The sequences of sampled frames are used to train the GCNN to capture joint dependencies in the spatial domain. Then, the parameters in the GCNN are used to train the FDNet and update the selected frames for each video in the temporal domain, which are then used to refine the GCNN. The two models build off one another, as the GCNN provides rewards for the FDNet and the FDNet selects key frames for refining the GCNN. The better the GCNN is, the more accurate the provided rewards will be. The higher the quality of selected frames, the better the GCNN can be refined. At the test time, each video goes through the FDNet to produce its corresponding sequence with the informative frames, which will then be sent to the GCNN to provide the action label.