Adapting Pre-trained 3D Models for Point Cloud Video Understanding via Cross-frame Spatio-temporal Perception

Baixuan Lv[1*] Yaohua Zha[1,2*] Tao Dai[3†] Xue Yuerong[1] Ke Chen[2] Shu-Tao Xia[1,2]
[1]Tsinghua University [2]Pengcheng Laboratory [3]Shenzhen University

Abstract

Point cloud video understanding is becoming increasingly important in fields such as robotics, autonomous driving, and augmented reality, as they can accurately represent object motion and environmental changes. Despite the progress made in self-supervised learning methods for point cloud video understanding, the limited availability of 4D data and the high computational cost of training 4D-specific models remain significant obstacles. In this paper, we investigate the potential of transferring pre-trained static 3D point cloud models to the 4D domain, pointing out the limitations of static models that capture only spatial information while neglecting temporal dynamics. To address this, we propose a novel Cross-frame Spatio-temporal Adaptation (CSA) strategy by introducing the Point Tube Adapter as the embedding layer and the Geometric Constraint Temporal Adapter (GCTA) to enforce temporal consistency across frames. This strategy extracts both short-term and long-term temporal dynamics, effectively integrating them with spatial features and enriching the model's understanding of temporal changes in point cloud videos. Extensive experiments on 3D action and gesture recognition tasks demonstrate that our method achieves state-of-the-art performance, establishing its effectiveness for point cloud video understanding. Code is available at: https://github.com/LvBaixuan/Point-CSA.

1. Introduction

Four-dimensional (4D) point cloud videos offer an accurate representation of object motion, making them increasingly valuable in fields such as robotics, autonomous driving, and augmented reality. A comprehensive understanding of point cloud videos is crucial for effectively perceiving environmental changes and facilitating interactions with the real world. Consequently, the learning-based point cloud video understanding [9–11, 20, 33] has attracted significant research interest in recent years.

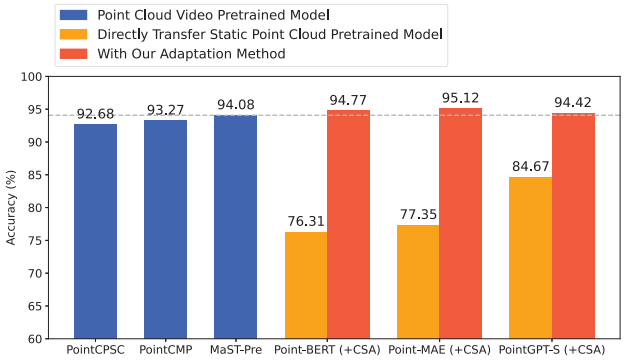


Figure 1. Comparison of direct transfer of static point cloud pre-training model and point cloud video pre-training model on MSR-Action dataset. Blue bars represent the performance of 4D-specific models, yellow bars indicate the results from directly transferring static point cloud pre-trained models, and orange bars depict the performance after applying our Adaptation Method.

Self-supervised representation learning has become the most popular approach for current point cloud video understanding, as it learns general representations using unlabeled 4D data. Although various self-supervised methods [25–28, 46] have achieved some success in point cloud video modeling, the difficulty in collecting 4D data results in relatively small pre-training datasets, which limits the ability to learn general representations. Furthermore, the introduction of the temporal dimension in point cloud videos increases the time and computational resources required during the pre-training phase, thereby raising the costs of training 4D-specific models.

Pre-trained models [4, 8, 23, 37–39, 42, 44] for 3D static point clouds have seen significant success within the community and have gained widespread adoption. Static and dynamic point clouds exhibit spatial consistency, raising the question of whether existing pre-trained models for static point clouds can be effectively leveraged for point cloud videos, thereby reducing the need for additional training of 4D-specific models. To explore this, we initially di-

*Equal contribution.
†Corresponding author. ✉ daitao.edu@gmail.com

rectly transfer existing popular pre-trained models for static point clouds to the 4D domain. However, as shown in Figure 1, the direct transfer of static point cloud pre-trained models results in significantly inferior performance compared to existing pre-training methods [25, 26, 28] specifically designed for point cloud videos. This performance gap is primarily due to the static models capturing only spatial information while neglecting critical temporal dynamics. These observations have motivated us to design an adaptation strategy that incorporates temporal awareness.

The straightforward idea is to incorporate temporal information into static point clouds. Inspired by previous work [9, 11], we introduce the Point Tube Adapter as our embedding layer, replacing the embedding layer in static point cloud pre-trained models. As shown in Table 3, although this modification significantly enhances the model's ability to understand point cloud videos, there remains a gap compared to dedicated 4D models. We believe this is due to the self-attention mechanism in the frozen backbone network being unable to dynamically capture the long-range spatio-temporal correlations between patches. To this end, we propose the Geometric Constraint Temporal Adapter (GCTA), which introduces geometric constraints on neighboring points to ensure the consistency of point cloud structures across different time frames during interaction, capturing cross-temporal geometric information. This strategy enables the model to perceive structural changes in spatial regions throughout the entire video and integrates global dynamic information into the pre-trained encoder, enriching the model's understanding of temporal dynamics.

