Dynamic Point Cloud Interpolation

Anique Akhtar[1], Zhu Li[1], Geert Van der Auwera[2], Jianle Chen[2]
[1]University of Missouri-Kansas City
[2]Qualcomm Technologies Inc.

Introduction

Dense photorealistic point clouds can depict real-world dynamic objects in high resolution and with a high frame rate. Frame interpolation of such dynamic point clouds would enable the distribution, processing, and compression of such content. In this work, we propose a first point cloud interpolation framework for photorealistic dynamic point clouds. Given two consecutive dynamic point cloud frames, our framework aims to generate intermediate frame(s) between them.

We evaluate our framework on high-resolution point cloud datasets used in MPEG, and JPEG Pleno standards. The quantitative and qualitative results demonstrate the effectiveness of the proposed method.

Contributions

- We propose a first of its kind dynamic point cloud interpolation framework for dense photo-realistic point clouds used in AR/VR/MR and telepresence.
- Given two consecutive dynamic point cloud frames, our framework aims to generate intermediate frame(s) between them.
- The work proposes three different modules: the encoder network, the fusion network, and the multi-scale point cloud synthesis module.
- The encoder module extracts discriminative features from frames at four different scales.
- The fusion network takes features at different scales from consecutive frames, concatenates them into 4D features ,then utilizes 4D convolutions to merge consecutive frame features.
- Finally, the multi-scale point cloud synthesis module hierarchically interpolates the target frame at different resolutions.

Encoder and Synthesis Module

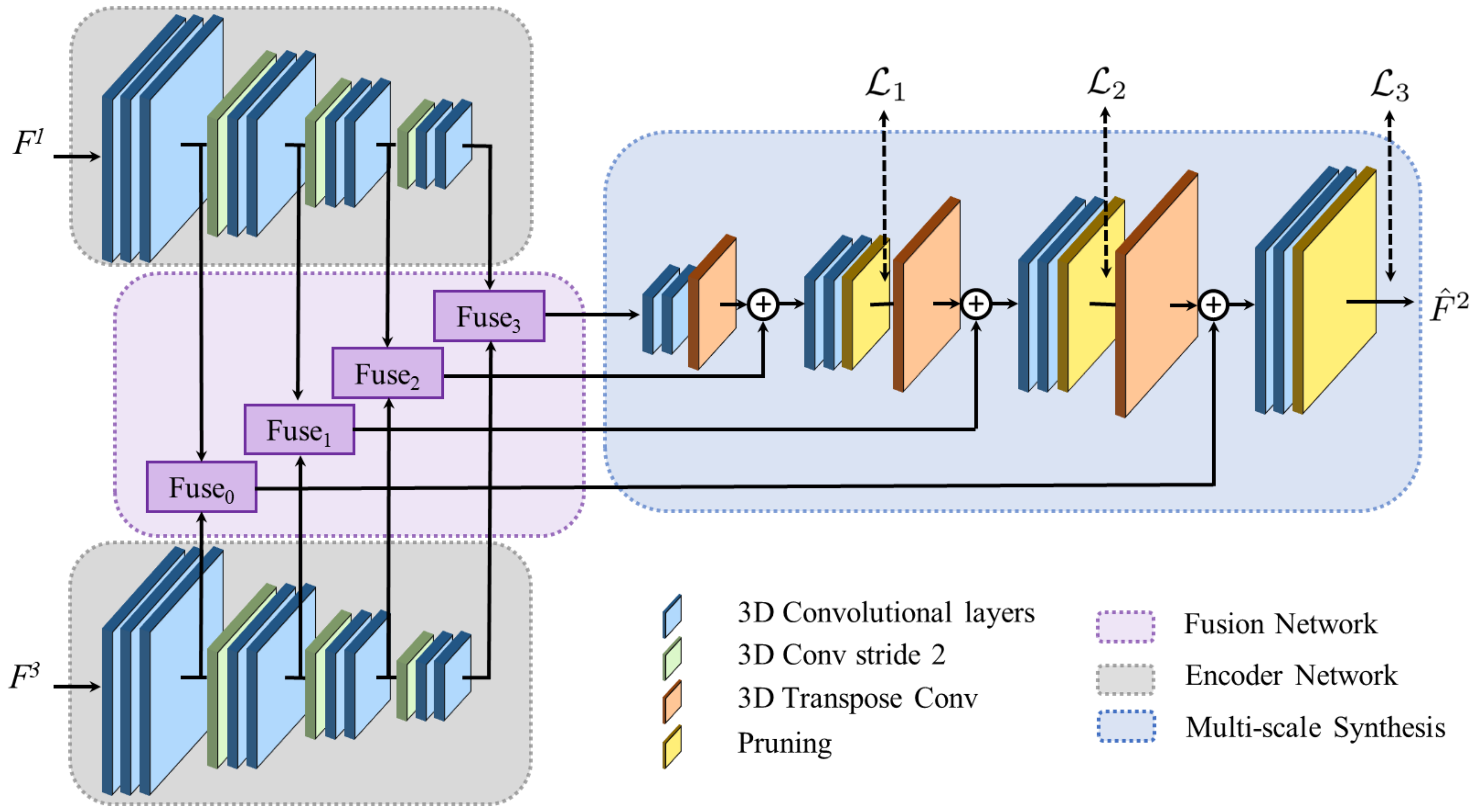
- Two encoder modules are used to extract multi-scale features from the frames. These encoders share the same weight.
- The multi-scale point cloud synthesis module hierarchically reconstructs the interpolated point cloud intermediate frame through progressive upscaling.
- The encoder and synthesis modules are pre-trained in a typical autoencoder way using reconstruction loss (binary voxel occupancy loss) on a static point cloud objects dataset (ShapeNet).

