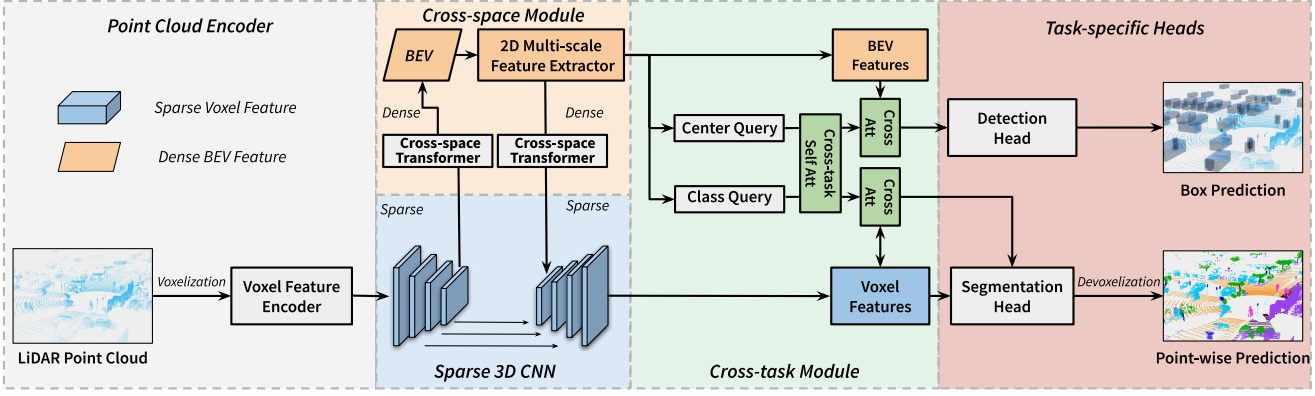

Figure 2: **The architecture of LiDARFormer.** Our network first transforms the point cloud into a sparse voxel map. Next, sparse 3D CNN is used to extract voxel feature representation. Between the encoder and the decoder, we use a Cross-space Transformer (**XSF**) module to learn long-range information in the BEV map. Additionally, we use a cross-task transformer decoder (**XTF**) to extract class-level and object-level feature representations, which are fed into task-specific heads to generate the detection and segmentation predictions.