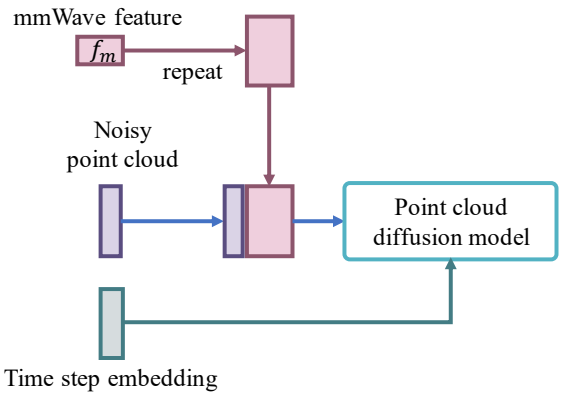
# mmDiffusion: mmWave Diffusion for Sequential 3D Human Dense Point Cloud Generation

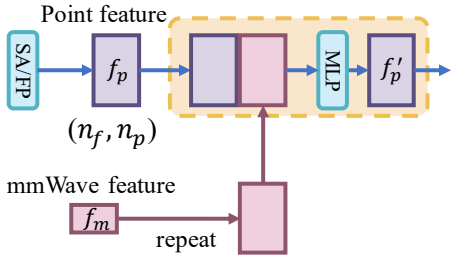
Supplementary Material

## 6. Architecture of Concat-V1 & Concat-V2

We provide additional details about architectures of the two condition strategies in the ablation study section: Concat-V1 and Concat-V2. As shown in Fig. 8, both of them fuse the mmWave feature into the point cloud feature via concatenation. The main difference lies in the order. Concat-V1 adopts a direct way which concatenates the mmWave feature to the original input point cloud before sent to the point cloud diffusion model. In contrast, Concat-V2 concatenates the mmWave feaure to the imtermediate point cloud feature during the diffusion process. As demonstrated in the result part, Concat-V2 is a more efficient way to condition the point cloud diffusion process. However, both of them are worse than the proposed MMC.




(a) Concat-V1




(b) Concat-V2

Figure 8. Architectures of the other two condition strategies in the ablation study part.

## 7. Visualization on diffusion process

Fig. 9 illustrates qualitative results of mmDiffusion applied to mmWave signals for the generation of 3D human point clouds. It showcases the gradual refinement process of the human figure reconstructions through the diffusion model stages. The columns depict the evolutionary stages of the point clouds, beginning with a randomly sampled Gaussian distribution and culminating in the finely detailed final model. We further test our method on some challenging cases. As shown in Fig. 10, mmDiffusion can still achieve promising results under dark and occlusion scenes, demonstrating the resilience of radar signals and our approach to challenging lighting and obstructed visual scenarios.

## 8. Temporal Consistency (TC) Visualization

The proposed metric, Temporal Consistency (TC), is employed to quantify the degree of variation between two consecutive frames. Given the brief time interval between these frames, one would expect the variation to be minimal. Consequently, the value (TC) representing this consistency should be large.

Figure 11 illustrates the computational process for two specific examples. In the upper row of the figure, a point cloud from frame $t$ comprising 1024 points is considered. Initially, correspondences for each point in frame $t$ are established with the points in frame $t+1$, resulting in a total of 1024 correspondences (depicted as green lines). However, not all these correspondences are accurate. A valid mask $M_t$ is applied to filter out 73 incorrect correspondences (depicted as pink lines). Subsequently, 139 filtered correspondences (depicted as blue lines) are identified, whose distances exceed the prescribed threshold $thr$ in Equation (8). TC is then computed as: $(1024-73-139)/(1024-93) = 812/951 = 0.854$ for this example. Similarly, the TC value for the second example is computed as $0.963$.

These TC values enable a quantitative comparison of the variations between the two examples. Specifically, the frames in the second example exhibit greater temporal stability compared to those in the first example, aligning with the qualitative observations.

## 9. Additional Visualization Tool

We provide additional qualitative examples of our method in a more photorealistic way. In Fig. 12, we put the generated 3D human point clouds in a 3D scene to make them look more like true 3D shapes.