Results Explained

Table 1 shows the results of the following frameworks:

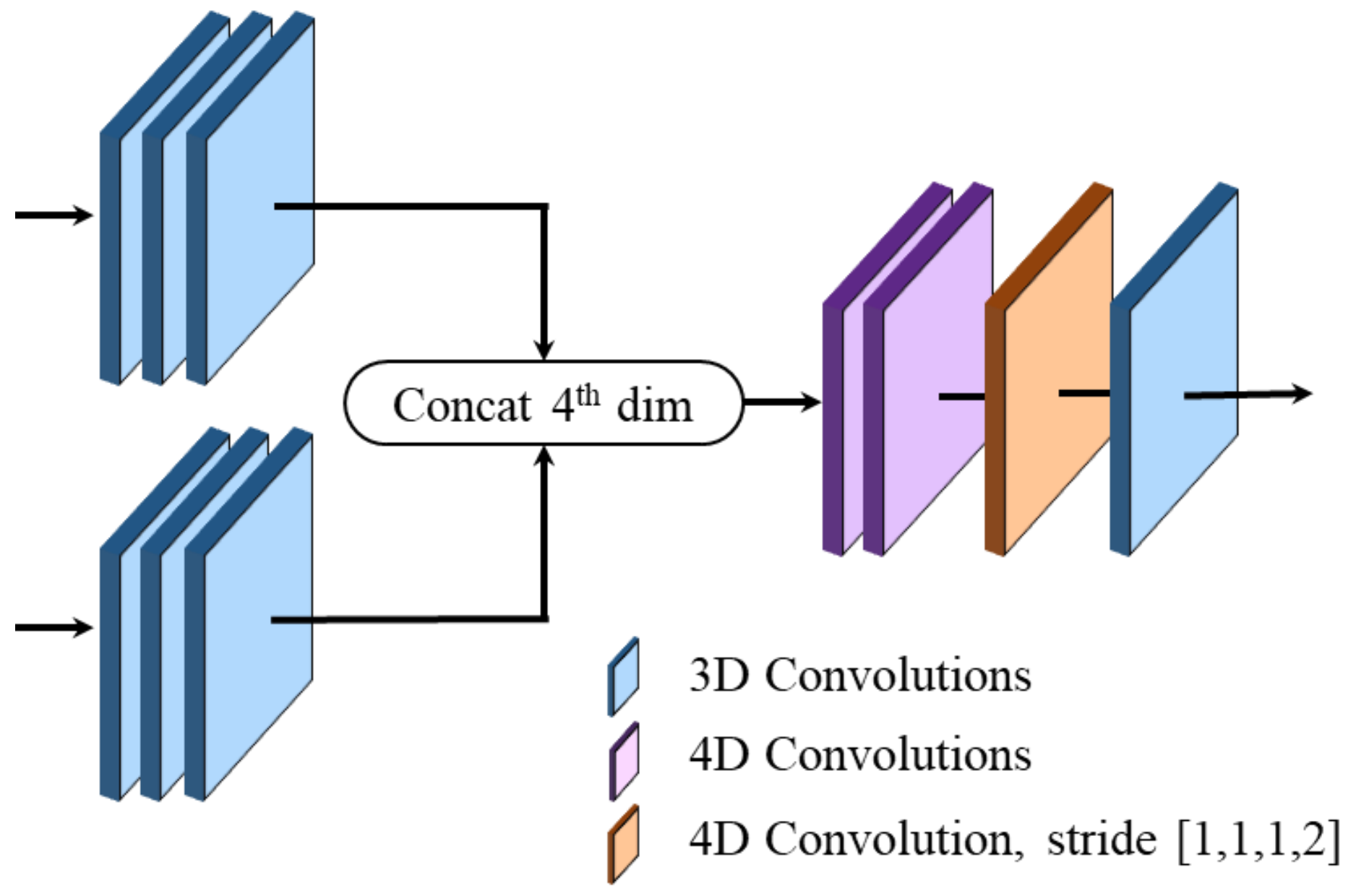
- **Ours** is the framework proposed in this work.
- **Ours-w/o Pretrained** is the framework where the Encode ris not pretrained on ShapeNet.
- **Ours-Single Loss** framework employs only a single loss (L_3) and no longer employs losses L_1 and L_2.
- **Ours-Fuse3D** framework utilizes 3D convolutions in the fuse block rather than the 4D convolutions.

