Spatial-Temporal Transformer for 3D Point Cloud Sequences

Yimin Wei\textsuperscript{1,2}, Hao Liu\textsuperscript{1,2}, Tingting Xie\textsuperscript{3}, Qiuhong Ke\textsuperscript{4}, Yulan Guo\textsuperscript{1,2,*}
\textsuperscript{1}Sun Yat-sen University, \textsuperscript{2}Shenzhen Campus of Sun Yat-sen University, \textsuperscript{3}Queen Mary University of London, \textsuperscript{4}The University of Melbourne
\texttt{weiym9@mail2.sysu.edu.cn, guoyulan@sysu.edu.cn}

Abstract

Effective learning of spatial-temporal information within a point cloud sequence is highly important for many downstream tasks such as 4D semantic segmentation and 3D action recognition. In this paper, we propose a novel framework named Point Spatial-Temporal Transformer (\textsc{PST}\textsuperscript{2}) to learn spatial-temporal representations from dynamic 3D point cloud sequences. Our \textsc{PST}\textsuperscript{2} consists of two major modules: a Spatio-Temporal Self-Attention (STSA) module and a Resolution Embedding (RE) module. Our STSA module is introduced to capture the spatial-temporal context information across adjacent frames, while the RE module is proposed to aggregate features across neighbors to enhance the resolution of feature maps. We test the effectiveness our \textsc{PST}\textsuperscript{2} with two different tasks on point cloud sequences, i.e., 4D semantic segmentation and 3D action recognition. Extensive experiments on three benchmarks show that our \textsc{PST}\textsuperscript{2} outperforms existing methods on all datasets. The effectiveness of our STSA and RE modules have also been justified with ablation experiments.

1. Introduction

Point cloud sequences provide both spatial and temporal information of a scene over a period and can be applied in various vision tasks such as 4D semantic segmentation and 3D action recognition. With the rapid development of dynamic point cloud acquisition technologies (such as LiDARs and depth cameras), point cloud sequences are becoming increasingly available and have been investigated for different applications, such as self-driving vehicles, robotics, and augmented reality.

4D semantic segmentation is challenging due to several reasons. First, different from regular and dense 2D image sequences, raw point cloud sequences usually have a regular order in the temporal domain but are unordered in the spatial domain. Besides, point clouds are usually highly sparse. Therefore, it is extremely difficult to model the spatio-temporal structure in a point cloud sequence. Second, it is hard to predict the motion of points within a point cloud sequence. However, the construction of spatio-temporal neighborhoods rely on tracking the motion of points across different frames. It is therefore, difficult to effectively extract and aggregate the spatio-temporal context information of points within inter-frame neighborhoods.

Recently, several methods have been proposed to capture the spatio-temporal dynamics of points from point cloud sequences by utilizing cross-frame information. These methods can be roughly divided into voxel-based methods [5, 15] and point-based methods [20, 2, 19, 8]. A voxel-based method converts a 3D point cloud sequence into regular sparse voxels and then extracts point-wise features via a network with 3D convolutions or sparse convolutions. Due to the introduction of quantization error, the performance of this approach is limited. A point-based method constructs a spatial-temporal neighborhood for each seed point (e.g., sampled by Farthest Point Sampling (FPS)), and then applies Multi-Layer Perceptron (MLP) to extract point-wise features. It then passes these features backward and aggregates them with the features of subsequent frames using an inter-frame feature fusion strategy.

Although existing point-based methods can avoid the quantization error by directly processing raw point cloud sequences, they still face two major challenges. First, existing methods usually adopt attention operations [2, 19] or Recurrent Neural Network (RNN) models [15, 8] to fuse inter-frame features. However, since these methods rely on long-term dependency, the inter-frame information fusion conducted at a frame depends on the fusion of the information of all its previous frames, leading to information redundancy. It is necessary for an inter-frame feature fusion strategy to capture spatio-temporal point context information while reducing the redundancy of previous frames. Besides, current semantic segmentation methods mainly adopt an encoder-decoder architecture. The encoder usually consists of multiple hierarchically stacked feature extraction layers (e.g., the set abstraction layer [24]) and extracts rich semantic features by reducing the resolution of its feature maps. Therefore, information loss is introduced and the
The architecture of our PST$^2$ model. Our PST$^2$ model is employed for two tasks: 4D semantic segmentation (as shown in (a)) and 3D action recognition (as shown in (b)). In the 4D semantic segmentation task, we apply two set abstraction layers to extract features, and then employ the Resolution Embedding (RE) module to recover the lost geometric information of the extracted features. Next, we perform convolution on the features of adjacent frames to obtain spatial-temporal patches, and then fed these patches into a Spatio-Temporal Self-Attention (STSA) module to capture spatial-temporal context information. Finally, we use multiple Feature Propagation (FP) layers to produce point-wise semantic predictions. In the 3D action recognition task, we insert our STSA module into the MeteorNet [20] model to achieve action recognition.

segmentation performance is decreased.

