2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

CrossVideo: Self-supervised Cross-modal Contrastive Learning for Point Cloud Video Understanding

Yunze Liu[1,2], Changxi Chen[1], Zifan Wang[1], Li Yi[1,2,3]

***Abstract*— This paper introduces a novel approach named CrossVideo, which aims to enhance self-supervised cross-modal contrastive learning in the field of point cloud video understanding. Traditional supervised learning methods encounter limitations due to data scarcity and challenges in label acquisition. To address these issues, we propose a self-supervised learning method that leverages the cross-modal relationship between point cloud videos and image videos to acquire meaningful feature representations. Intra-modal and cross-modal contrastive learning techniques are employed to facilitate effective comprehension of point cloud video. We also propose a multi-level contrastive approach for both modalities. Through extensive experiments, we demonstrate that our method significantly surpasses previous state-of-the-art approaches, and we conduct comprehensive ablation studies to validate the effectiveness of our proposed designs.**

I. INTRODUCTION

Recently, there has been a wide interest in understanding point cloud sequences in 4D (3D space + 1D time) [1], [2], [3], [4]. Point cloud video is a dynamic sequence constructed based on three-dimensional point cloud data. It plays an important role in the fields of computer vision and machine learning. Point cloud video contains rich geometric and topological information, which can accurately describe objects and scenes in the real world. It has wide applications in areas such as autonomous driving, robot navigation, and augmented reality.

Despite the ubiquity of 4D data, annotating such data on a large scale with detailed information is costly. Therefore, we need to find ways to make use of massive amounts of unlabeled data. Among possible solutions, self-supervised representation learning has demonstrated its effectiveness in various fields, including image, video, and point cloud.

The existing self-supervised point cloud video representation learning solution is 4D distillation [5]. The core of this method is to compare the differences between different frames of the same object or scene, which can capture features such as shape, size, and dynamic changes of objects. A teacher network that can obtain full information is used to guide a student network that can only obtain partial information. However, real point cloud videos are often hindered by sampling patterns, noise, and other interferences, which often prevent the student network from recovering full information. Besides, this method also requires manually designing inputs for both the teacher network and the student network, which significantly affects pretraining performance.

*This work was supported by Shanghai Qi Zhi Institute
[1]Tsinghua University, [2] Shanghai Qi Zhi Institute, [3] Shanghai Artificial Intelligence Laboratory. Mails:liuyzchina@gmail.com

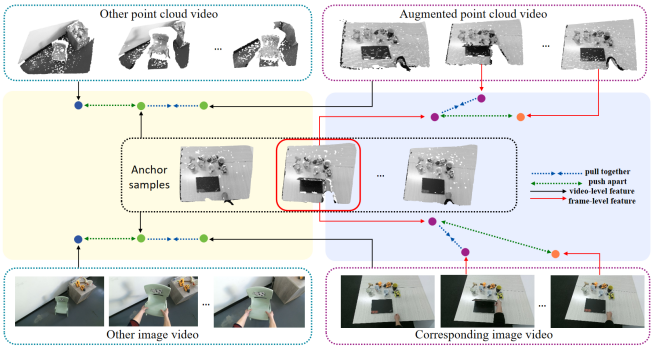

Fig. 1. Our core idea is to learn 4D point cloud video representations by leveraging the synergy from image video under a self-supervised manner. We propose intra-modal contrastive objective and cross-modal contrastive objective to learn the spatio-temporal invariance of point cloud videos and the correlation between point cloud videos and image videos.

We assume the appearance information in image videos can complement point cloud videos. For example, when a person is "graffitiing on a wall", it is difficult to perceive the changes on the wall from the point cloud alone, but images can provide information about the changes. Additionally, the color information and texture details in image video can also provide additional features to assist in motion and action recognition. We assume that point cloud video and image video can help each other, so we propose to learn 4D representation by leveraging the synergy from image video.

But learning 4D representations by leveraging the synergy from image video under a self-supervised manner is not an easy task. It is well known that a point cloud video is composed of multiple point clouds, and the same applies to image videos. We need to enable the network to understand information at different granularities, including understanding at the video level and at the video frame level. Besides, to conduct effective contrastive learning, it is also necessary to design a suitable backbone for cross-modal pre-training.

We have studied the characteristics of point cloud videos and image videos, and have obtained two main observations. Firstly, for point cloud videos, as they provide accurate three-dimensional coordinate information, they are able to capture the spatial position, shape, and motion trajectory of objects. Appearance information in image videos can complement point cloud videos, like the example of "graffiti on a wall" mentioned above. Continuous sequences of images in image videos can reflect the motion patterns and action states of objects from color information and texture details. Therefore, these two modalities naturally complement each other. To facilitate the synergy between image videos and point cloud videos, we introduce a self-supervised cross-modal

contrastive learning method. Secondly, there exists a strong correlation between video and frames, thus understanding videos can enhance the understanding of frames, and vice versa. Therefore, we need to design objective to reflect such synergy by contrasting features at different granularity levels.

