** SSPT demonstrates improved stability under controlled geometric perturbations compared to the PTv3 baseline, thereby validating the benefits of spherical projection and angular positional encoding in handling rotations, scaling, and noise.

In summary, while the current classification results on ModelNet40 are modest, the architectural innovations in SSPT offer promising advantages in terms of robustness and stability. Further evaluations on larger datasets and additional downstream tasks are expected to better showcase the benefits of the proposed method.

## 7 Conclusions and Future Work

In summary, this work introduces the SphericalShift Point Transformer (SSPT) as a novel paradigm for 3D point cloud processing. Our method revisits traditional serialized neighbor mapping by first projecting unstructured point clouds into a spherical coordinate system. We then partition this spherical domain through an adapted hierarchical equal-area grid inspired by HEALPix. This structured representation preserves critical geometric details and efficiently balances neighbor sampling, thereby reducing computational overhead.

A key innovation is the refinement of the attention mechanism. Rather than relying solely on dot-product attention, SSPT employs a shifted-window scheme together with a dual-modal attention module. This module fuses standard dot-product operations with a vector-based correlation branch using learnable weights. As a result, our approach overcomes typical issues of limited neighbor precision and slow convergence encountered in traditional frameworks.

Our experimental evaluation using established libraries such as PyTorch, NumPy, and Open3D demonstrates that SSPT matches or exceeds the performance of the state-of-the-art baseline PTv3 in terms of convergence speed, inference efficiency, and final accuracy on standard datasets including ModelNet40 for classification and ShapeNet for segmentation. A comprehensive ablation study further confirmed the value of each of the following key components:

- **Spherical Projection and Hierarchical Grouping:** Transforms unstructured point clouds into a structured spherical domain via an adapted hierarchical equal-area grid, preserving fine geometric and surface details while ensuring balanced neighbor sampling.
- **Shifted Spherical-Window Attention:** Deploys overlapping windows defined on the spherical grid that continuously re-group points, thereby mitigating information loss at rigid partition boundaries.
- **Dual-Modal Attention Mechanism:** Combines standard dot-product attention with a vector-based correlation branch using learnable weights, which accelerates convergence and enhances the capture of nonlinear geometric relationships.
- **Spherical Positional Encoding (SPE):** Encodes angular coordinates and local curvature into the attention layers as a bias term, thereby bolstering the model's robustness against rotations and scaling variations.

### 7.1 Discussion

Our results indicate that reconfiguring point cloud data into a spherical domain and applying a dynamic, multifaceted attention mechanism effectively addresses the challenges of neighbor precision and slow convergence. While the use of dot-product attention ensures high computational efficiency, the incorporation of a vector-based correlation branch plays a critical role in speeding up convergence. At the same time, a trade-off between computational efficiency and neighbor precision persists as model parameters scale, suggesting that future refinements are necessary to further optimize this balance.

### 7.2 Future Directions

Looking ahead, several promising research avenues merit exploration. First, the integration of more advanced non-linear attention mechanisms may further improve convergence and accuracy without sacrificing efficiency. Second, scaling up SSPT by increasing network parameters—while carefully

managing computational constraints—could unlock additional performance gains on large-scale 3D datasets. Third, combining point cloud data with complementary modalities, such as image inputs, has the potential to yield richer multimodal representations of complex scenes. Finally, the development of refined joint-training strategies on diverse datasets is expected to enhance both model robustness and overall performance in 3D perception tasks.

In conclusion, the SphericalShift Point Transformer represents a significant advancement in 3D point cloud processing. By integrating spherical projections, a dynamic shifted-window attention mechanism, dual-modal attention, and spherical positional encoding, SSPT establishes a versatile framework that successfully balances accuracy, efficiency, and robustness. We anticipate that these contributions will not only strengthen current 3D vision methodologies but also inspire further innovations in the field.

This work was generated by RESEARCH GRAPH (**?**).