The Cross-frame Spatio-temporal Adaptation (CSA) strategy not only allows the model to capture short-term structure changes but also emphasizes long-term dynamics, significantly enriching its understanding of temporal information. We evaluated our model on 3D action recognition and gesture recognition tasks. Experiments on the MSR-Action3D [18] and SHREC'17 [7] datasets demonstrate the effectiveness of our approach. The contributions of this paper include:

- We explore methods for transferring spatial semantic information from static point clouds to point cloud videos. To our knowledge, this is the first exploration of applying static point cloud pre-trained models in point cloud video tasks.
- We propose a novel Cross-frame Spatio-temporal Adaptation (CSA) strategy by introducing the Point Tube Adapter as the embedding layer and the Geometric Constraint Temporal Adapter (GCTA) to enforce temporal consistency across frames. This strategy extracts both short-term and long-term temporal dynamics, effectively integrating them with spatial features and enriching the model's understanding of temporal changes in point cloud videos.
- Extensive experiments on 3D action and gesture recognition tasks demonstrate that our method achieves state-of-the-art performance, highlighting its efficacy in point cloud video understanding.

2. Related Work

2.1. Static Point Cloud Pre-trained Model

Recent research on pre-training methods for 3D point cloud analysis has made significant progress. These methods typically involve learning latent semantic information from point clouds through pretext tasks, followed by fine-tuning the model for specific applications to enhance performance. Existing pretext tasks can be categorized into two types: discriminative tasks [2, 5, 12, 34, 45], and generative tasks [13, 19, 30, 36, 40, 41]. Discriminative methods enable the model to distinguish similar point clouds by contrasting various views or instances. For example, techniques such as PointContrast [35] and CrossPoint [1] leverage both in-domain and cross-domain features. Generative methods learn the latent features of the data relying on autoencoders to reconstruct the original input. For example, Point-BERT [37] generates discrete point tokens that contain meaningful local information derived from masked tokens. Point-MAE [23] employs an autoencoder to learn high-level latent features from unmasked patches, with the aim of reconstructing the masked point patches. ACT [8] significantly improves the performance of 3D self-supervised networks on real scanned point clouds by combining pre-trained language and image models. PointGPT [4] introduces an autoregressive generation task for point clouds to pre-train Transformer models, offering high scalability.

However, these pre-trained models have primarily been fine-tuned for downstream tasks associated with static point clouds. Our research aims to extend their applicability by transferring these models to the domain of point cloud video understanding.

2.2. Point Cloud Video Understanding

Point cloud videos exhibit a dual nature, being unordered in the spatial dimension while ordered in the temporal dimension. This characteristic necessitates the design of specialized methods to aggregate spatio-temporal information to perceiving both geometry and dynamics. Current approaches in this domain can be broadly classified into voxel-based [15, 43] and point-based [9–11, 20, 21, 33] techniques. Voxel-based methods, such as MinkowskiNet [6], involve voxelizing the original point clouds and then employing 4D convolution to extract features from 4D voxels. However, this voxelization process inherently results in quantization loss. In contrast, point-based methods operate directly on the raw point data. MeteorNet [20] introduces the temporal dimension and tracks motion at the point level,

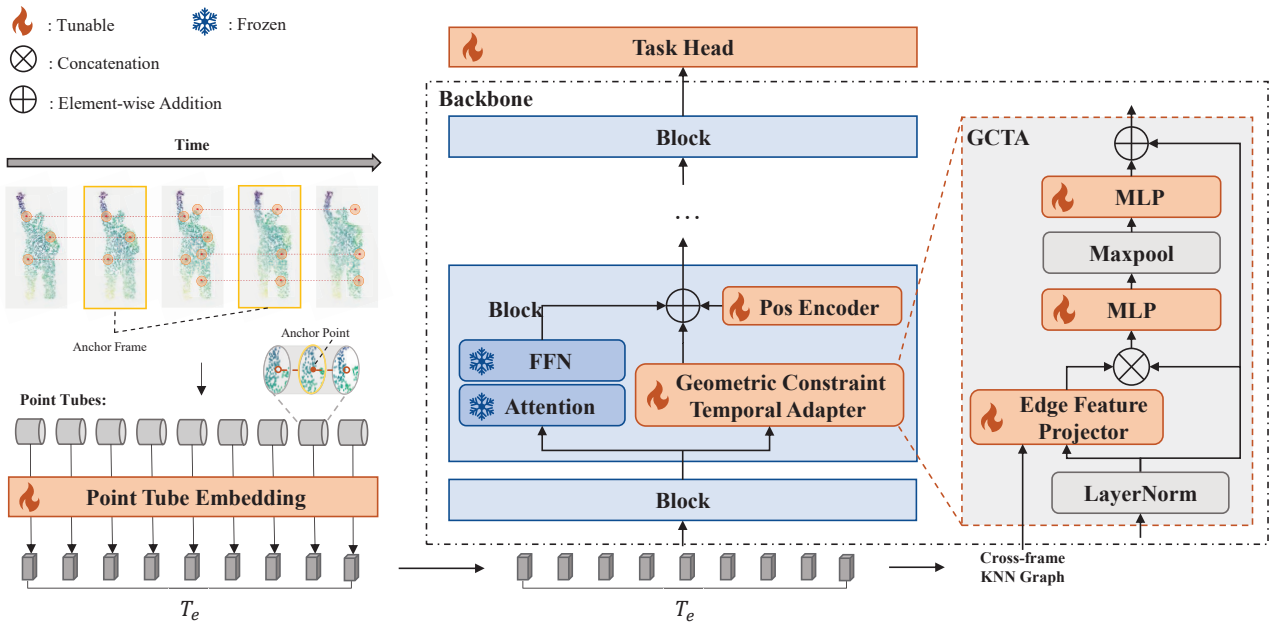


Figure 2. Illustration of the proposed Cross-frame Spatio-temporal Adaptation strategy. We divide point cloud videos into point tubes. Aggregate points within the point tube to obtain local spatio-temporal embedding, and construct a cross-frame KNN graph based on the coordinates of the anchor point of the point tubes. Our proposed Geometric Constraint Temporal Adapter (GCTA) aggregates long-term changes based on the cross-frame KNN graph.During fine-tuning, the parameters of Attention and Feed Forward Network(FFN) of the pre-trained model are frozen.