In this paper, we propose a novel Point Spatial-Temporal Transformer (PST$^2$) network to tackle the above two challenges. First, we introduce a self-attention based module, i.e., Spatio-Temporal Self-Attention (STSA), to capture inter-frame spatial-temporal context information. More specifically, our PST$^2$ model adaptively aggregates these correlated inter-frame neighborhood features based on the self-attention operation instead of fusing all previous frames. Consequently, the redundancy is reduced, the training speed is increased, and the robustness is improved (with residual connection and layer normalization). Second, we introduce a Resolution Embedding (RE) module to enhance the resolution (which is lost in the encoder-decoder architecture). By aggregating inter-neighborhood features with the attention weights, the resolution can be enhanced and the segmentation performance can be improved significantly. To further demonstrate the effectiveness of our point spatial-temporal transformer, we introduce a new framework to leverage the proposed STSA module for action recognition from 3D point clouds. Experiments on the Synthia [25], SemanticKITTI [1], and MSR-Action3D [18] datasets demonstrate that our PST$^2$ model outperforms existing methods [20, 2].

The contributions of this paper are summarized as follows:

1) We propose a self-attention based module (namely, STSA) to capture point dynamics in point cloud sequences. It is shown that this module can encode the spatio-temporal features from point cloud sequences.

2) We introduce an RE module to enhance the resolution of feature maps. This module clearly improves the segmentation performance.

3) Our PST$^2$ model is tested with two different tasks: 4D semantic segmentation and 3D action recognition. Experiments on three datasets show that the proposed STSA and RE modules can easily be plugged into exiting static point cloud processing pipelines to achieve improved performance.

2. Related Work

In this section, we briefly review those works that are highly related to our method, i.e., deep learning method on static point clouds, deep learning method on 3D point cloud sequences, and the transformer.

2.1. Deep Learning on Static Point Clouds

Deep learning has been widely used in many point cloud tasks, such as classification [23, 24, 13], part segmentation [37, 14, 36], semantic segmentation [35, 17, 31, 12], reconstruction [6, 39], and object detection [4, 22, 11]. Existing methods for deep learning on static point clouds can be roughly divided into multi-view based, volumetric-based, and point-based methods. Multi-view based meth-
ods [27, 34] first project a point cloud into multiple views and extract view-wise features, and then fuse these features for static point cloud processing. Volumetric-based methods [21, 32] convert a point cloud into regular 3D voxels and then employ 3D Convolution Neural Network (CNN) or sparse convolution on the volumetric representation. These two approaches introduce quantization errors and exhibit performance declination, while point-based methods directly work on raw point clouds and reduce the quantization error.

PointNet [23] is a pioneering work of point-based methods. Qi et al. [23] used several MLP layers to extract point-wise local features and adopted a max-pooling layer to aggregate these features. The local features and the aggregated global features are then concatenated to predict the category of each point. In follow-up work, Qi et al. [24] introduced a hierarchical structure and multi-scale feature learning mechanism into PointNet++, thereby improving the performance of scene semantic segmentation. Subsequent works have been proposed based on the improvement of PointNet++. However, these methods cannot directly process dynamic point cloud sequences due to the lack of the inter-frame feature fusion strategy.

2.2. Deep Learning on 3D Point Cloud Sequences

In recent years, two major mainstream methods focusing on 3D dynamic point cloud sequences have been explored.

Several voxelization-based methods have been proposed to handle dynamic point cloud sequences. MinkowskiNet et al. [5] first converted 3D point cloud sequences into 4D occupancy grids, and then employed sparse 4D convolution to process 4D occupancy grids. In 3D Sparse Conv LSTM, Huang et al. [15] voxelized 3D point cloud sequences into sparse 4D voxels and convolved them on a sparse grid, then passed the memory and hidden features through a sparse convolution inside the LSTM network. The LSTM network produces hidden and memory features that will be passed to the next frame and fused with features of subsequent frames.

