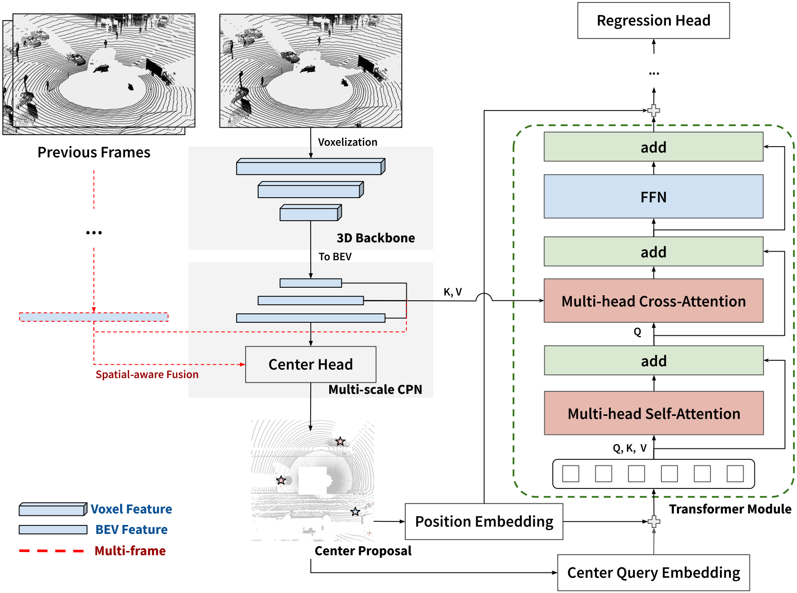

Fig. 2: The overall architecture of CenterFormer. The network consists of four parts: a voxel feature encoder that encodes the raw point cloud into a BEV feature representation, a multi-scale center proposal network (CPN), the center-based transformer decoder, and a regression head to predict the bounding box.