while PSTNet [11] constructs point tubes along the time axis to facilitate 4D point convolution. Recent advances, including P4Transformer [9], PPTr [33], and LeaF [21], have integrated transformer architectures to expand the receptive field. PST-Transformer [10] builds on P4Transformer, enhancing its ability to model spatio-temporal structures through spatial displacement encoding. To effectively learn the inherent information in point cloud videos, rather than focusing solely on downstream tasks, several methods employ self-supervised learning to enhance generalization capabilities. One approach [31] pretrains an encoder by predicting the temporal order of shuffled segments. C2P [46] introduces a complete-to-part 4D distillation method to predict representations of point cloud frames over short time windows. Additionally, several methods [26–28] utilize spatial and inter-frame contrastive learning to extract inherent geometric and dynamic features. However, these techniques often rely on clip-level or frame-level pretext tasks to learn motion information, which may be limited by inconsistencies in points across frames. MaST-Pre [25] proposes a masked prediction approach that segments continuous point clouds into point tubes for reconstruction. While this method effectively encodes relative spatio-temporal information within the point tubes, it falls short in capturing global semantics.

Our approach preserves the model's strong encoding capability for spatial structures learned from static point clouds while also emphasizing the capture of short-term morphological changes and long-term dynamics.

3. Method

3.1. Overview

The general framework of our method is shown in Figure 2. To enable the static point cloud pre-trained model to perceive dynamic information within point cloud videos, we introduce the Cross-frame Spatio-temporal Adaptation (CSA) strategy, which allows the model to comprehend point cloud videos through both short-term and long-term spatial structural changes. Initially, we incorporate point tubes during the feature embedding phase. This grouping strategy combines local points along both spatial and temporal dimensions. By integrating relative spatio-temporal information, the model effectively captures temporal dynamics along with spatial structures. Subsequently, to enhance attention to the spatial structural changes of point clouds across the entire video, we propose a Geometric Constraint Temporal Adapter(GCTA). This module facilitates feature aggregation over a global temporal range, allowing each point tube to traverse time and identify relevant features from other point tubes throughout the video. By aggregating spatial information from various frames, we in-

fuse global dynamic knowledge into the pre-trained model.

Through this approach, we successfully adapt the static point cloud pre-trained model for point cloud video understanding tasks while keeping its parameters frozen.

3.2. Point Tube Adapter

Some widely used static point cloud pretraining methods process input point clouds using farthest point sampling (FPS) and K-Nearest Neighbors (KNN) algorithms to obtain center points and point patches. This encoding technique effectively aggregates information from local spatial neighborhoods. However, point cloud video data exhibit dynamic characteristics, necessitating a neighborhood partitioning method that can extract spatial structures while concurrently encoding temporal information.

Building on previous work [9, 11], we introduce the concept of point tubes. Point cloud videos can be represented as a series of 3D point coordinates $P \in \mathbb{R}^{L\times N\times 3}$, where L is the length of the video and N is the number of points per frame. First, we uniformly select anchor frames within the point cloud video, obtaining anchor points using FPS within these frames and extending them to adjacent frames. Next, we search for neighbors within a spatial radius r_s around the anchor points in both the anchor frame and its adjacent frames, forming a point tube, denoted as $T \in \mathbb{R}^{l\times k\times 3}$, where l is the temporal length of the point tube and k is the number of points per frame in the point tube.

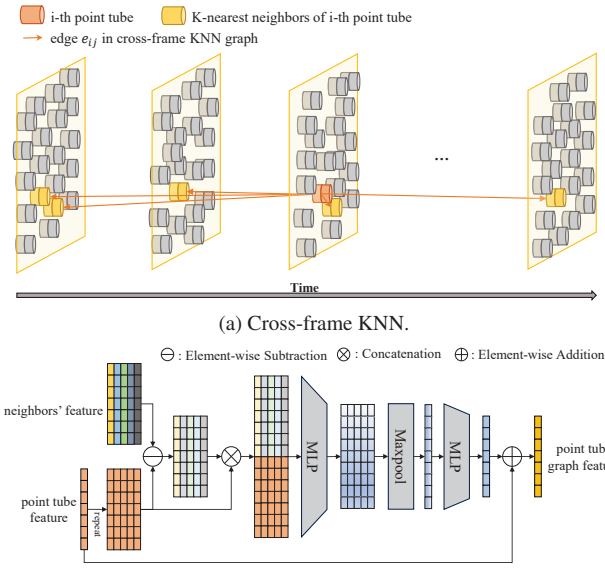
Similarly to the embedding methods for static point clouds, we explicitly calculate the relative distances between the points in the tube and the anchor points, which aids the model in capturing local spatial structures. In addition, we incorporate relative temporal information to encode short-term local structural dynamics. The obtained spatio-temporal displacements $(\delta_x, \delta_y, \delta_z, \delta_t)$ serve as point-level features F_p . The embedding of point tubes is accomplished by an extended lightweight 4D PointNet [24], which mainly consists of feature mapping functions and pooling layers. This embedding can be represented as:

$$T_e = maxpool[\psi(F_p)], \tag{1}$$

where $\psi : \mathbb{R}^{lk\times 4} \longrightarrow \mathbb{R}^{lk\times C}$ aims to increase the dimensionality of point features to enhance feature representation capabilities. We directly input the embeddings T_e into the pre-trained encoder.