Cross-modal contrastive learning for point cloud video understanding has been rarely discussed in the self-supervised representation learning literature. To the best of our knowledge, we are the first to propose to learn 4D representations by leveraging the synergy from image video under a self-supervised manner. We have carefully designed a multi-level objective function and pre-training backbone.

In particular, our method is as shown in Fig. 1. First, we design a network to extract video-level and frame-level features from point cloud videos and image videos. In addition, for point cloud videos, we apply spatio-temporal data augmentation to obtain a new version. Then, we propose intra-modal contrastive objective and cross-modal contrastive objective to learn the spatio-temporal invariance of point cloud videos and the correlation between point cloud videos and image videos. Finally, we can get a strong point cloud video encoder by self-supervised pretraining.

We evaluate our method on two downstream point cloud video tasks: 4D action segmentation on HOI4D [6], 4D semantic segmentation on HOI4D [6]. We demonstrate significant improvements over the previous method($+2.6\%$ accuracy on HOI4D action segmentation, $+1.5\%$ mIoU on HOI4D semantic segmentation). At the same time, we also demonstrate that the image video encoder can also integrate information from the point cloud video, thereby enhancing its representation capability.

The contributions of this paper are fourfold: First, we propose the first 4D self-supervised cross-modal representation learning method which facilitates the synergy of image video and point cloud video learning. Second, we propose to use intra-modal and cross-modal contrastive learning to facilitate an effective point cloud video understanding. Third, we propose to contrast the features of both modalities at different levels. Fourth, extensive experiments show that our method outperforms previous state-of-the-art methods by a large margin and we provide comprehensive ablation studies to validate our designs.

II. RELATED WORK

A. 4D Point cloud video understanding.

Compared to understanding static 3D point clouds, comprehending 4D point cloud sequences requires a stronger focus on aggregating and utilizing spatial-temporal information to perceive both the geometry and dynamics. To tackle these challenges, several 4D backbones have been proposed, which can be divided into two categories based on their representations. The first category involves voxelizing raw point clouds and extracting features from 4D voxels. An example is MinkowskiNet [7], which applies 4D spatial-temporal convolutions on 4D voxels. The second category operates directly on raw points. For instance, MeteorNet [8] extends PointNet++ [9] by introducing a temporal dimension and explicitly tracking points' motion for grouping. PSTNet [10], on the other hand, constructs a point tube along the temporal dimension for 4D point convolution. State-of-the-art methods like Point 4D Transformer [1] and PPTr [2] belong to the second category. They incorporate transformer architecture to avoid point tracking and better capture spatio-temporal correlation.

B. 3D representation learning

Due to advances in 2D representation learning [11], [12], [13], similar progress has been made in the field of 3D representation learning [14], [15], [16], [17]. Existing methods fall into two main categories: generative-based methods [18], [19], [15], [20], such as Point-Bert [15] which recover masked object parts to learn Transformer representations, and context-based methods [21], [22], [23], [24], [25], like SelfCorrection [24] which distinguish and restore destroyed objects to learn informative representations. Existing methods face challenges when extended to high-dimensional 4D data due to the difficulty of optimization and computational cost. We aim to explore the extraction of high-quality spatio-temporal features at the scene level for enhancing 4D downstream tasks, so we introduce the utilization of multimodal data into the field of 4D point cloud pretraining.

C. 4D Representation Learning.

4D Representation Learning is an emerging field upon 3D research. For instance, 4Dcontrast [26] exploits 4D motion information and STRL [27] uses temporal-spatial contrastive learning. Although pre-trained on 4D data, they only learn static 3D representations. C2P [5] is a new pre-training method for 4D point cloud sequence representation learning. Our approach builds upon the foundations laid by STRL and C2P, which represents a pioneering effort in incorporating multimodal data into 4D point cloud pretraining.

D. Cross-modal Representation Learning.

Cross-modal learning leverages diverse data sources to yield rich contextual information and effective semantic comprehension. Recent studies [28], [29], [30] have demonstrated the efficiency of cross-modal pretraining in generating various representations suitable for many subsequent tasks. CLIP [30] learns a multimodal embedding space by maximizing cosine similarity between images and text, while Morgado et al. [31] combines audio and video modalities to achieve notable performance improvements in action and sound recognition tasks through cross-modal agreement.

In addition, the research conducted by [32] expands the pretrained 2D image model to a point-cloud model by utilizing filter inflation. [33] suggest rendering images for each point cloud and implementing cross-modal contrastive loss to enhance performance. However, these approaches cannot be directly applied to 4D point cloud videos due to the unsuitability and ineffectiveness of existing pre-training backbones and contrastive objectives.

III. METHOD

A. Overview

We propose a novel multimodal point cloud video pretraining method to obtain a powerful point cloud video encoder in an unsupervised manner. Our key idea is to learn 4D representations by leveraging the synergy from image video. And we propose to align point cloud videos with image videos at both the video-level and frame-level in the feature space through cross-modal contrastive learning. We note that since our main focus is on the understanding of point cloud videos, we will primarily introduce how to train a powerful point cloud video encoder in the following sections. At the same time, in the experimental section, we also demonstrate the 2D video encoder can also obtain strong representations.