Several methods directly working on raw point cloud sequences have also been explored. MeteorNet [20] is a leading deep learning method for dynamic 3D point cloud sequences. The structure of MeteorNet [20] is inherited from PointNet [23]. By proposing two grouping approaches (including direct grouping and chained-flow grouping) to construct spatio-temporal neighborhoods, inter-neighborhood features are aggregated to capture point dynamics in the point cloud sequence. Cao et al. [2] proposed an Attention and Structure Aware (ASAP) model, which contains a novel grouping approach (i.e., Spatio-Temporal Correlation) to construct spatio-temporal neighborhoods and a novel attentive temporal embedding layer to fuse the related inter-frame local features in a recurrent fashion. Our Point Spatial-Temporal Transformer PST2 is also directly performed on raw point cloud sequences, but aims at enhancing the resolution when capturing spatio-temporal dynamics across entire point cloud sequences.

2.3. Transformer

Transformer is a typical deep neural network based on the self-attention mechanism and particularly suitable for modeling long-range dependencies.

Transformer was first proposed in [28] for the sequence-to-sequence machine translation task, and then extended to various computer vision tasks. In computer vision tasks, image features are obtained from a backbone network and converted into sequences, and then a transformer-based network is employed to process sequence-form features. DETR Transformer (DETR) [3] is a landmark work that considers object detection as a set prediction problem. Specifically, an image is first divided into several candidate regions, and then the object in candidate regions is classified. Recently, transformer has been used in various computer vision applications, including image classification, high/mid-level vision, low-level vision, and video processing. In these applications, Transformer excels to process video sequences and has extended many video tasks, such as video retrieval [26], action recognition [9], and video object detection [38].

Very recently, Transformer is extended from video tasks to point cloud tasks with promising results. Guo et al. [40] proposed a novel transformer-based framework (namely, Point Cloud Transformer (PCT)) for static point cloud learning. They used self-attention to design an offset-attention layer to model irregular point clouds. Zhao et al. [10] proposed a Point Transformer (PT) layer to fuse point cloud features.

However, there are few prior Transformer applications to model point cloud sequences. Intuitively, Transformer has the advantage of modeling long-range dependencies and should be an ideal candidate for point cloud sequences tasks. Therefore, we propose the PST2 model to model spatial-temporal information in point cloud sequences with Transformer.

Our PST2 model is different from the transformers designed in PCT [10] and PT [40]. First, our STSA module introduces skip connection and layer normalization operations based on the self-attention operation. Besides, our STSA module divides the input features of the adjacent frame into spatio-temporal patches for further processing, while the input feature is directly processed in [10, 40]. Consequently, our module is better at capturing the spatial-
temporal context information than [10, 40].

3. Point Spatial-Temporal Transformer $PST^2$

In this section, we first use 4D semantic segmentation as an example to describe our Point Spatial-Temporal Transformer network $PST^2$, including the overall architecture, the step for spatial-temporal neighborhood construction, the Resolution Embedding (RE) module, and the Spatio-Temporal Self-Attention (STSA) module. We then adapt the proposed module for 3D action recognition.

3.1. 4D Semantic Segmentation

3.1.1 Overview Architecture

For 4D semantic segmentation, which is a point-level classification task, our $PST^2$ model takes a multi-frame point cloud sequence as its input, and predicts the category of each point in the point cloud sequence. Our $PST^2$ model adopts an encoder-decoder architecture. Specifically, the encoder consists of a backbone, an RE module and an STSA module, and the decoder includes multiple Feature Propagation (FP) layers. The FP layer is originally introduced in PointNet++ [24] to interpolate the semantic features extracted by the encoder upwards and predict the semantic category of each point.

As illustrated in Fig. 1 (a), given a point cloud sequence, three steps are adopted to capture point dynamics in our $PST^2$ model. First, we use a backbone network (including two set abstraction layers [24]) to construct the spatial-temporal neighborhoods and extract point features in each frame (Sec. 3.1.2). Then, we adopt the proposed RE module to enhance the resolution of features in each frame (Sec. 3.1.3). Finally, we employ the proposed STSA module to fuse the inter-frame features and produce the predictions (Sec. 3.1.3).

