TEXAS
The University of Texas at Austin

Multi-View Representation is What You Need for Point-Cloud Pre-Training

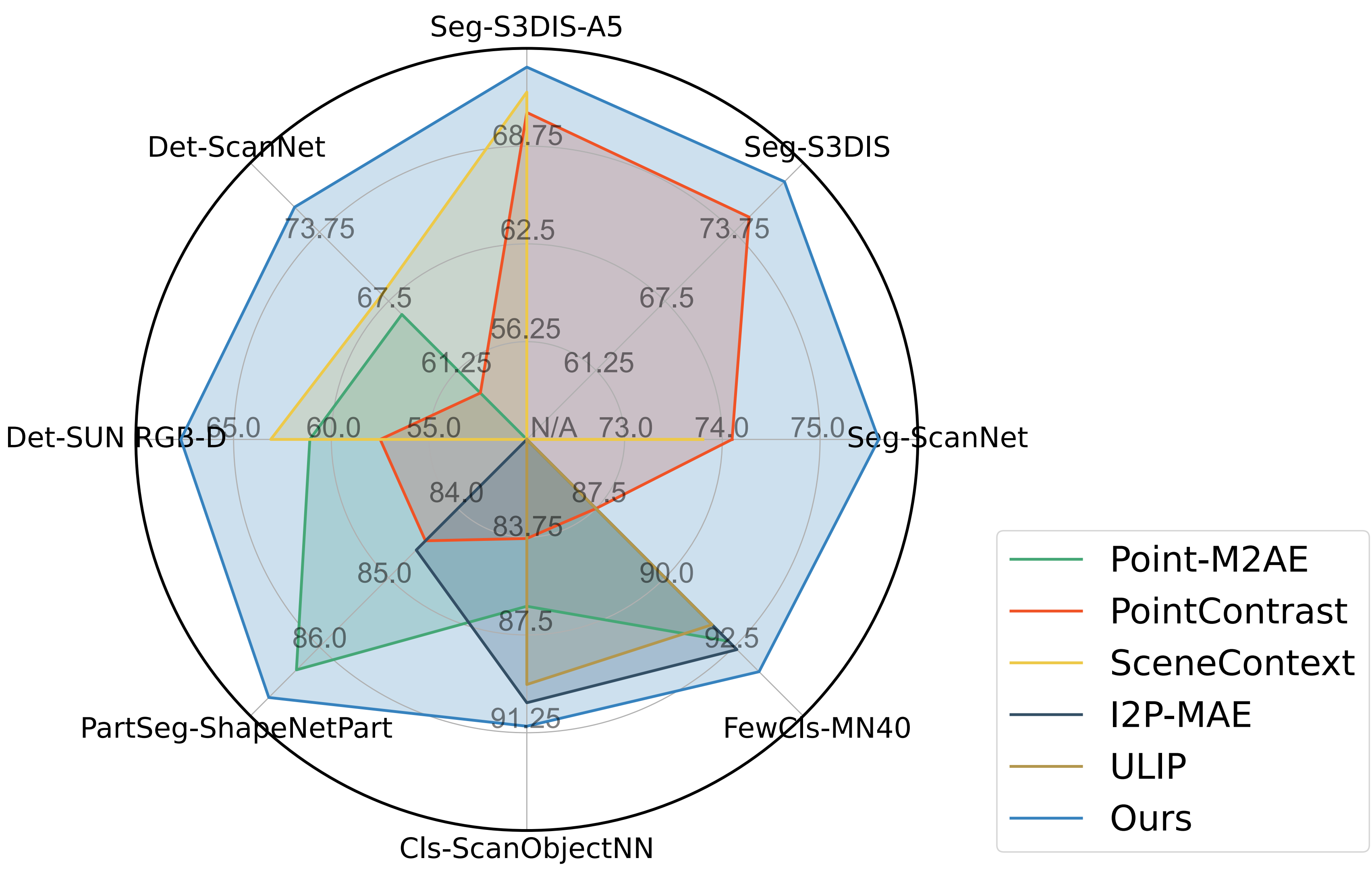
Siming Yan, Chen Song, Youkang Kong, Qixing Huang