The overview of our method is shown in the Fig. 2. Given a point cloud video, we first perform data augmentation to obtain an augmented version of the point cloud video. A 4Dconv module is used to extract features, followed by a transformer to further facilitate feature communication across different time steps. Consequently, we can obtain the features of the point cloud video and its features for each frame. Similarly, for the image video, we first use a 3Dconv module to extract features and a transformer to facilitate feature communication, obtaining the features of the image video and its features for each frame. At the video-level and frame-level, we perform cross-modal and intra-modal contrastive learning respectively to obtain powerful encoders for point cloud video and image video. Cross-modal pre-training allows knowledge to be transferred from image video to the point cloud video encoder, while intra-modal pre-training enables the encoder to have invariance to a set of geometric transformations. The feature spaces at both the video-level and frame-level provide constraints on the encoder at different granularities, thereby further enhancing its representation ability. In the remaining part of this section, we will introduce our method in detail. First, we present the network architecture details of the proposed method. Then, we describe the contrastive learning loss functions formulated under intra-modal and cross-modal settings. Finally, we propose our overall training objective.

B. Preliminaries

Suppose we are given a dataset, $\mathcal{D} = \{(\mathbf{P}_i, \mathbf{I}_i)\}_{i=1}^{|\mathcal{D}|}$ with $\mathbf{P}_i \in \mathbb{R}^{L\times N\times 3}$ and $\mathbf{I}_i \in \mathbb{R}^{L\times H\times W\times 3}$. We define each $\mathbf{P}_i$ with sequence length L as $\mathbf{P}_i = \{s_1, s_2, \ldots, s_j, \ldots, s_L\}$ where each s denotes one frame. We aim to train a point cloud video feature extractor $f_{\theta_\mathbf{P}}(.)$ in a self-supervised manner to be effectively transferable to downstream tasks. To this end, we use an image video feature extractor $f_{\theta_\mathbf{I}}(.)$, multi-layer perceptron (MLP) projection heads $g_{\phi_\mathbf{P}}(.)$ and $g_{\phi_\mathbf{I}}(.)$ for point cloud and image respectively.

C. Pre-training Backbone

For point cloud videos, due to providing accurate 3D coordinate information, they can capture the spatial position, shape, and motion trajectory of objects. This advantage makes point cloud videos beneficial for motion analysis and action recognition.

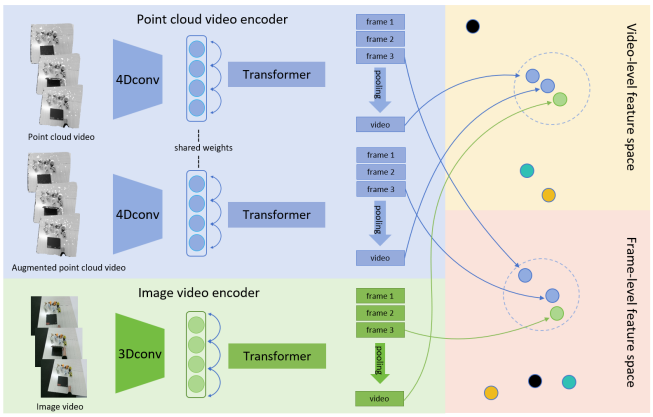

Fig. 2. Cross-modality pretraining enables distillation of knowledge from 2D videos into point cloud videos, while intra-modality pretraining allows the encoder to possess invariance to a set of geometric transformations. The feature space at the video level and the frame level can both provide constraints on the encoder at different levels of granularity.

Although image video cannot directly provide three-dimensional geometric information, the captured continuous image sequences can reflect the motion patterns and action states of objects. By utilizing the temporal relationship of image sequences, motion analysis and action recognition can be performed. Additionally, the color information and texture details in image video can also provide additional features to assist in motion and action recognition.

In order to support our pre-training, we need a network that can extract features from both image videos and point cloud videos and can support two types of features(video-level and frame-level). Based on this, we first design a three-tower structure as shown in Fig 2. The whole network is divided into two branches, namely the point cloud-video branch and the image-video branch.

1) Point cloud video encoder: The point cloud video branch consists of two shared-weight networks, each taking the original point cloud video and the enhanced point cloud video as inputs. Point cloud video encoder is divided into 4D convolution module and Transformer module. The 4D convolution is mainly used to extract spatiotemporal features, while the Transformer is used to enhance information exchange across different time sequences.

2) Image video encoder: Image video encoder is based on Asformer, which is mainly divided into 3D Convolution and Transformer modules. 3D Convolution is used to extract spatiotemporal features, while Transformer is used to enhance the model's representation ability and increase cross-temporal communication capability.

D. Intra-Modal Contrastive learning

Inspired by contrastive learning in the fields of image and static point cloud, we believe that intra-modal contrastive learning is crucial for encoding invariant representations. At the video level, we model the problem to keep invariance

to spatiotemporal data augmentation including geometric transformations $\mathbf{T_g}$ and temporal transformations $\mathbf{T_t}$.

