Appendix 2

Image Segmentation Model Based on Wearable Ultrasound

The DSTA-Net model was specifically developed to automatically measure the IJV Max Area, and IJV Min Area, CCA Area, IJV Max/CCA Area, IJV Ratio over one respiratory cycle. The architecture comprises a shared encoder (E) and two decoders (D1 and D2) with distinct roles. Encoder E extracts deep features from video frames using alternating convolutional and pooling layers, halving spatial dimensions, while doubling feature channels to reduce complexity and enhance abstraction. This process integrates labeled and unlabeled image data for accurate video reconstruction. Decoder D1 restores image resolution and spatial dimensions by combining up-sampling and convolutional layers; it contains a temporal attention module to capture video time information, and utilizes skip connections to merge the encoder and decoder features, preserving context. This structure is based on U-Net, but optimized for spatiotemporal video data. Decoder D2 manages unlabeled frames, omitting temporal attention and skip connections to simplify the model and reduce resource use. During training, labeled frames and their neighbors are input into D1 for supervised learning, while unlabeled frames are transmitted to both decoders. D1 generates pseudo-labels for D2, guiding its learning. Both decoders use cross-entropy loss for training. In inference, D2 processes input images for final high-quality segmentation. This dual-decoder design effectively leverages limited labeled data and enhances temporal information usage, excelling in scenarios with sparse annotations.