3.1.2 Spatial-Temporal Neighborhood Construction

Given a point cloud sequence $\{S_1, S_2, ..., S_T\}$ of length $T$, each frame of the point cloud is represented as $S_t = \{p^{(t)}_i\}_{i=1}^n$, where $n$ is the number of points. Each point $p^{(t)}_i$ consists of its Euclidean coordinates $x^{(t)}_i \in \mathbb{R}^3$ and hand-crafted feature $f^{(t)}_i$. A number of seed points $\{c^{(t)}_j\}_{j=1}^m$ can be sampled from $S_t$ by the FPS method [23, 24], where $m$ is the number of seed points and $m < n$.

We first sample the seed points in the first frame $S_1$, and groups all points within a certain radius around these seed points to form a neighborhood. Next, we use the same seed points in subsequent frames $\{S_2, ..., S_T\}$ to construct spatial-temporal neighborhoods with stable spatial locations at different times. Then, we apply two set abstraction layers [24] to extract local features $h^{(t)}_i \in \mathbb{R}^{s \times d}$ from the constructed neighborhood, where $s$ and $d$ are the sizes of the spatial dimension and the feature dimension, respectively.

![Figure 2. The architecture of our RE module.](image)

3.1.3 Resolution Embedding (RE)

Given the feature representation $h^{(t)}_i$, we propose a novel resolution block in the RE module to enhance the resolution of point features in each frame. The RE module consists of a feature block and a resolution block. The feature block was designed to further extract semantic features, while the resolution block extracts the spatial context information from the input features, and treats these context information as an enhanced resolution to incorporate with the features produced by the feature block.

(1) Feature block: As shown in the Fig. 2(a), we perform a set abstraction layer on $h^{(t)}_i$ to obtain $n^{(t)}_i$. The set abstraction layer is used in the encoder-decoder architecture to hierarchically extract local (even global) features.

(2) Resolution block: As shown in Fig. 2(b), we divide $h^{(t)}_i$ into $m^{(t)}_1$ and $m^{(t)}_2$ along the spatial dimension, and then concatenate them along the feature dimension to obtain $g^{(t)}_i$. We employ a learnable multi-layer perception $f$ to update the feature dimension of $g^{(t)}_i$, resulting in $k^{(t)}_i$. This process aims at effectively extracting the spatial inter-neighborhood related information.

Once $n^{(t)}_i$ and $k^{(t)}_i$ are obtained, we calculate two scalar attentions $[a_1, a_2]$ according to the correlation of $n^{(t)}_i$ and $k^{(t)}_i$. By feeding $n^{(t)}_i$ and $k^{(t)}_i$ into a shared MLP function $\gamma$ and a Softmax function, a weighted-sum feature $I^{(t)}_i$ is calculated based on these two attentions:

$$I^{(t)}_i = a_1 \cdot k^{(t)}_i + a_2 \cdot n^{(t)}_i,$$

where the attention weights $a_1$ and $a_2$ are calculated as:

$$[a_1, a_2] = \text{Softmax}(\gamma(k^{(t)}_i, n^{(t)}_i)).$$
Given spatial features $I_i(t)$ of each frame, the STSA module is used to incorporate inter-frame features along the temporal dimension and capture the spatio-temporal context information. Intuitively, different inter-frame spatial-temporal neighbors contribute differently to the final results. Specifically, for objects with slow-moving speeds, inter-frame spatial-temporal neighbors that are close in space contributed more than those neighbors that are far apart. In contrast, for objects with fast-moving speeds, inter-frame spatial-temporal neighbors that are far in space should also be considered. Transformer-based methods have the potential to learn the correlation within inter-frame spatial-temporal neighborhoods.

As shown in Fig. 3, our STSA module consists of two blocks: spatial-temporal patch division, and self-attention.

(1) **Spatial-Temporal Patch Division**

This block is designed to correlate multiple frame features and generate spatial-temporal patches. Inspired by [7], we first divide spatial features $I_i(t)$ into patches and then align these inter-frame spatial patches along the temporal domain based on the same seed points to form spatial-temporal patches. These spatial-temporal patches $F_{input}$ are adjacent along the temporal dimension, and neighborhoods in each patch is consistent along the spatial dimension.