Given an input 4D point cloud video $\mathbf{P}_i$, we denote the raw point cloud video and the augmented versions $\mathbf{P}_i^{t_1}$ and $\mathbf{P}_i^{t_2}$ of it. We compose t_2 by randomly combining transformations from $\mathbf{T_g}$ and $\mathbf{T_t}$. For geometric transformations $\mathbf{T_g}$, we use rotation, scaling and translation. For temporal transformation, we extract some frames by down-sampling.

The point cloud video feature extractor $f_{\theta_\mathbf{P}}$ take $\mathbf{P}_i$ as input, and output both video-level and frame-level feature. The feature vectors are projected to an invariant space $\mathbb{R}^d$ where the contrastive loss is applied, using the projection head $g_{\phi_\mathbf{P}}$. We denote the projected video-level feature of $\mathbf{P}_i^{t_1}$ and $\mathbf{P}_i^{t_2}$ as $\mathbf{z}_{v_i}^{t_1}$ and $\mathbf{z}_{v_i}^{t_2}$ and projected frame-level feature as $\mathbf{z}_{f_i}^{t_1}$ and $\mathbf{z}_{f_i}^{t_2}$ respectively where, $\mathbf{z}_i^t = g_{\phi_\mathbf{P}}\left(f_{\theta_\mathbf{P}}\left(\mathbf{P}_i^t\right)\right)$. Here, $\mathbf{z}_{v_i}^t$ is obtained by applying the max-pooling to $\mathbf{z}_{f_i}^t$.

For video-level contrastive learning, the goal is to maximize the similarity of $\mathbf{z}_{v_i}^{t_1}$ with $\mathbf{z}_{v_i}^{t_2}$ while minimizing the similarity with all the other projected vectors in the mini-batch of point clouds. Similarly, for frame-level contrastive learning, the goal is to maximize the similarity of $\mathbf{z}_{f_i}^{t_1}$ with $\mathbf{z}_{f_i}^{t_2}$ while minimizing the similarity with others. We leverage NT-Xent loss proposed in SimCLR for instance discrimination.

For video-level, We compute the loss function $L_v(i, t_1, t_2)$ for the positive pair of examples $\mathbf{z}_{v_i}^{t_1}$ and $\mathbf{z}_{v_i}^{t_2}$ as:

$$L_v = -\log \frac{\exp(s(\mathbf{z}_{v_i}^{t_1}, \mathbf{z}_{v_i}^{t_2})/\tau)}{\sum\limits_{\substack{k=1 \\ k\neq i}}^{N} \exp(s(\mathbf{z}_{v_i}^{t_1}, \mathbf{z}_{v_k}^{t_1})/\tau) + \sum\limits_{k=1}^{N} \exp(s(\mathbf{z}_{v_i}^{t_1}, \mathbf{z}_{v_k}^{t_2})/\tau)} \tag{1}$$

where N is the mini-batch size, τ is the temperature coefficient and $s(.)$ denotes the cosine similarity function.

Similarly, for frame-level, the objective can be written as:

$$L_f = -\log \frac{\exp(s(\mathbf{z}_{f_i}^{t_1}, \mathbf{z}_{f_i}^{t_2})/\tau)}{\sum\limits_{\substack{k=1 \\ k\neq i}}^{N} \exp(s(\mathbf{z}_{f_i}^{t_1}, \mathbf{z}_{f_k}^{t_1})/\tau) + \sum\limits_{k=1}^{N} \exp(s(\mathbf{z}_{f_i}^{t_1}, \mathbf{z}_{f_k}^{t_2})/\tau)} \tag{2}$$

Our intra-modal instance discrimination loss function $\mathcal{L}_{intra}$ for a mini-batch can be described as:

$$\mathcal{L}_{intra} = \frac{1}{2N} \sum_{i=1}^{N} [L_v(i, t_1, t_2) + L_f(i, t_1, t_2)] \tag{3}$$

E. Cross-Modal Contrastive learning

We designed a cross-modal pre-training method to empower the encoders of both modalities, facilitating communication and integration between the two modalities. To the best of our knowledge, we are the first study to systematically investigate the pretraining of multi-modal point cloud videos. And we propose a simple and effective method. We empirically validate with the experimental results that our objective outperforms existing unsupervised representation methods, thus facilitating an effective representation learning of 4D point clouds video.

