



Fig. 3: Illustration of the cross-attention layer. (Left) Multi-scale cross-attention. (Right) Multi-scale deformable cross-attention.