(2) **Self-Attention**

The self-attention mechanism measures intra-term semantic affinities within a data sequence by mapping a query and a set of key-value pairs. Following this mechanism, the self-attention block is designed to learn the similarities of the above spatio-temporal patches. Given query matrix $Q$, key matrix $K$, and value matrix $V$, the query, key and value are generated by linear transformations based on the input features $F_{input} \in \mathbb{R}^{N \times d}$:

$$ (Q, K, V) = F_{input} \cdot (W_q, W_k, W_v) $$ (3)

where $Q, K, V \in \mathbb{R}^{N \times d}, W_q, W_k, W_v \in \mathbb{R}^{d \times d}$ are the shared learnable linear transformation implemented by an MLP function, and $d$ is the dimension of the query and key vectors.

We then use the query and key matrices to calculate the attention weights $A_1$:

$$ A_1 = (a_1)_{i,j} = Q \cdot K^T $$ (4)

These weights are scaled by $\sqrt{d}$ to produce $A_2 = (a_2)_{i,j}$:

$$ A_2 = (a_2)_{i,j} = \frac{(a_1)_{i,j}}{\sqrt{d}} $$ (5)

$A_2$ is further then normalized to produce $A_3 = (a_3)_{i,j}$:

$$ A_3 = (a_3)_{i,j} = \text{softmax}(a_2)_{i,j} = \frac{\exp[(a_2)_{i,j}]}{\sum_k \exp[(a_2)_{i,k}]} $$ (6)

Then, the features $F_{sa\_out}$ with self-attention are obtained as the weighted sums of the value vector using their attention weights:

$$ F_{sa\_out} = A_3 \cdot V $$ (7)

Finally, the features produced by self-attention $F_{sa\_out}$ and the input features $F_{input}$ are further concatenated to obtain the feature $F_{output}$ through residual connection, a feed-forward network $\text{FeedForward}()$, and a layer normalization function $\text{LayerNorm}()$:

$$ F_{output} = \text{LayerNorm} [\text{FeedForward}(F_{sa\_out} + F_{input})] $$ (8)

The self-attention operation introduces randomness in the initial weights and tends to make the output features different from the input feature semantics. To reduce the influence of this randomness and accelerate the model fitting process, skip connection is used to incorporate the input feature and the output feature. Moreover, the difficulty to converge is a generally recognized problem with the original transformer [28]. Layer normalization is employed to accelerate the model training speed and to improve the robustness of the model by normalizing the output features.

### 3.2. Adaption to 3D Action Recognition

To further evaluate the ability of PST$^2$ for modeling dynamic point cloud sequences, we also employ PST$^2$ to perform 3D action recognition.

In a sequence-level classification task (e.g., action recognition), our PST$^2$ model takes a multi-frame point cloud...
Table 1. The main hyper-parameters we adopted in the experiments.

<table>
<thead>
<tr>
<th>Experiment</th>
<th>learning rate</th>
<th>nframe</th>
<th>batch size</th>
<th>point number</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSR-Action3D</td>
<td>0.001</td>
<td>4</td>
<td>16</td>
<td>2048</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>8</td>
<td>8</td>
<td>8192</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>12</td>
<td>8</td>
<td>8192</td>
</tr>
<tr>
<td></td>
<td>0.001</td>
<td>16</td>
<td>8</td>
<td>10240</td>
</tr>
<tr>
<td>Synthia</td>
<td>0.0016</td>
<td>3</td>
<td>2</td>
<td>16384</td>
</tr>
<tr>
<td>SemanticKITTI</td>
<td>0.012</td>
<td>3</td>
<td>2</td>
<td>16384</td>
</tr>
</tbody>
</table>

sequence as its input, and predicts the label of the sequence (e.g., action category). Since the RE module is specifically designed to enhance the resolution of a segmentation network (implemented with an encoder-decoder structure), we remove the RE module in the classification task. In addition, to intuitively show the effectiveness and easy portability of our STSA module, we improve an existing model MeteorNet [20] with our STSA module for action recognition. As shown in Fig. 1(b), we insert the STSA module into the encoding layer of MeteorNet. The remaining settings are the same as those in [20]. Note that, the meteor module proposed in [20] adopts a direct grouping method to roughly group the points in a spatial-temporal neighborhood and further extracts features, which ignore the correlation between these spatial-temporal neighbors. After inserting our STSA module, more spatial-temporal context information can be obtained and performance can be improved. The effectiveness of the STSA module for a long sequence is evaluated with experimental comparison (see Sec. 4.4).