Microsoft

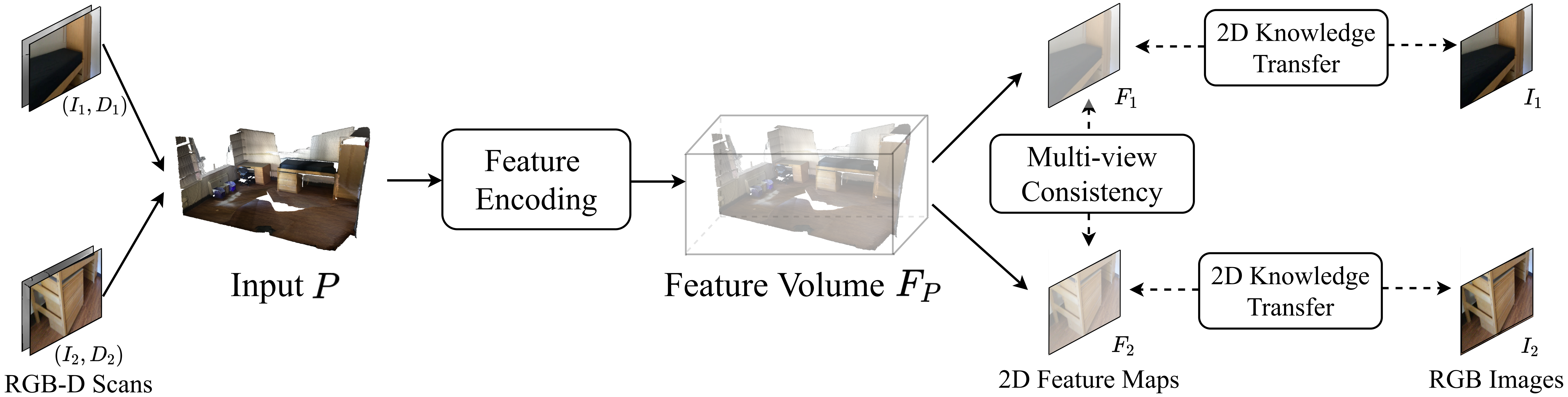
Introduction

- Formulate point-cloud pre-training as learning a multi-view consistent 3D feature volume.
- Leverage pre-trained 2D image-based models to supervise 3D pre-training
- Develop an auxiliary pre-task where the goal is to predict the multi-view pixel-wise correspondences from the 2D pixel embeddings.
- Conduct extensive experiments to demonstrate the effectiveness of our approach