System Model



System Model. F^1 is the previous frame and F^3 is the next frame, whil$\hat{F}^2$ is the interpolated current frame.

Fusion Network



The fusion network utilizes a novel 4D fuse block that merges features from two consecutive point cloud frames into a single feature. In the fuse block, the features pass through 3D convolutions and then are concatenated in the 4th dimension. Afterward, a 4D convolution with stride in only the 4th dimension (stride = [1, 1, 1, 2]) is applied to downsample the features back to 3D.

Loss Functions

Rather than upscaling the features directly to the full scale, our network synthesizes the interpolated point cloud at different resolutions by employing multiple loss functions. The multi-scale synthesis module produces the interpolated point cloud at three different resolutions. As shown in system model, we train our network using three different loss functions:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3 \quad ,$$

We perform voxel classification using binary cross-entropy loss to compare the voxel occupancy prediction from the network and the ground truth (original) point cloud.

Each of the loss function is defined as:

$$L_{BCE} = -\frac{1}{N}\sum_i (x_i log(p_i) + (1-x_i)log(1-p_i))$$

Results

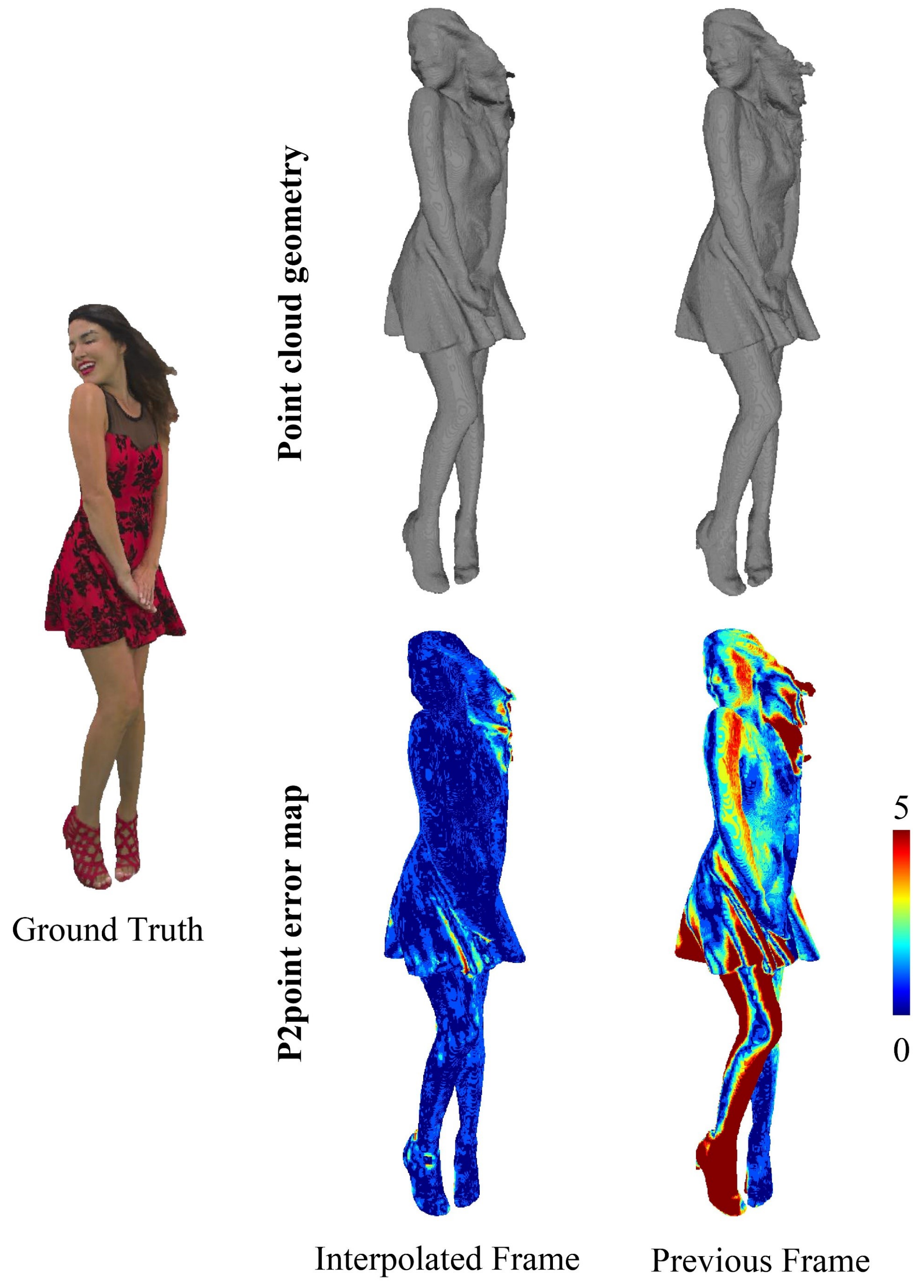
Method	*redandblack*		*queen*		*basketball*		*exercise*		Avg	
	CD↓	PSNR↑	CD↓	PSNR↑	CD↓	PSNR↑	CD↓	PSNR↑	CD↓	PSNR↑
Identity	1623.69	53.68	45.29	70.75	113.65	71.73	146.54	71.19	482.29	66.84
Ours	**386.22**	**56.88**	**30.44**	**76.08**	**80.96**	**74.54**	**110.20**	**75.48**	**151.96**	**70.75**