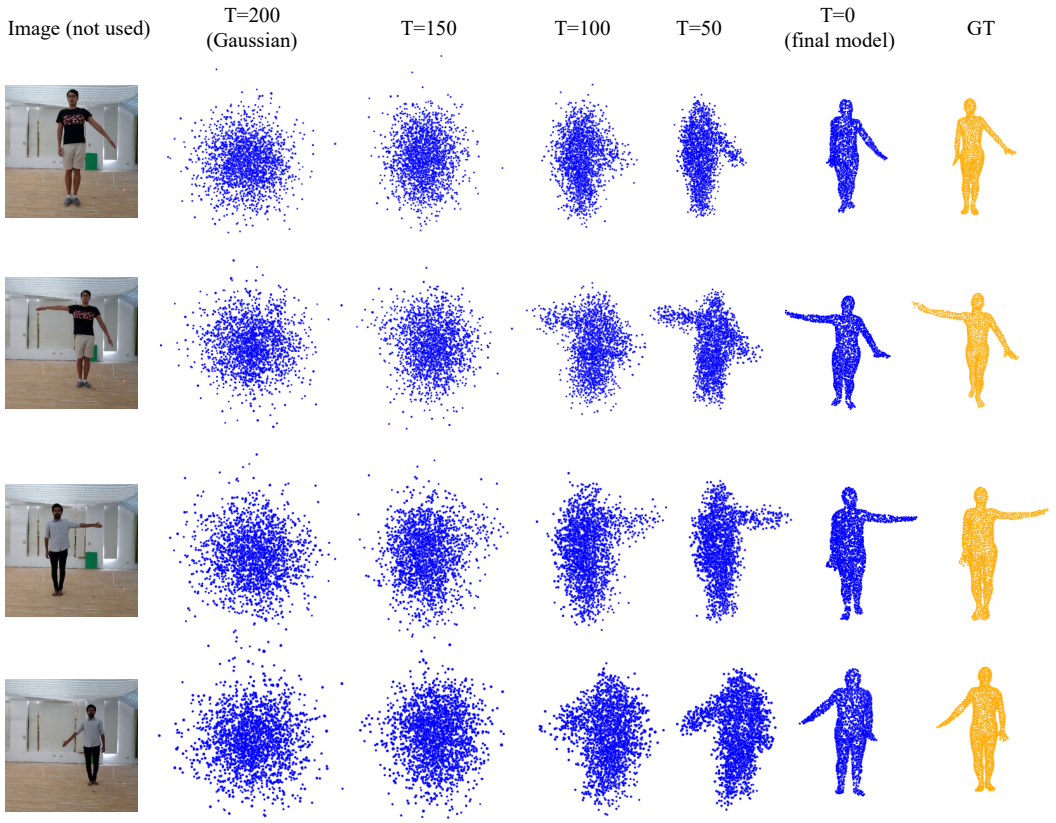



Figure 9. Examples of intermediate steps in the conditioned diffusion process. The first column shows the images which are not used in our method. The subsequent five columns show the evolution of the point cloud from a randomly sampled Gaussian 3D point cloud to a final shape. The last column shows ground truth.