3.3. Geometric Constraint Temporal Adapter

The classical Transformer [29] relies on the self-attention mechanism to perceive long-range correlations among all patches globally and has achieved great success in static 3D domains. However, simply replacing point patch embeddings with point tube embeddings still makes it challenging to adapt static point cloud pre-trained models for point cloud video. The temporal scope of point tubes remains limited by their length, capturing only short-term changes. Moreover, static point clouds only contain spatial structural information, which results in the frozen parameter static point cloud pre-trained models having a limited ability to perceive temporal relationships between point tubes.



(a) Cross-frame KNN.

(b) Aggregation of point tube features through cross-frame KNN Graph.

Figure 3. Illustration of Geometric Constraint Temporal Adapter.

To address this issue, we propose Geometric Constraint Temporal Adapter (GCTA) to capture long-term dynamic information from point cloud videos. The structure of GCTA is shown in Figure 3. Inspired by EdgeConv [32], We construct a directed cross-frame KNN graph $G = (T, E)$ based on the spatial distance between all point tubes, allowing each point tube to explore more relevant point tubes across the entire video without being constrained by time. Consider a point cloud video divided into N' point tubes with features denoted as $F = \{f_1, f_2, ..., f_{N'}\}$, we define the edge feature between the i-th and the j-th point tube as :

$$e_{ij} = \phi(concat(f_j - f_i, f_i)), \tag{2}$$

where $\phi : \mathbb{R}^{2C} \longrightarrow \mathbb{R}^{C'}$ is a function parameterized by a set of learnable parameters. We use a symmetric operation, max pooling, to aggregate the edge features and project them into feature dimension C using $\xi : \mathbb{R}^{C'} \longrightarrow \mathbb{R}^{C}$:

$$f'_i = \underset{j:(i,j)\in E}{\xi(maxpool(e_{ij}))}. \tag{3}$$

GCTA explicitly combines global variation, captured by $f_j - f_i$, with local structural features, captured by the point tube embedding f_i . By aggregating point tubes distributed across different timestamps, GCTA can learn the global dynamics of spatial structures, thereby enhancing long-term

Methods	Reference	8 Frames	12 Frames	16 Frames	24 Frames
Supervised Learning					
MeteorNet[20]	ICCV2019	81.14	86.53	88.21	88.50
PSTNet[11]	ICLR2021	83.50	87.88	89.90	91.20
P4Transformer[9]	CVPR2021	83.17	87.54	89.56	90.94
Kinet[47]	CVPR2022	83.84	88.53	91.92	93.27
PPTr[33]	ECCV2022	84.02	89.89	90.31	92.33
LeaF[21]	ICCV2023	84.50	-	91.50	93.84
PST-Transformer[10]	TPAMI2023	83.97	88.15	91.98	93.73
X4D-SceneFormer[17]	AAAI 2024	86.47	-	92.56	93.90
Self-Supervised Learning (End-to-End Fine-Tuning)					
CPR[27]	AAAI2023	86.53	91.00	92.15	93.03
C2P[46]	CVPR2023	87.16	-	91.89	94.76
PointCMP[26]	CVPR2023	89.56	91.58	92.26	93.27
PointCPSC[28]	ICCV2023	88.89	90.24	92.26	92.68
MaST-Pre[25]	ICCV2023	-	-	-	94.08
Adaptation for Pre-Trained Models					
Point-BERT+CSA	-	90.72	92.39	93.38	94.77
Point-MAE+CSA	-	**91.41**	92.04	93.73	**95.12**
PointGPT-S+CSA	-	**91.41**	**93.08**	**94.08**	94.42

Table 1. Action recognition accuracy (%) on the MSR-Action3D.

temporal perception.

3.4. Feature Interaction

In this section, we fuse local and global spatio-temporal information. The Point Tube Adapter provides the model with local spatio-temporal information, while the GCTA learns global dynamic knowledge. The global information related to spatial positions is provided by a customized position encoder (PE) implemented as a simple multilayer perceptron (MLP). The position encoders in different layers maintain the same architecture but have different parameters. The specific architecture is illustrated in the figure. The feature interaction process at the i-th layer can be represented as:

$$\begin{aligned}\mathcal{F}_{i+1} =& FFN(Attention(\mathcal{F}_i)) \\ &+ GCTA(\mathcal{F}_i) + PE(\mathcal{F}_i),\end{aligned} \tag{4}$$

where FFN is a feedforward neural network whose parameters, along with the attention parameters, are pre-trained on static point cloud data and frozen during the adaptation process.

4. Experiments

We evaluated the performance of our proposed Cross-frame Spatio-temporal Adapting (CSA) strategy on 3D action recognition and gesture recognition tasks. We also discuss the benefits of our adaptation strategy and architecture in Sec 4.4. We used three classic pre-trained models, Point-Bert [37], Point-MAE [23], and PointGPT-S [4], as our baseline.

4.1. Experiment Setting

We apply our CSA strategy on three baseline models: Point-BERT [37], Point-MAE [23], and PointGPT-S [4]. All three are pre-trained on ShapeNet [3] dataset for 300 epochs.

To ensure a fair comparison, we use the same experimental settings for each baseline. We freeze the weights of the pre-trained static point cloud model and only update the parameters of the Point Tube Embedding module, the GCTA, and the head during downstream task training. We employ the SGD optimizer. The learning rate warms up to 0.01 in the first 10 epochs and decays at a rate of 0.1 in the 20th and 30th epochs. The model is trained for a total of 50 epochs.

