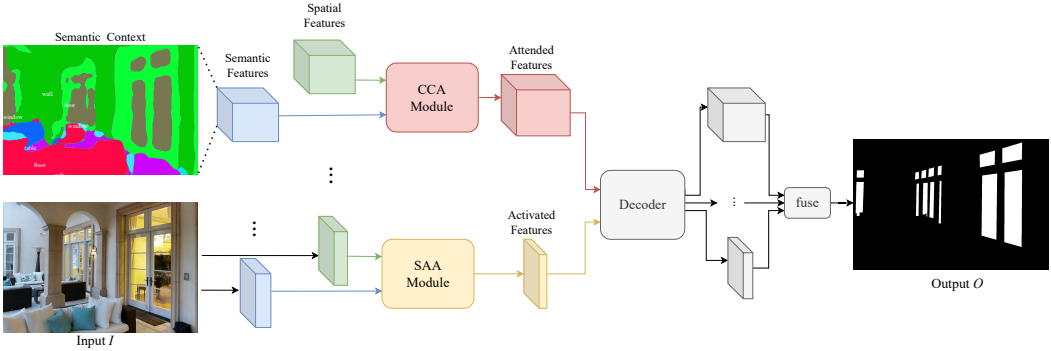

Figure 4: Model Architecture. We first feed the input image into two backbones to capture semantic knowledge, and spatial location features separately. Together with the semantic encodings, low-level features first get selectively activated by the SAA Module with respect to the decoupled features. The CCA Module is placed at a higher level to learn the relationships between contextual meanings and locations of objects. Features from multiple stages are aggregated by the UPerNet decoder to produce the output map, along with the intermediate feature maps for supervision.