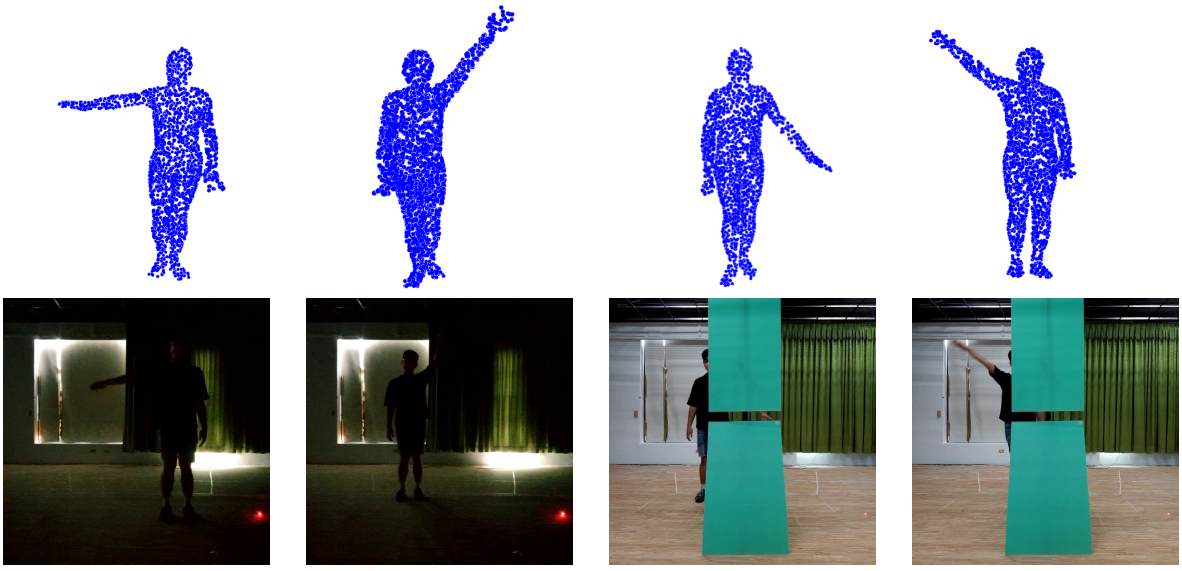


Figure 10. Our results under challenging scenes.

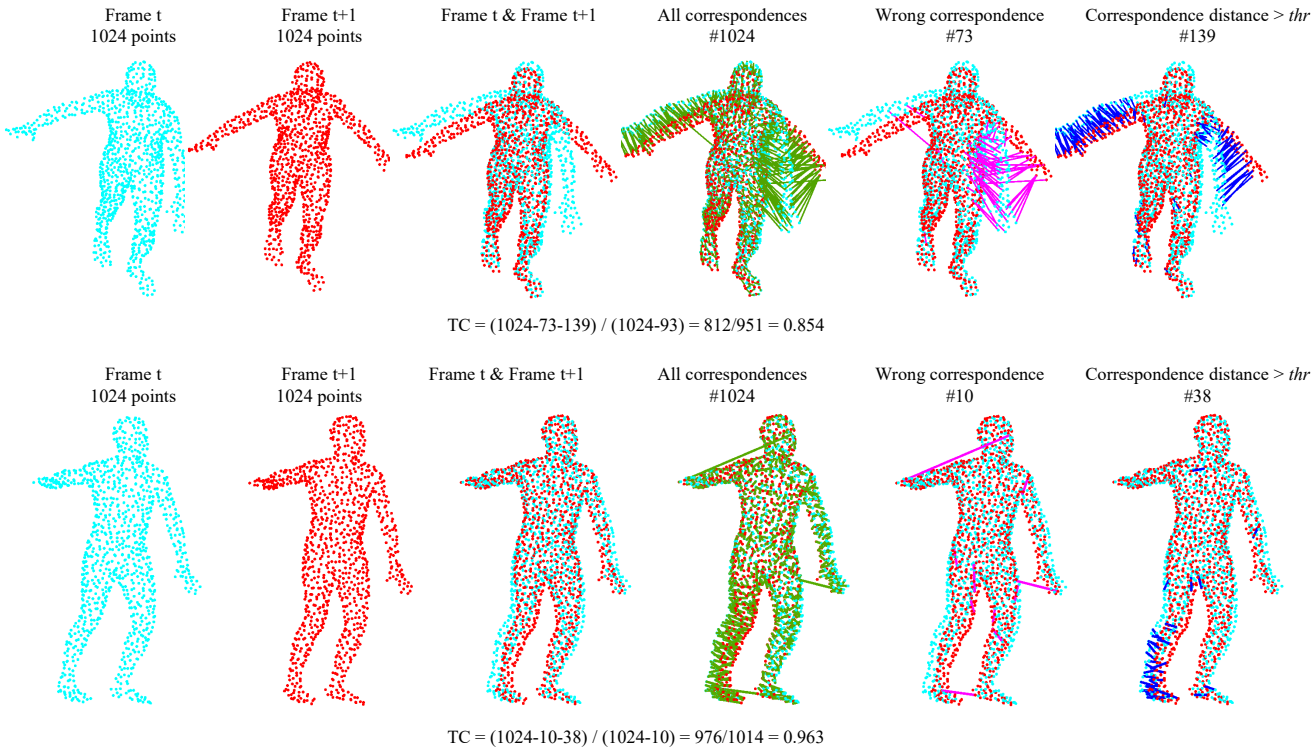



Figure 11. Visualization of the temporal consistency calculation for two examples. 1-st and 2-nd columns: two adjacent human point cloud frames with 1024 points. 3-rd column: two adjacent frames overlapped together. 4-th column: correspondences for all points. 5-th column: wrong correspondences filtered by the valid mask $M_t$ in Sec. 3.3. 6-th column: valid correspondences whose distances are larger than the threshold $thr$ in Eq. (8).

Figure 12. Visualization of results of the proposed mmDiffusion in a more 3D way.