4.2. 3D Action Recognition

We only use the coordinates of the points, without incorporating color information. The point cloud videos are divided into multiple clips with a fixed number of frames as input. During evaluation, we average the prediction probabilities at the clip level to obtain the video-level prediction result.

The experiment is conducted on the MSR-Action3D [18] dataset. MSR-Action3D consists of 567 depth videos, with 20 action categories. Each video in the MSR-Action3D dataset contains an average of 40 frames. Following pre-

vious work [9], we partition the dataset into 270 training videos and 297 testing videos. To evaluate the performance of our adaptation strategy under different temporal resolutions, we conduct experiments by extracting 8, 12, 16, and 24 frames, with each frame randomly sampled to 2048 points.

We select an anchor frame every two frames and sample 64 anchor points from it. We set the temporal length of the point tube $l = 3$, the neighborhood space radius $r_s = 0.3$, and randomly sample 32 points in the neighborhood space. The results summarized in Table 1 demonstrate that our adaptation method outperforms both supervised and self-supervised learning approaches across all temporal configurations. These results underscore the potential of transferring knowledge from static point cloud pre-trained models to point cloud video understanding. The robust spatial feature extraction capabilities inherent in these models provide valuable insights that enhance action recognition in dynamic settings. The CSA strategy effectively bridges the gap between static and dynamic representations, ensuring that the model can capitalize on both short-term and long-term temporal dynamics. The consistent improvements across varying input configurations suggest that our approach adapts flexibly to different temporal resolutions. In the experiments with 8, 12, and 16 frames, the application of our CSA to the three baseline models outperformed existing methods, demonstrating the effectiveness of this approach at low temporal resolutions.

Overall, our results not only validate the effectiveness of the proposed adaptation method but also highlight the importance of leveraging existing pre-trained models to improve performance in downstream tasks.

4.3. Gesture Recognition

We conduct gesture recognition experiments on the SHREC'17 [7] dataset. SHREC'17 consists of 2800 videos in 28 classes for gesture recognition, of which 70% (2960 videos) are training data and the other 30% (840 videos) are the test set. To adapt our models, we freeze the parameters of the pre-trained models and retain the same architecture used in the action recognition tasks from the MSR-Action3D dataset.

As shown in Table 2, our method achieved higher recognition accuracy compared to both supervised and self-supervised methods. Notably, pre-trained models using the Cross-Frame Spatio-temporal Adaptation (CSA) strategy achieved impressive accuracies: PointGPT-S+CSA reached 96.5%, Point-BERT+CSA attained 96.2%, and Point-MAE+CSA achieved 95.2%. The high performance across various tasks reinforces the adaptability of our approach. By effectively leveraging the consistency in spatial structures between static point clouds and point cloud videos, our method successfully adapts static pre-trained models to a variety of downstream tasks. This indicates that the structural insights gained from static point clouds can be effectively transferred to dynamic contexts.

Methods	Reference	Acc.(%)
Supervised Learning		
PLSTM-base [22]	CVPR2020	87.6
PLSTM-early [22]	CVPR2020	93.5
PLSTM-PSS [22]	CVPR2020	93.1
PLSTM-middle [22]	CVPR2020	94.7
PLSTM-late [22]	CVPR2020	93.5
Kinet [47]	CVPR2022	95.2
Self-Supervised Learning (End-to-End Fine-Tuning)		
PointCMP [26]	CVPR2023	93.3
MaST-Pre [25]	ICCV2023	92.4
Adaptation for Pre-Trained Models		
Point-BERT+CSA	-	96.2
Point-MAE+CSA	-	95.2
PointGPT-S+CSA	-	**96.5**

Table 2. Gesture recognition accuracy (%) on the SHREC'17.

4.4. Ablation Studies

To analyze the impact of each module in our proposed method on performance, we utilize Point-MAE as the baseline model and conduct ablation experiments on the MSR-Action3D dataset.

4.4.1. Architecture Design

Our proposed Cross-frame Spatio-temporal Adaptation strategy enhances static point cloud pre-trained models with spatio-temporal information in three key ways: (1) Point Tube Adapter: This component aggregates local spatio-temporal neighborhood information. (2) Position Encoders: These encoders are independently learned for each layer, providing adaptive global spatial context. (3) Geometric Constraint Temporal Adapter(GCTA): This module extends the temporal field of view, enabling the model to perceive global dynamics effectively. we conduct extensive ablation studies, as shown in Table 3.

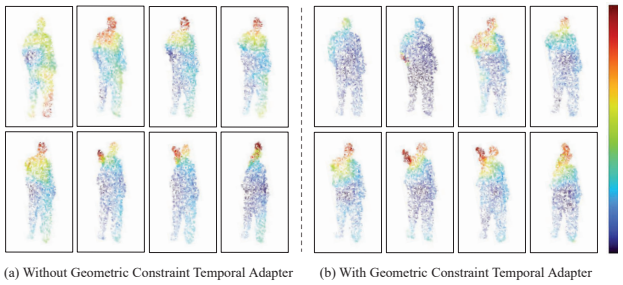
We first present the performance of model A0 as the baseline (freeze the pre-trained model and only tuned the task head), achieving an accuracy of 77.35%. Model A1, which integrates Point Tube Adapter, increases the accuracy by 6.62%, highlighting the benefit of local spatio-temporal aggregation. Adding Position Encodings in model A2 further enhances accuracy to 87.11%, emphasizing the importance of adaptive global spatial context. Model A3, which combines Point Tube Adapter with the GCTA, achieves an accuracy of 94.08%, demonstrating the value of extend-

