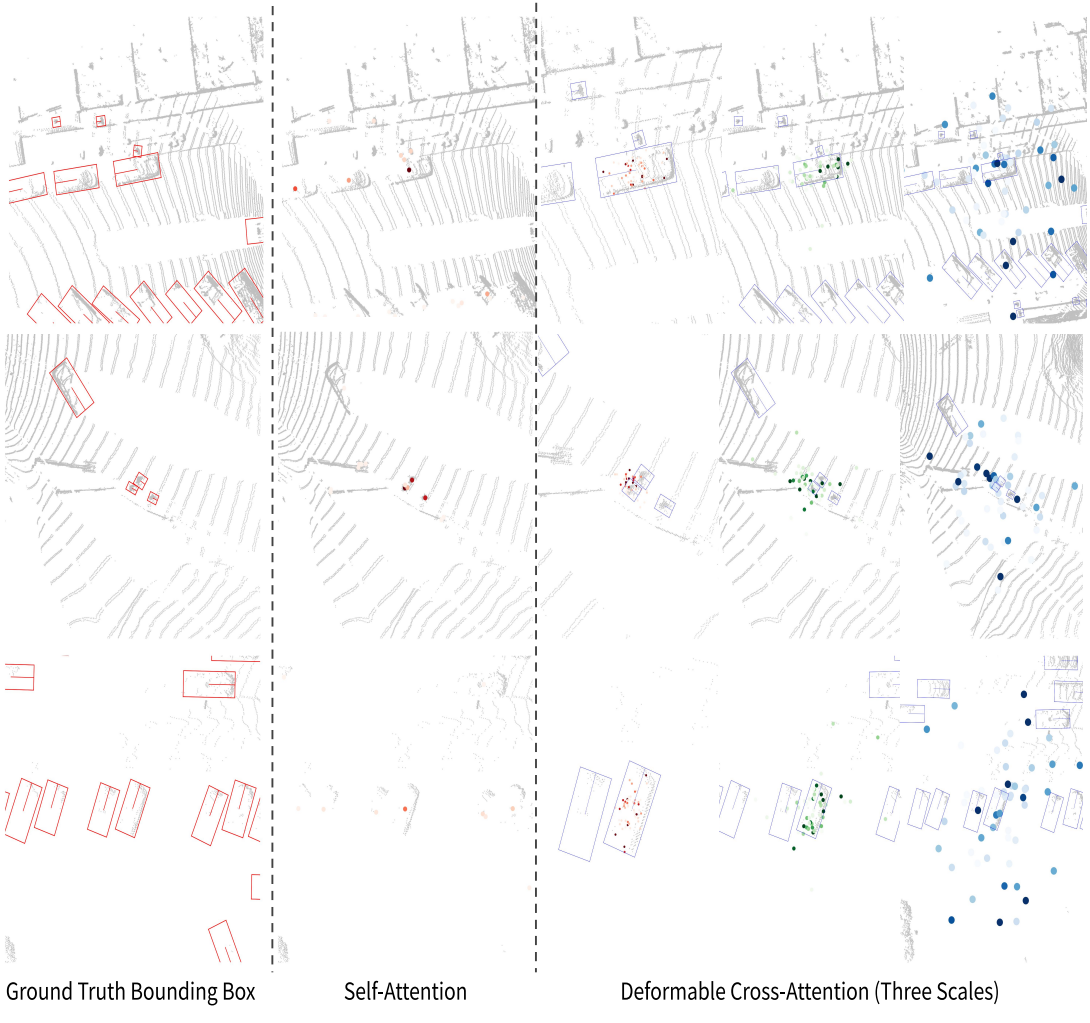

Fig. 5: **The visualization of the learned self- and cross-attention weight.** The lightness or darkness of the color represents the value of attention weight. The red box denotes the ground truth box and blue box denotes the predicted box. In cross-attention, the sampled keypoints are drawn with red, green and blue for the scale from low to high. Best viewed in color.