Ours-w/o Pretrained	502.86	55.44	35.64	74.81	90.37	73.30	117.62	74.34	186.63	69.47
Ours-Single Loss	575.91	55.40	36.64	74.00	89.74	73.03	121.54	73.37	205.96	68.95
Ours-Fuse3D	865.72	54.64	37.01	73.64	100.18	72.14	125.47	73.12	282.10	68.39

Table 1. Evaluation results of our interpolation method using Chamfer Distance (CD (10^{-2})) and MSE PSNR (dB).

Implementation

- The model is implemented by employing sparse tensors and sparse convolution using Minkowski Engine [1]
- The input frames are pre-processed where they are voxelized and converted into sparse tensors.

Visual Results



Conclusion

This work proposes the first dynamic point cloud interpolation framework for dense high-resolution point clouds. While the previous point cloud interpolation methods are limited to point cloud scenes, our framework is able to process and interpolate frames for high-resolution dynamic point clouds. We employ a pretrained multi-scale encoder module to extract features at multiple scales. We introduce a novel 4D feature fusion module that utilizes 4D learning to merge 3D features from two consecutive frames at multiple scales. Finally, our multi-scale point cloud synthesis module hierarchically reconstructs the interpolated point cloud frame at different resolutions. We test our framework on a diverse set of high-resolution dynamic point cloud sequences. The evaluation results validate our network design and demonstrate the effectiveness of our method.

References

1. 4D Spatio-Temporal ConvNets: Minkowski Convolutional Neural Networks. C. Choy, J. Gwak, and S. Savarese, In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (CVPR’19)

UMKC School of Computing and Engineering