	PTA	PE	GCTA	Acc.(%)
A0	×	×	×	77.35
A1	✓	×	×	83.97
A2	✓	✓	×	87.11
A3	✓	×	✓	94.08
A4(ours)	✓	✓	✓	**95.12**

Table 3. Ablation studies on architecture designs. PTA denotes Point Tube Adapter, and PE denotes additional Position Encoders.

ing the temporal field of view. Finally, our full model A4, incorporating all components, reaches an accuracy of 95.12%, confirming the effectiveness of our Cross-frame Spatio-temporal Adaptation strategy in leveraging both local and global information to improve recognition performance in dynamic point cloud scenarios.

We visualize the attention distribution of the model with and without the GCTA. As shown in the Figure 4, for the same input video, the model adding the GCTA focuses more on the key frames where the action occurs, and allocates less attention to the frames that are weakly related to the motion (the first two frames). This demonstrates that that the GCTA helps the model understand long-term motion.



(a) Without Geometric Constraint Temporal Adapter (b) With Geometric Constraint Temporal Adapter

Figure 4. Visualization of the attention of model with/without the Geometric Constraint Temporal Adapter. Warmer color indicates higher attention. The model with the Geometric Constraint Temporal Adapter can correctly identify important frames and moving areas.

4.4.2. Advantages of CSA compared to Other Tuning strategies

To demonstrate the effectiveness of our Cross-frame Spatio-temporal Adaptation strategy, we conduct comprehensive comparative experiments against various tuning strategies while keeping the parameters of the pre-trained model frozen. As shown in the Figure 5, we compare our method with the vanilla adapter [14], VPT-Deep [16], and IDPT [39]. The baseline is consistent with that described in Sec 4.4.1. Our CSA outperforms other tuning strategies in terms of performance. We fully consider the spatially disordered and temporally ordered characteristics of point cloud

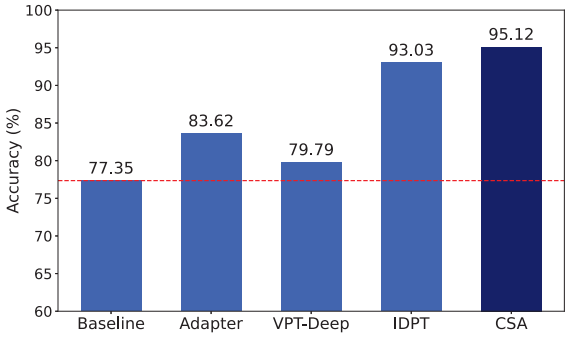
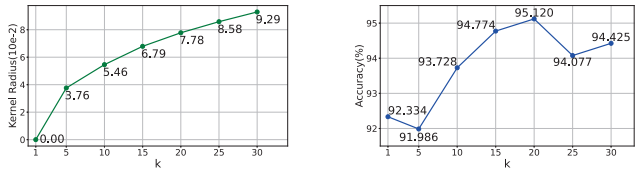


Figure 5. Comparison of different tuning strategies.

videos, and enhance the ability of models to capture dynamics by combining spatio-temporal information. This focus on dynamics is crucial for point cloud video tasks such as action recognition and gesture detection, where temporal changes play a significant role.

4.4.3. Receptive Field of Geometric Constraint Temporal Adapter

The size of the convolutional kernel significantly influences the ability to model local structures. In the Geometric Constraint Temporal Adapter, the kernel size of EdgeConv [32] is determined by the number of neighbors k selected for each anchor point during graph construction. We investigate the impact of varying k values on spatio-temporal modeling. The kernel radius is defined as the distance between the center point of a point tube and the farthest point tube among its k-nearest neighbors. Figure 6a shows the average kernel radius corresponding to different k values. Figure 6b shows the video-level recognition accuracy under different k values.



(a) Kernel radius. (b) Accuracy.

Figure 6. The effect of number of edge per point tube in the Geometric Constraint Temporal Adapter.

As k increases, the kernel radius also increases, indicating that a larger k allows the model to consider a broader spatial context when processing local structures. The kernel radius (<0.1) is much smaller than the point tube spatial radius (0.3), which means that the GCTA can focus on changes in spatial structure at similar locations in different frames.

In the case of $k = 1$, point tubes do not aggregate infor-

mation from others. Increasing k to 5 results in a slight decrease in accuracy to 91.99%, suggesting that too few neighbors may limit the ability to capture relevant features or even confuse its own information. The peak accuracy of 95.12% is observed at $k = 20$, demonstrating optimal performance when the model has access to a sufficient number of neighbors. However, as k increases to 25 and 30, the accuracy decreases by 1.04% and 0.69%, respectively, suggesting that excessive neighborhood information may introduce redundancy that negatively impacts performance.

We further cut the point cloud video into slices, which limits the process of searching for spatial K-nearest neighbors within those slices. We evaluate the impact of the temporal range on the GCTA. The results shown in Table 4 indicate a clear trend: as the temporal range increases, model performance tends to improve, peaking when the full video is utilized. The model achieves the highest accuracy of 95.12% when using the full video as the slice length. This result underscores the advantage of utilizing the entire temporal span of the video, enabling the model to integrate global context and dynamics effectively.

Slice Lenth	Acc.(%)
4	93.38
6	93.03
8	94.08
12	94.77
24(full clip)	**95.12**

Table 4. The effect of temporal range.