To this end, we first embed the corresponding image video $\mathbf{I}_i$ to a feature space using the image video feature extractor $f_{\theta_\mathbf{I}}$. We then project the feature vectors to the invariant space $\mathbb{R}^d$ using the image projection head $g_{\phi_\mathbf{I}}$. Note that, we can obtain both video-level features and frame-level features here. The projected image video-level and frame-level feature is defined as $\mathbf{h}_{v_i}$ and $\mathbf{h}_{f_i}$ where $\mathbf{h}_i = g_{\phi_\mathbf{I}}\left(f_{\theta_\mathbf{I}}\left(\mathbf{I}_i\right)\right)$. Here, $\mathbf{h}_{v_i}$ is obtained by applying the max-pooling to $\mathbf{h}_{f_i}$. Then, we compute the mean of the projected vectors $\mathbf{z}_{v_i}^{t_1}$ and $\mathbf{z}_{v_i}^{t_2}$ to obtain the projected prototype video-level feature $\mathbf{z}_{v_i}$ of $\mathbf{P}_i$. Similarly, we can also obtain frame-level features $\mathbf{z}_{f_i}$ of $\mathbf{P}_i$.

$$\mathbf{z}_{v_i} = \frac{1}{2}\left(\mathbf{z}_{v_i}^{t_1} + \mathbf{z}_{v_i}^{t_2}\right); \mathbf{z}_{f_i} = \frac{1}{2}\left(\mathbf{z}_{f_i}^{t_1} + \mathbf{z}_{f_i}^{t_2}\right) \tag{4}$$

In the invariance space, we aim to maximize the similarity of $\mathbf{z}_{v_i}$ and $\mathbf{z}_{f_i}$ with $\mathbf{h}_{v_i}$ and $\mathbf{h}_{f_i}$, respectively. We compute the loss function $l(i, \mathbf{z}, \mathbf{h})$ for the positive pair of examples as:

$$C_v = -\log \frac{\exp(s(\mathbf{z}_{v_i}, \mathbf{h}_{v_i})/\tau)}{\sum\limits_{\substack{k=1 \\ k\neq i}}^{N} \exp(s(\mathbf{z}_{v_i}, \mathbf{z}_{v_k})/\tau) + \sum\limits_{k=1}^{N} \exp(s(\mathbf{z}_{v_i}, \mathbf{h}_{v_k})/\tau)} \tag{5}$$

$$C_f = -\log \frac{\exp(s(\mathbf{z}_{f_i}, \mathbf{h}_{f_i})/\tau)}{\sum\limits_{\substack{k=1 \\ k\neq i}}^{N} \exp(s(\mathbf{z}_{f_i}, \mathbf{z}_{f_k})/\tau) + \sum\limits_{k=1}^{N} \exp(s(\mathbf{z}_{f_i}, \mathbf{h}_{f_k})/\tau)} \tag{6}$$

where s, N, τ refers to the same parameters as in Eq. 1. The cross-modal loss function $\mathcal{L}_{cmid}$ for a mini-batch is then formulated as:

$$\mathcal{L}_{cross} = \frac{1}{2N} \sum_{i=1}^{N} [C_v(i, \mathbf{z}, \mathbf{h}) + C_f(i, \mathbf{z}, \mathbf{h})] \tag{7}$$

F. Overall Contrastive Objective

Finally, we obtain the resultant loss function during training as the combination of $\mathcal{L}_{imid}$ and $\mathcal{L}_{cmid}$ where $\mathcal{L}_{imid}$ imposes invariance to point cloud transformation while $\mathcal{L}_{cmid}$ injects the 3D-2D correspondence.

$$\mathcal{L}_{total} = \mathcal{L}_{intra} + \mathcal{L}_{cross} \tag{8}$$

IV. EXPERIMENT

In this section, following other representation learning methods, we use the pre-trained network weights as initialization and fine-tune them in 4D downstream tasks. The performance gain will be a good indicator of measuring the quality of the learned feature. In this section, we cover two 4D point cloud sequence understanding tasks: action segmentation on HOI4D, semantic segmentation on HOI4D in Section IV-B, IV-C respectively. For these tasks, some basic pretraining settings will first be introduced in Section IV-A. We demonstrate that the 2D video encoder involved in pretraining can also obtain strong representations in Section IV-D. We also provide extensive ablation studies to validate our design choices in Section IV-E.

A. Pre-training

We use HOI4D[6] as the dataset for pretraining. HOI4D consists of 2.4M RGB-D egocentric video frames over 4000 sequences interacting with 800 different object instances from 16 categories over 610 different indoor rooms. For a given RGB-D sequences, we first project each depth map to the point cloud. We also use the corresponding RGB images for pretraining. We use 2048 points for each point cloud while we resize the corresponding RGB image to 224×224. In addition to the augmentations applied for point cloud as described in Sec, we perform random crop and color jittering for rendered images as data augmentation.

To draw a fair comparison with the existing methods, we deploy PPTr[2] and P4Transformer[1] as the point cloud feature extractors. Our experiments show that our method consistently outperforms previous approaches in both feature extractors. We use a 3Dconv and Transformer module as the image video feature extractor. We employ a 2-layer MLP as the projection heads which yields a 256-dimensional feature vector projected in the invariant space $\mathbb{R}^d$.

