Recent studies have witnessed the successes of using 3D CNNs for video action recognition. However, most 3D models are built upon RGB and optical flow str