4.4.4. Location of Geometric Constraint Temporal Adapter

We add the GCTA at different locations to explore how spatio-temporal information can be effectively combined with static point cloud pre-trained models. As shown in the Figure 7, we conducted experiments at 4 locations: (a) Bypass (b) Prepend (c) Post (d) Middle.

The model achieves the highest accuracy of 95.12% when the GCTA is positioned at the Bypass location. This suggests that integrating spatio-temporal information in a way that allows the model to process it alongside static features leads to optimal performance. The Bypass configuration allows the model to maintain the integrity of the pre-trained static features while benefiting from additional temporal dynamics. Compared to the result of placing at the Middle location, those at the Post location indicates that applying the GCTA after adding the global spatial position information still maintains a strong impact on the model's performance, suggesting that the temporal dynamics introduced at this stage enhance the overall understanding of the data. The lowest accuracy is observed at the Prepend location. This placement disrupts the initial processing of static features, indicating that introducing temporal dynamics too early in the pipeline can hinder the model's ability to effectively leverage pre-trained static information.

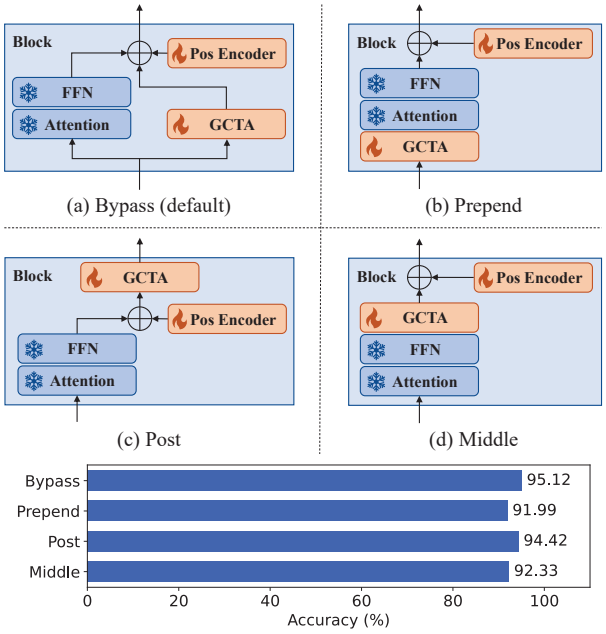


Figure 7. Different location of Geometric Constraint Temporal Adapter.

5. Conclusion

In this paper, we explored transferring pre-trained static point cloud models to point cloud video tasks and identified the limitations of static models in capturing temporal dynamics. We proposed a novel Cross-frame Spatio-temporal Adaptation (CSA) strategy, integrating the Point Tube Adapter and Geometric Constraint Temporal Adapter (GCTA) to capture both short-term and long-term temporal dynamics. Our method outperforms existing approaches in 3D action and gesture recognition tasks, demonstrating its effectiveness in bridging the gap between static and dynamic point cloud representations. This work advances point cloud video understanding and offers valuable insights for future research.

Acknowledgment

This work is supported in part by the National Natural Science Foundation of China, under Grant (62302309,62171248), Shenzhen Science and Technology Program (JCYJ20220818101014030, JCYJ20220818101012025), and the PCNL KEY project (PCL2023AS6-1).

References

[1] Mohamed Afham, Isuru Dissanayake, Dinithi Dissanayake, Amaya Dharmasiri, Kanchana Thilakarathna, and Ranga Rodrigo. Crosspoint: Self-supervised cross-modal contrastive learning for 3d point cloud understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9902–9912, 2022. 2

[2] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356), 1992. 2

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 5

[4] Guangyan Chen, Meiling Wang, Yi Yang, Kai Yu, Li Yuan, and Yufeng Yue. Pointgpt: Auto-regressively generative pre-training from point clouds. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 5

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 2

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 2

[7] Quentin De Smedt, Hazem Wannous, Jean-Philippe Vandeborre, Joris Guerry, Bertrand Le Saux, and David Filliat. Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset. In *3DOR-10th Eurographics Workshop on 3D Object Retrieval*, pages 1–6, 2017. 2, 6

[8] Runpei Dong, Zekun Qi, Linfeng Zhang, Junbo Zhang, Jianjian Sun, Zheng Ge, Li Yi, and Kaisheng Ma. Autoencoders as cross-modal teachers: Can pretrained 2d image transformers help 3d representation learning? *arXiv preprint arXiv:2212.08320*, 2022. 1, 2

[9] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14204–14213, 2021. 1, 2, 3, 4, 5, 6

[10] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point spatio-temporal transformer networks for point cloud video modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2181–2192, 2022. 3, 5

[11] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. *arXiv preprint arXiv:2205.13713*, 2022. 1, 2, 3, 4, 5

[12] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. 2

[14] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019. 7

[15] Jian-Fang Hu, Wei-Shi Zheng, Jiahui Pan, Jianhuang Lai, and Jianguo Zhang. Deep bilinear learning for rgb-d action recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 335–351, 2018. 2

[16] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 7

[17] Linglin Jing, Ying Xue, Xu Yan, Chaoda Zheng, Dong Wang, Ruimao Zhang, Zhigang Wang, Hui Fang, Bin Zhao, and Zhen Li. X4d-sceneformer: Enhanced scene understanding on 4d point cloud videos through cross-modal knowledge transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2670–2678, 2024. 5

[18] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 9–14. IEEE, 2010. 2, 5

[19] Haotian Liu, Mu Cai, and Yong Jae Lee. Masked discrimination for self-supervised learning on point clouds. In *ECCV*, 2022. 2

