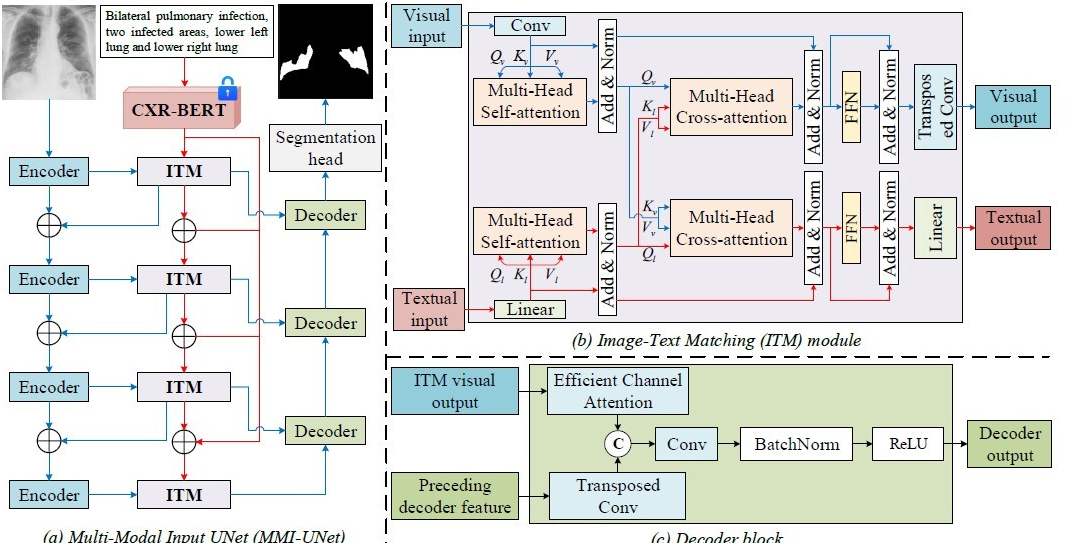

Fig. 1: The overview of the proposed Multi-Modal Input UNet (a) with Image-Text Matching module (b) and decoder block (c).