Experiment Results

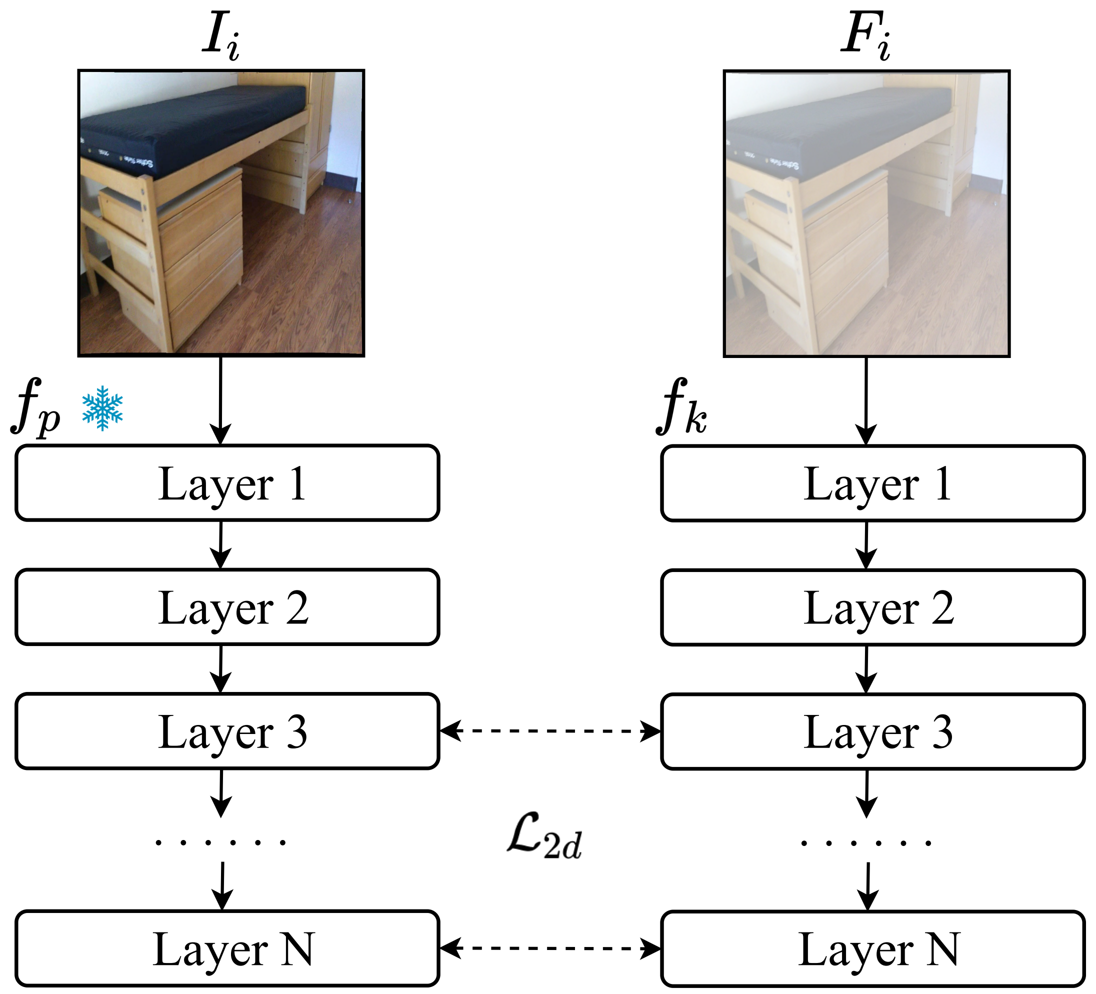


Method Overview



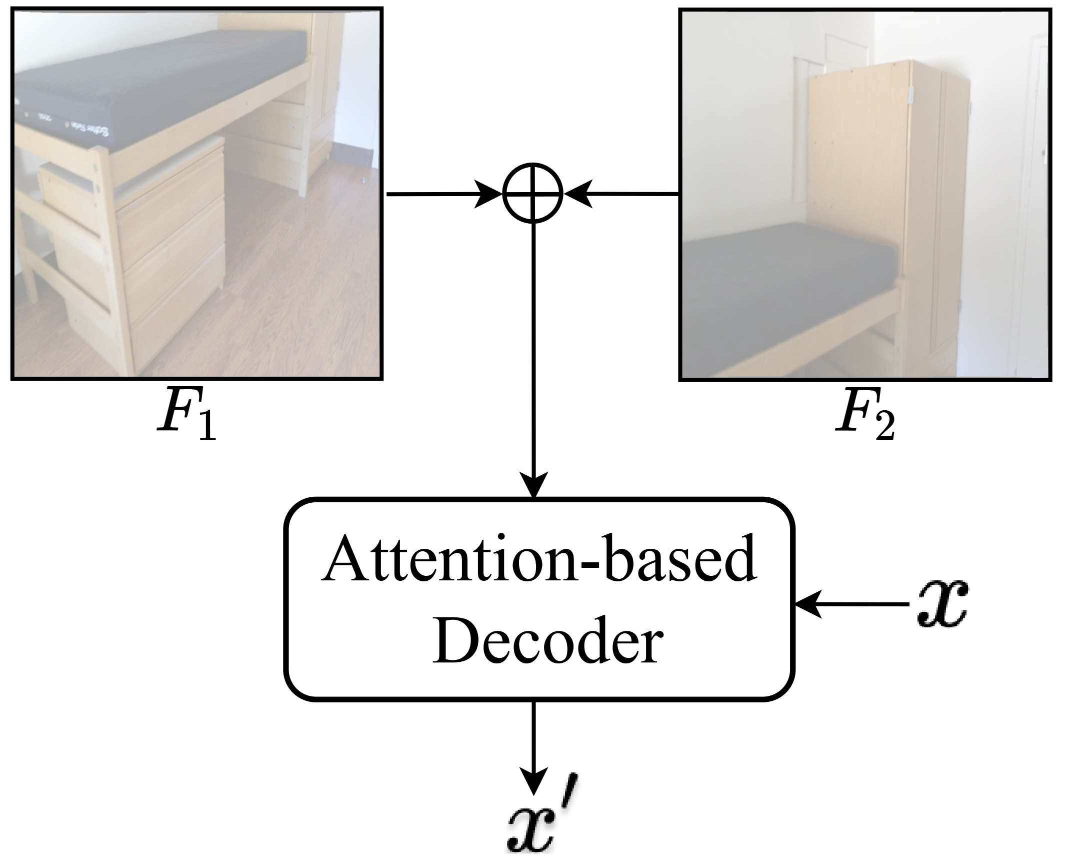
2D Knowledge Transfer Module

- Frozen DINOv2 as 2D pre-trained model
- Knowledge distillation loss



Multi-view Consistency Module

- Predict feature correspondences from 2D embedding
- Cross-attention layer input: 1.concat 2-view feature maps; 2. query point position from first view


F_1
F_2
Attention-based Decoder
x
x'