4. Experiments

In this section, we first introduce the Synthia [25], SemanticKITTI [1], and MSRAction3D [18] datasets and their experimental settings in Sec. 4.1. Then, we report our results on these datasets and compare our model with several existing methods (Secs. 4.2 ~ 4.4). Finally, we provide ablation analyses on two spatial-temporal correlation strategies (i.e., STSA and RE) proposed in this paper and justify their contributions in Section 4.5.

4.1. Datasets and Settings

We evaluate our PST² model on the Synthia and SemanticKITTI datasets for 4D semantic segmentation, and the MSRAAction3D dataset for 3D action recognition.

4.1.1 4D Semantic Segmentation

We conducted two 4D semantic segmentation experiments on the large-scale synthetic dataset Synthia [25] and the largest publicly available real LiDAR dataset SemanticKITTI [1]. Then, we compare our PST² model with the MeteorNet [20], ASAP-Net [2], and MinkNet [5] baselines and perform an ablation study on the SemanticKITTI dataset.

Synthia The dataset consists of six sequences of driving scenarios in nine different weather conditions. Each sequence includes 4 stereo RGB-D images from four viewpoints captured from the top of a moving car. Following [20], we reconstruct 3D point clouds from RGB and depth images, and then employ the FPS method to downsample the point cloud to 16,384 points per frame. We employ the mean Intersection over Union (mIoU) and the mean Accuracy (mAcc) as the evaluation metrics in our experiments.

SemanticKITTI The dataset provides 23,201 frames for training and 20,351 frames for testing. To compare our method with the state-of-the-art method ASAP-Net [2], we follow the multi-frame experiments proposed in [2]. We only use the mean Intersection over Union (mIoU) as the evaluation metric in our experiments.

4.1.2 3D Action Recognition

MSR-Action3D The MSR-Action3D dataset consists of 567 depth map sequences acquired by Kinect v1. The depth map sequences include 10 different people, 20 categories of actions, and 23,797 frames. We reconstruct point cloud sequences from these depth maps, and use the same train/test split as previous work [20, 30]. The overall accuracy is used as the evaluation metric in this experiments.

4.1.3 Implementation Details

The main hyper-parameters adopted in the experiments are reported in Table 1. In our PST² model, we set the number of sampling points in the stacked feature encoding layers as [2048, 512, 128, 64] in the Synthia experiment. Other implementation details, including those for the experiments on Synthia and SemanticKITTI, are set the same as in ASAP-Net [2]. For 3D action recognition, we improve the MeteorNet [20] baseline with our STSA module, while other settings of the network are the same as MeteorNet [20]. Moreover, we follow the implementation of FPS provided by PointNet++ [24], which is implemented in GPU. All experiments are conducted on an NVIDIA RTX 2080Ti GPU using the ADAM optimizer.

4.2. Evaluation on Synthia

Table 2 reports the semantic segmentation results on the Synthia dataset. Our PST² consistently outperforms MinkowskiNet [5], MeteorNet [20], and ASAP-Net [2]. It establishes a new state-of-the-art result with a mIoU of 81.86 and a mAcc of 87.03. Further, our method achieves the best performance under 5 out of 12 categories. This is mainly caused by the fact that our PST² model can capture
Table 2. Semantic segmentation results on the Sythia dataset. Mean accuracy and mean IoU (%) are used as the evaluation metrics.

<table>
<thead>
<tr>
<th>Method</th>
<th>param (M)</th>
<th>nframe</th>
<th>mAcc</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>3D MiniNet [5]</td>
<td>19.31</td>
<td>1</td>
<td>89.31</td>
<td>76.24</td>
</tr>
<tr>
<td>4D MiniNet [5]</td>
<td>23.72</td>
<td>3</td>
<td>88.01</td>
<td>77.46</td>
</tr>
<tr>
<td>PointNet++ [24]</td>
<td>0.88</td>
<td>1</td>
<td>85.43</td>
<td>79.35</td>
</tr>
<tr>
<td>Meteornet [20]</td>
<td>1.78</td>
<td>3</td>
<td>86.78</td>
<td>81.80</td>
</tr>
<tr>
<td>ASAP-Net [2]</td>
<td>1.84</td>
<td>3</td>
<td>87.02</td>
<td>82.73</td>
</tr>
<tr>
<td>PST2 (ours)</td>
<td>2.99</td>
<td>3</td>
<td>87.03</td>
<td>81.86</td>
</tr>
</tbody>
</table>

Table 3. Semantic segmentation results on the SemanticKITTI dataset. Per-class and average IoU (%) are used as the evaluation metrics.