We use SGD optimizer to train the network. The learning rate is set to be 0.01 and we use a learning rate warmup for 5 epochs, where the learning rate increases linearly for the first 10 epochs. The dimension of the features used to calculate the contrastive loss is set to 2048. The temperature coefficient used when calculating contrastive loss is set to 0.07. The time window size is set to 3. As for P4DConv, we set the spatial stride to 32, the radius of the ball query is set to 0.9 and the number of samples is set to 32 by default. With batch-size set to 8, our pre-training method can be implemented on a NVIDIA GeForce 3090 GPUs. After pre-training we discard the image feature extractor $f_{\theta_\mathbf{I}}(.)$ and projection heads $g_{\phi_\mathbf{P}}(.)$ and $g_{\phi_\mathbf{I}}(.)$. All downstream tasks are performed on the pre-trained point cloud feature extractor $f_{\theta_\mathbf{P}}(.)$.

B. Fine-tuning on HOI4D 4D action segmentation

Setup. To demonstrate the effect of our approach, we conduct experiments on the HOI4D action segmentation task. For each point cloud sequence, we need to predict the action label for each frame. We follow the official data split of HOI4D with 2971 training scenes and 892 test scenes. Each sequence has 150 frames, and each frame has 2048 points. We also conduct experiments with other 4D pre-training strategies including STRL(a 4D spatial-temporal contrastive learning method)[27] and VideoMAE(an MAE-based method for video representation learning)[34] and C2P(a 4D Distillation method)[5]. The following metrics are reported: framewise accuracy (Acc), segmental edit distance, as well as segmental F1 scores at the overlapping thresholds of 10%, 25%, and 50%. Overlapping thresholds are determined by the IoU ratio.

Result. As reported in Tab. I, our method has big improvement on both two backbones. For the state-of-the-art backbone PPTr, it can be seen that our method consistently outperforms STRL and VideoMAE by a big margin for all metrics. Considering STRL, its short sequence augmentation can not well guide the model to notice motion cues so it is very hard to leverage temporal information which is actually very important in action segmentation tasks. Simple extension of VideoMAE to point cloud sequence also shows very little improvement. Unlike video pixel tokens which are regular and compact, high down-sampled point cloud tokens have very irregular and sparse patterns. It is hard for the model to learn to predict raw points. Also notice that VideoMAE do self-supervised learning on a point level, which is very hard to learn motion features. C2P formulate 4D self-supervised representation learning as a teacher-student knowledge distillation framework and let the student learn useful 4D representations with the guidance of the teacher. However, it requires complex and time-consuming data generation steps to generate a partial observed point cloud video as a form of data augmentation. This significantly increases the computational cost during pre-training and relies on continuous camera viewpoint generation. Our method is introduced to leverage image video to help the understanding of point cloud videos which finally comes in a satisfying result.

TABLE I
4D ACTION SEGMENTATION ON HOI4D DATASET

Method	Frames	Acc	Edit	F1@10	F1@25	F1@50
P4Transformer	150	71.2	73.1	73.8	69.2	58.2
P4Transformer+C2P	150	73.5	76.8	77.2	72.9	62.4
P4Transformer+ours	150	**76.2**	**79.2**	**78.8**	**75.4**	**65.0**
PPTr	150	77.4	80.1	81.7	78.5	69.5
PPTr+STRL	150	78.4	79.1	81.8	78.6	69.7
PPTr+VideoMAE	150	78.6	80.2	81.9	78.7	69.9
PPTr+C2P	150	81.1	84.0	85.4	82.5	74.1
PPTr+ours	150	**83.7**	**86.0**	**87.3**	**83.2**	**76.0**

C. Fine-tuning on HOI4D semantic segmentation

Setup. To verify that our approach can also be effective on fine-grained tasks, we conducted further experiments on HOI4D for 4D semantic segmentation. For each point cloud frame, there are 8192 points. We follow the official data split of HOI4D with 2971 training scenes and 892 test scenes. During representation learning and training/fine-tuning, we randomly select 1/5 of the whole data to form one epoch for efficient training. We use mean IoU(mIoU) % as the evaluation metric and 39 category labels are used to calculate it. Considering our representation learning method prefers long sequence which has relatively abundant temporal information while the limitation of GPU memory, we set the sequence length as 10 and num points per frame as 4096. Fine-tuning and testing are performed on sequence length of 3 to be consistent with the baseline.

Result. As reported in Tab. II, there is still a performance improvement on the 4D semantic segmentation task, which also shows the effectiveness of our approach for fine-grained feature understanding. Compared with previous methods, our method can effectively extract features with high representational capabilities by introducing cross-modal learning.

D. Finetune image video encoder on action segmentation

Setup. To demonstrate the representational capabilities of the image video encoder, we conduct experiments on

TABLE II

SEMANTIC SEGMENTATION ON HOI4D DATASET

Method	Frames	mIoU
P4Transformer	3	40.1
P4Transformer+C2P	3	41.4
P4Transformer+ours	3	**42.1**
PPTr	3	41.0
PPTr+STRL	3	41.2
PPTr+VideoMAE	3	41.3
PPTr+C2P	3	42.3
PPTr+ours	3	**43.8**

TABLE III

3D ACTION SEGMENTATION ON HOI4D DATASET

Method	Frames	Acc	Edit	F1@10	F1@25	F1@50
MS-TCN	150	44.2	74.7	55.6	47.8	31.8
MS-TCN++	150	42.2	75.8	54.7	46.5	30.3
Asformer	150	46.8	80.3	58.9	51.3	35.0
Asformer + Ours	150	**48.8**	**81.3**	**60.2**	**54.8**	**37.9**