[20] Xingyu Liu, Mengyuan Yan, and Jeannette Bohg. Meteornet: Deep learning on dynamic 3d point cloud sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9246–9255, 2019. 1, 2, 5

[21] Yunze Liu, Junyu Chen, Zekai Zhang, Jingwei Huang, and Li Yi. Leaf: Learning frames for 4d point cloud sequence understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 604–613, 2023. 2, 3, 5

[22] Yuecong Min, Yanxiao Zhang, Xiujuan Chai, and Xilin Chen. An efficient pointlstm for point clouds based gesture recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5761–5770, 2020. 6

[23] Yatian Pang, Wenxiao Wang, Francis EH Tay, Wei Liu, Yonghong Tian, and Li Yuan. Masked autoencoders for point cloud self-supervised learning. In *European conference on computer vision*, pages 604–621. Springer, 2022. 1, 2, 5

[24] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 4

[25] Zhiqiang Shen, Xiaoxiao Sheng, Hehe Fan, Longguang Wang, Yulan Guo, Qiong Liu, Hao Wen, and Xi Zhou. Masked spatio-temporal structure prediction for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16580–16589, 2023. 1, 2, 3, 5, 6

[26] Zhiqiang Shen, Xiaoxiao Sheng, Longguang Wang, Yulan Guo, Qiong Liu, and Xi Zhou. Pointcmp: Contrastive mask prediction for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1212–1222, 2023. 2, 3, 5, 6

[27] Xiaoxiao Sheng, Zhiqiang Shen, and Gang Xiao. Contrastive predictive autoencoders for dynamic point cloud self-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9802–9810, 2023. 5

[28] Xiaoxiao Sheng, Zhiqiang Shen, Gang Xiao, Longguang Wang, Yulan Guo, and Hehe Fan. Point contrastive prediction with semantic clustering for self-supervised learning on point cloud videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16515–16524, 2023. 1, 2, 3, 5

[29] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[30] Hanchen Wang, Qi Liu, Xiangyu Yue, Joan Lasenby, and Matt J Kusner. Unsupervised point cloud pre-training via occlusion completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9782–9792, 2021. 2

[31] Haiyan Wang, Liang Yang, Xuejian Rong, Jinglun Feng, and Yingli Tian. Self-supervised 4d spatio-temporal feature learning via order prediction of sequential point cloud clips. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3762–3771, 2021. 3

[32] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019. 4, 7

[33] Hao Wen, Yunze Liu, Jingwei Huang, Bo Duan, and Li Yi. Point primitive transformer for long-term 4d point cloud video understanding. In *European Conference on Computer Vision*, pages 19–35. Springer, 2022. 1, 2, 3, 5

[34] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *CVPR*, 2018. 2

[35] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020. 2

[36] Siming Yan, Zhenpei Yang, Haoxiang Li, Chen Song, Li Guan, Hao Kang, Gang Hua, and Qixing Huang. Implicit autoencoder for point-cloud self-supervised representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14530–14542, 2023. 2

[37] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19313–19322, 2022. 1, 2, 5

[38] Yaohua Zha, Rongsheng Li, Tao Dai, Jianyu Xiong, Xin Wang, and Shu-Tao Xia. Sfr: Semantic-aware feature rendering of point cloud. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[39] Yaohua Zha, Jinpeng Wang, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Instance-aware dynamic prompt tuning for pre-trained point cloud models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14161–14170, 2023. 1, 7

[40] Yaohua Zha, Huizhen Ji, Jinmin Li, Rongsheng Li, Tao Dai, Bin Chen, Zhi Wang, and Shu-Tao Xia. Towards compact 3d representations via point feature enhancement masked autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6962–6970, 2024. 2

[41] Yaohua Zha, Naiqi Li, Yanzi Wang, Tao Dai, Hang Guo, Bin Chen, Zhi Wang, Zhihao Ouyang, and Shu-Tao Xia. Lcm: locally constrained compact point cloud model for masked point modeling. *arXiv preprint arXiv:2405.17149*, 2024. 2

[42] Renrui Zhang, Ziyu Guo, Peng Gao, Rongyao Fang, Bin Zhao, Dong Wang, Yu Qiao, and Hongsheng Li. Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training. *Advances in neural information processing systems*, 35:27061–27074, 2022. 1

[43] Shiwen Zhang, Sheng Guo, Weilin Huang, Matthew R Scott, and Limin Wang. V4d: 4d convolutional neural networks for video-level representation learning. *arXiv preprint arXiv:2002.07442*, 2020. 2

[44] Yabin Zhang, Jiehong Lin, Ruihuang Li, Kui Jia, and Lei Zhang. Point-dae: Denoising autoencoders for self-supervised point cloud learning. *arXiv e-prints*, pages arXiv–2211, 2022. 1

[45] Zaiwei Zhang, Rohit Girdhar, Armand Joulin, and Ishan Misra. Self-supervised pretraining of 3d features on any point-cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10252–10263, 2021. 2

[46] Zhuoyang Zhang, Yuhao Dong, Yunze Liu, and Li Yi. Complete-to-partial 4d distillation for self-supervised point cloud sequence representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17661–17670, 2023. 1, 3, 5

[47] Jia-Xing Zhong, Kaichen Zhou, Qingyong Hu, Bing Wang, Niki Trigoni, and Andrew Markham. No pain, big gain: classify dynamic point cloud sequences with static models by fitting feature-level space-time surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8510–8520, 2022. 5, 6