<table>
<thead>
<tr>
<th>Method</th>
<th>nframe</th>
<th>meanIoU</th>
<th>c1</th>
<th>c2</th>
<th>c3</th>
<th>c4</th>
<th>c5</th>
<th>c6</th>
<th>c7</th>
<th>c8</th>
<th>c9</th>
<th>c10</th>
<th>c11</th>
<th>c12</th>
<th>c13</th>
<th>c14</th>
<th>c15</th>
<th>c16</th>
<th>c17</th>
<th>c18</th>
<th>c19</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNv2 [24]</td>
<td>1</td>
<td>20.1</td>
<td>53.7</td>
<td>19</td>
<td>0.2</td>
<td>0.9</td>
<td>0.2</td>
<td>0.9</td>
<td>1.0</td>
<td>0.0</td>
<td>72.0</td>
<td>18.7</td>
<td>41.8</td>
<td>5.6</td>
<td>62.3</td>
<td>16.9</td>
<td>46.5</td>
<td>13.8</td>
<td>20.0</td>
<td>6.0</td>
<td>8.9</td>
</tr>
<tr>
<td>ASAP-Net [2]</td>
<td>3</td>
<td>33.3</td>
<td>84.1</td>
<td>11.6</td>
<td>7.5</td>
<td>3.2</td>
<td>1.1</td>
<td>4.8</td>
<td>18.5</td>
<td>3.0</td>
<td>81.8</td>
<td>28.1</td>
<td>53.1</td>
<td>7.8</td>
<td>74.9</td>
<td>37.6</td>
<td>64.4</td>
<td>27.2</td>
<td>51.7</td>
<td>22.8</td>
<td>30.8</td>
</tr>
<tr>
<td>PST2 (ours)</td>
<td>3</td>
<td>36.5</td>
<td>84.3</td>
<td>3.4</td>
<td>14.7</td>
<td>14.8</td>
<td>14.5</td>
<td>8.7</td>
<td>31.0</td>
<td>22.4</td>
<td>81.8</td>
<td>29.6</td>
<td>62.1</td>
<td>14.2</td>
<td>78.8</td>
<td>41.8</td>
<td>63.9</td>
<td>30.6</td>
<td>56.6</td>
<td>23.8</td>
<td>17.5</td>
</tr>
</tbody>
</table>

Table 4. 3D action recognition accuracy on the MSRAAction3D dataset (%).

<table>
<thead>
<tr>
<th>Method</th>
<th>Input</th>
<th>nframe</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vieira et al. [29]</td>
<td>depth</td>
<td>20</td>
<td>78.20</td>
</tr>
<tr>
<td>Klaser et al. [16]</td>
<td>depth</td>
<td>18</td>
<td>81.43</td>
</tr>
<tr>
<td>Actionlet [30]</td>
<td>skeleton</td>
<td>all</td>
<td>88.21</td>
</tr>
<tr>
<td>PointNet++ [24]</td>
<td>point</td>
<td>1</td>
<td>61.61</td>
</tr>
<tr>
<td>MeteorNet [20]</td>
<td>point</td>
<td>4</td>
<td>78.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>81.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>86.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>88.21</td>
</tr>
<tr>
<td>MeteorNet+STSA</td>
<td>point</td>
<td>4</td>
<td>81.14 (↑3.03%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>86.53 (↑5.39%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>12</td>
<td>88.55 (↑2.02%)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>16</td>
<td>89.22 (↑1.01%)</td>
</tr>
</tbody>
</table>

The spatial-temporal context information in a point cloud sequence.

4.3. Evaluation on SemanticKITTI

In the SemanticKITTI experiment, the state-of-the-art method ASAP-Net [2] adopts PointNet++ (PNv2) [24] as its backbone network to be incorporated with our ASAP module. For rigorous comparison, we employ the same encoder as ASAP-Net [2] to incorporate with the proposed STSA and RE module in our $PST^2$.