HOI4D action segmentation using image video as input. We consider three representative methods as baseline backbone: MS-TCN[35], MS-TCN++[36] and Asformer[37]. Followed by these work, we also use the I3D[38] feature to train the network and use the pre-trained Transformer as initialization. We use the temporal resolution at 15 fps, and the dimension of the I3D feature for each frame is 2048-d. The following three metrics are reported: framewise accuracy (Acc), segmental edit distance, as well as segmental F1 scores at the overlapping thresholds of 10%, 25%, and 50%. Overlapping thresholds are determined by the IoU ratio.

Result. The results in Tab. III indicate that our method achieves powerful representation capabilities through image video encoder pre-training. It is worth noting that the action segmentation performance on image videos is significantly lower than that of point cloud videos in Table I. This may be due to the accurate three-dimensional coordinate information, point cloud video can provide the spatial position, shape, and motion trajectory of objects which gives point cloud videos an advantage in motion analysis.

E. Ablations and Analysis

In this section, we first conduct an ablation study to verify the design of our method. we conduct all experiments on HOI4D 4D action segmentation task. We also conduct experiments to show whether image video and point cloud video can benefit each other when both modalities are available. We also evaluate our method under limited training data.

1) Ablation study: Our core idea is that point cloud video and image video can complement each other, so we introduce cross-modal contrastive learning to facilitate communication between the two modalities for obtaining powerful representations. To verify our idea, we do ablation studies to show the effectiveness of each objective. We can find the accuracy drops 3.6% without cross-modal contrastive loss and 2.0% without intra-modal contrastive learning. Besides, we can find the accuracy drops 1.4% without frame level contrastive learning and 2.8% without video-level contrastive learning. The above results demonstrate the effectiveness of each objective.

TABLE IV

ACTION SEGMENTATION ON HOI4D DATASET

Method	Frames	Acc	Edit	F1@10	F1@25	F1@50
Asformer	150	46.8	80.3	58.9	51.3	35.0
Asformer + Ours	150	48.8	81.3	60.2	54.8	37.9
PPTr	150	77.4	80.1	81.7	78.5	69.5
PPTr+ours	150	**83.7**	**86.0**	**87.3**	**83.2**	**76.0**
PPTr + Asformer	150	82.7	84.7	85.2	81.9	74.3

TABLE V

DATA-EFFICIENT LEARNING ON HOI4D ACTION SEGMENTATION.

%Data	Scratch	C2P	Ours
10%	44.0	53.4	57.8
20%	53.9	69.9	74.1
40%	69.9	75.4	76.9
80%	76.7	79.0	80.2

2) Multi-modal Fusion v.s Cross-modal pre-training: We conducted experiments to verify whether the fusion of two modalities can also improve performance. As shown in the Tab. IV, when using a single modality, point cloud videos achieve the best performance. When we cascade the features of the two modalities, noticeable performance improvement can be observed, but the fusion result is still lower than the performance that pre-trained point cloud video networks can achieve. This indicates that our pre-training method significantly promotes communication between cross-modalities and provides a good pre-trained model for point cloud video encoders. Additionally, it is worth mentioning that in real testing scenarios, it may not be possible to obtain data from both modalities simultaneously, which also hinders the possibility of feature fusion.

3) Data-efficient 4D Representation Learning: We evaluate our method under limited training data on the HOI4D Action Segmentation task. For all data-efficient experiments, our limited data are randomly sampled from the full dataset of HOI4D Action Segmentation dataset. As shown in Tab. V, our pre-training method shows consistently outstanding performance in the case of lack of data, compared with previous methods. This indicates that our method can still formulate a strong 4D representation under limited data.

V. CONCLUSION

This paper introduces CrossVideo, which applies self-supervised cross-modal contrastive learning in point cloud video understanding. We propose to leverage the cross-modal relationship between point cloud videos and image videos to learn meaningful feature representations. We employ intra-modality and cross-modality contrastive learning to facilitate effective point cloud video understanding. Additionally, we propose to contrast features of the two modalities at different levels. Experimental results demonstrate that our method outperforms previous state-of-the-art approaches.

REFERENCES

[1] H. Fan, Y. Yang, and M. Kankanhalli, "Point 4d transformer networks for spatio-temporal modeling in point cloud videos," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2021.

[2] H. Wen, Y. Liu, J. Huang, B. Duan, and L. Yi, "Point primitive transformer for long-term 4d point cloud video understanding," in *European Conference on Computer Vision*. Springer, 2022, pp. 19–35.

[3] H. Shi, J. Wei, R. Li, F. Liu, and G. Lin, "Weakly supervised segmentation on outdoor 4d point clouds with temporal matching and spatial graph propagation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 840–11 849.

[4] T. Khurana, P. Hu, D. Held, and D. Ramanan, "Point cloud forecasting as a proxy for 4d occupancy forecasting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] Z. Zhang, Y. Dong, Y. Liu, and L. 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*, 2023, pp. 17 661–17 670.

