The image annotation process consists of five sequential modules, as illustrated in Fig. 1:

1. Double-channel based low-level shallow and deep modality feature extraction from image without DSM data.
   1) Extract multiple types of low-level shallow modality features including the LBP, SIFT, and Color features for each pixel of the input image;
   2) Extract low-level deep modality features which consist of the feature maps of Pool2, Conv4, and Pool5 layers from CNN network.

2. Mid-level feature construction within superpixels;
   Generate a mid-level feature vector for each superpixel by integrating the low-level features of all the pixels within superpixel segmentations;

3. Deep belief network (DBN) based high-level feature learning;
   Use DBNs model to further construct a high-level feature vector from mid-level feature vector for each superpixel.

4. Restricted boltzmann machine (RBM) based feature fusion;
   Employ a RBM model to generate the final representation by fusing high-level shallow and deep modality feature.

5. After obtaining high-level feature, we perform one-versus-all annotation by using softmax regression.

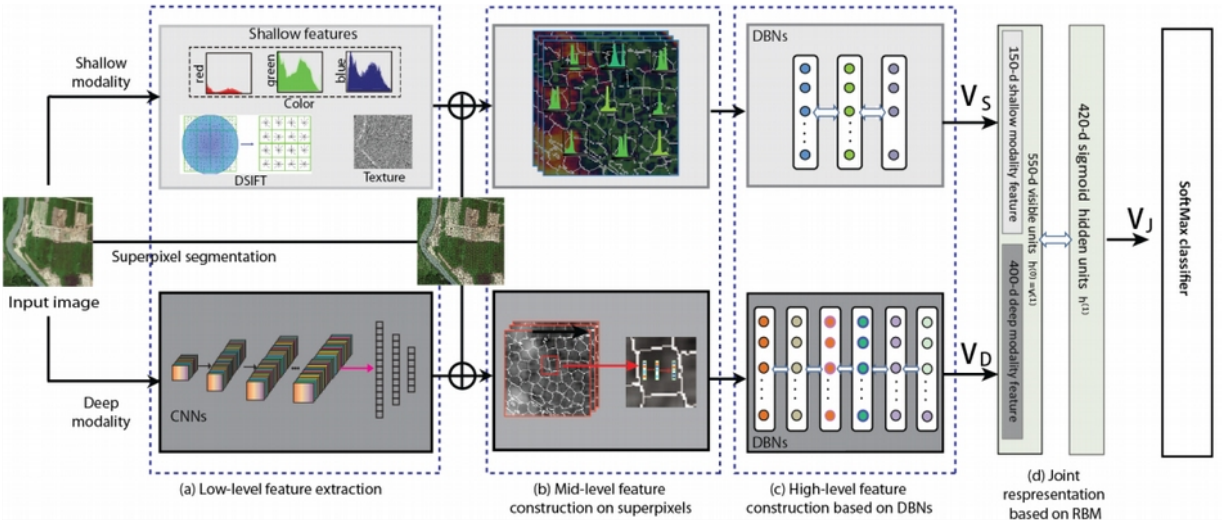

Figure 1: Flowchart of image annotation.