Table 3 shows the semantic segmentation results on the SemanticKITTI dataset. As shown in Table 3, it can be seen that our $PST^2$ outperforms PNv2 by a large margin (16.4% in mIoU). Further, our $PST^2$ surpasses the state-of-the-art method ASAP-Net [2] by 3.2% in mIoU and achieves a consistent performance improvement in 15 out of 19 categories. To provide an intuitive comparison, we visualize two segmentation results from the SemanticKITTI dataset, as shown in Fig. 4. It can be seen that our $PST^2$ can accurately segment most objects, even in some complex scenes. For example, as can be seen from the first row of Fig. 4, most points on the objects of the motorcyclist category (c8) and the trunk category (c16) are misclassified as unlabeled category by ASAP-Net [2], but correctly classified by our $PST^2$. This is mainly because our RE module can enhance the resolution of the features in each frame, and our STSA module can better incorporate the spatial-temporal context information between adjacent frames.

4.4. Evaluation on MSRAAction3D

Table 4 presents the classification results on the MSRAAction3D dataset. MeteorNet [20] achieves a significant improvement under the different number of input frames in a sequence by plugging our STSA module. Specifically, given 16 frames in a sequence, our classification pipeline achieves the best performance among these methods, with a mAcc of 89.22. Given 8 frames in a sequence, the most remarkable improvement of 5.39% is achieved by our method. Moreover, it is observed that better classification results can be achieved with more frames in a sequence. That is because, our STSA module can better capture the inter-frame spatial-temporal context information. Given more frames in a sequence, more information can be encoded by our method.

4.5. Ablation Study

In this section, we analyze the impacts of two proposed modules on the semanticKITTI dataset.

4.5.1 The RE Strategy

To demonstrate the effectiveness of our RE strategy, we employ two comparative experiments to quantify the contribu-
Table 5. Ablation results on the SemanticKITTI dataset. c1-c19 represent the 19 categories provided in the SemanticKITTI dataset, namely, car (c1), bicycle (c2), motorcycle (c3), truck (c4), other-vehicle (c5), person (c6), bicyclist (c7), motorcyclist (c8), road (c9), parking (c10), sidewalk (c11), other-ground (c12), building (c13), fence (c14), vegetation (c15), trunk (c16), terrain (c17), pole (c18), and traffic-sign (c19).

<table>
<thead>
<tr>
<th>Method</th>
<th>param (M)</th>
<th>nframe</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>PNv2 [24]</td>
<td>3.62</td>
<td>1</td>
<td>20.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>c1 53.7</td>
</tr>
<tr>
<td>PNv2+RE</td>
<td>5.13</td>
<td>1</td>
<td>38.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>c1 85.4</td>
</tr>
<tr>
<td>PNv2+STSA</td>
<td>3.87</td>
<td>3</td>
<td>32.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>c1 83.4</td>
</tr>
<tr>
<td>PST² (PNv2+STSA+RE)</td>
<td>5.63</td>
<td>3</td>
<td>36.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>c1 84.3</td>
</tr>
</tbody>
</table>

Figure 4. Visualization of 4D semantic segmentation results on the SemanticKITTI dataset. From left to right: ground truth, segmentation results of ASAP-Net [2] and our PST². Red circles are used to highlight the failure cases produced by ASAP-Net [2], while green circles are used to highlight the successful cases produced by our PST².

4.5.2 STSA Strategy

The STSA strategy not only improves the static point cloud pipeline with the ability to process point cloud sequences, but also makes the model achieve better performance. The qualification results are reported in Table 5. It can be seen that PNv2+STSA outperforms PNv2 on all categories with our STSA module. Specifically, the mIoU gains introduced by the RE strategy on all these categories vary from 0.1% to 10.5%. Large improvements can be observed on parking, building, and motorcyclist.

5. Conclusion

In this paper, we proposed a point spatial-temporal transformer framework to effectively modeling raw 3D point cloud sequences. Our PST² model consists of two core modules: the spatio-temporal self-attention module and the resolution embedding module. We apply the STSA module to capture inter-frame spatial-temporal context information, and the RE module to aggregate inter-neighborhood features to enhance the resolution. Extensive experiments demonstrate that our PST² model outperforms existing methods on the tasks of 4D semantic segmentation and 3D action classification. Further, these two modules can be easily plugged into existing static point cloud processing frameworks to achieve remarkable performance improvements.

Acknowledgement

This work was partially supported by the National Natural Science Foundation of China (No. U20A20185, 61972435), the Natural Science Foundation of Guangdong Province (2019A1515011271), and the Shenzhen Science and Technology Program (No. RCYX2020071414641140, JCYJ20190807152209394).
References


[25] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The Synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In Proceedings of the IEEE con-
ference on computer vision and pattern recognition, pages 3234–3243, 2016.