[6] Y. Liu, Y. Liu, C. Jiang, K. Lyu, W. Wan, H. Shen, B. Liang, Z. Fu, H. Wang, and L. Yi, "Hoi4d: A 4d egocentric dataset for category-level human-object interaction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 21 013–21 022.

[7] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.

[8] X. Liu, M. Yan, and J. Bohg, "Meteornet: Deep learning on dynamic 3d point cloud sequences," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9246–9255.

[9] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[10] H. Fan, X. Yu, Y. Ding, Y. Yang, and M. Kankanhalli, "Pstnet: Point spatio-temporal convolution on point cloud sequences," *arXiv preprint arXiv:2205.13713*, 2022.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[12] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[13] Z. Liu, Y. Chen, J. Li, P. S. Yu, J. McAuley, and C. Xiong, "Contrastive self-supervised sequential recommendation with robust augmentation," *arXiv preprint arXiv:2108.06479*, 2021.

[14] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *European conference on computer vision*. Springer, 2022, pp. 604–621.

[15] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. 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*, 2022, pp. 19 313–19 322.

[16] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li, "Pointclip: Point cloud understanding by clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8552–8562.

[17] Y. Liu, L. Yi, S. Zhang, Q. Fan, T. Funkhouser, and H. Dong, "P4contrast: Contrastive learning with pairs of point-pixel pairs for rgb-d scene understanding," *arXiv preprint arXiv:2012.13089*, 2020.

[18] D. Bear, C. Fan, D. Mrowca, Y. Li, S. Alter, A. Nayebi, J. Schwartz, L. F. Fei-Fei, J. Wu, J. Tenenbaum, *et al.*, "Learning physical graph representations from visual scenes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6027–6039, 2020.

[19] H. Wang, Q. Liu, X. Yue, J. Lasenby, and M. J. Kusner, "Unsupervised point cloud pre-training via occlusion completion," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9782–9792.

[20] R. Zhang, Z. Guo, P. Gao, R. Fang, B. Zhao, D. Wang, Y. Qiao, and H. Li, "Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training," *Advances in neural information processing systems*, vol. 35, pp. 27 061–27 074, 2022.

[21] S. Xie, J. Gu, D. Guo, C. R. Qi, L. Guibas, and O. 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*. Springer, 2020, pp. 574–591.

[22] Y. Rao, B. Liu, Y. Wei, J. Lu, C.-J. Hsieh, and J. Zhou, "Randomrooms: Unsupervised pre-training from synthetic shapes and randomized layouts for 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3283–3292.

[23] Z. Zhang, R. Girdhar, A. Joulin, and I. Misra, "Self-supervised pretraining of 3d features on any point-cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 252–10 263.

[24] Y. Chen, J. Liu, B. Ni, H. Wang, J. Yang, N. Liu, T. Li, and Q. Tian, "Shape self-correction for unsupervised point cloud understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8382–8391.

[25] J. Hou, B. Graham, M. Nießner, and S. Xie, "Exploring data-efficient 3d scene understanding with contrastive scene contexts," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 587–15 597.

[26] Y. Chen, M. Nießner, and A. Dai, "4dcontrast: Contrastive learning with dynamic correspondences for 3d scene understanding," in *European Conference on Computer Vision*. Springer, 2022, pp. 543–560.

[27] S. Huang, Y. Xie, S.-C. Zhu, and Y. Zhu, "Spatio-temporal self-supervised representation learning for 3d point clouds," *arXiv preprint arXiv:2109.00179*, 2021.

[28] R. Arandjelovic and A. Zisserman, "Look, listen and learn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 609–617.

[29] K. Desai and J. Johnson, "Virtex: Learning visual representations from textual annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 162–11 173.

[30] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[31] P. Morgado, N. Vasconcelos, and I. Misra, "Audio-visual instance discrimination with cross-modal agreement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 475–12 486.

[32] C. Xu, S. Yang, T. Galanti, B. Wu, X. Yue, B. Zhai, W. Zhan, P. Vajda, K. Keutzer, and M. Tomizuka, "Image2point: 3d point-cloud understanding with 2d image pretrained models," *arXiv preprint arXiv:2106.04180*, 2021.

[33] M. Afham, I. Dissanayake, D. Dissanayake, A. Dharmasiri, K. Thilakarathna, and R. 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 (CVPR)*, June 2022, pp. 9902–9912.

[34] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems*, 2022.

[35] Y. A. Farha and J. Gall, "Ms-tcn: Multi-stage temporal convolutional network for action segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3575–3584.

[36] S.-J. Li, Y. AbuFarha, Y. Liu, M.-M. Cheng, and J. Gall, "Ms-tcn++: Multi-stage temporal convolutional network for action segmentation," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[37] F. Yi, H. Wen, and T. Jiang, "Asformer: Transformer for action segmentation," *arXiv preprint arXiv:2110.08568*, 2021.

[38] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6299–6308.