Short description

We use irCSN-152 as our backbone which was designed by many aspects such as large-scale,pre-training label space and the temporal dimension of vide.
Channel-separated convolutional networks (CSN) as 3D CNNs in which all convolutional layers (except for conv1) are either 1×1×1 conventional convolutions or k×k×k depthwise convolutions (where, typically, k = 3). Conventional convolutional networks model channel interactions and local interactions (i.e., spatial or spatiotemporal) jointly in their 3D convolutions. Instead, channel-separated networks decompose these two types of interactions into two distinct layers: 1×1×1 conventional convolutions for channel interaction (but no local interaction) and k×k×k depthwise convolutions for local spatiotemporal interactions (but not channel interaction). Channel separation may be applied to any k×k×k traditional convolution by decomposing it into a 1×1×1 convolution and a depthwise k×k×k convolution.