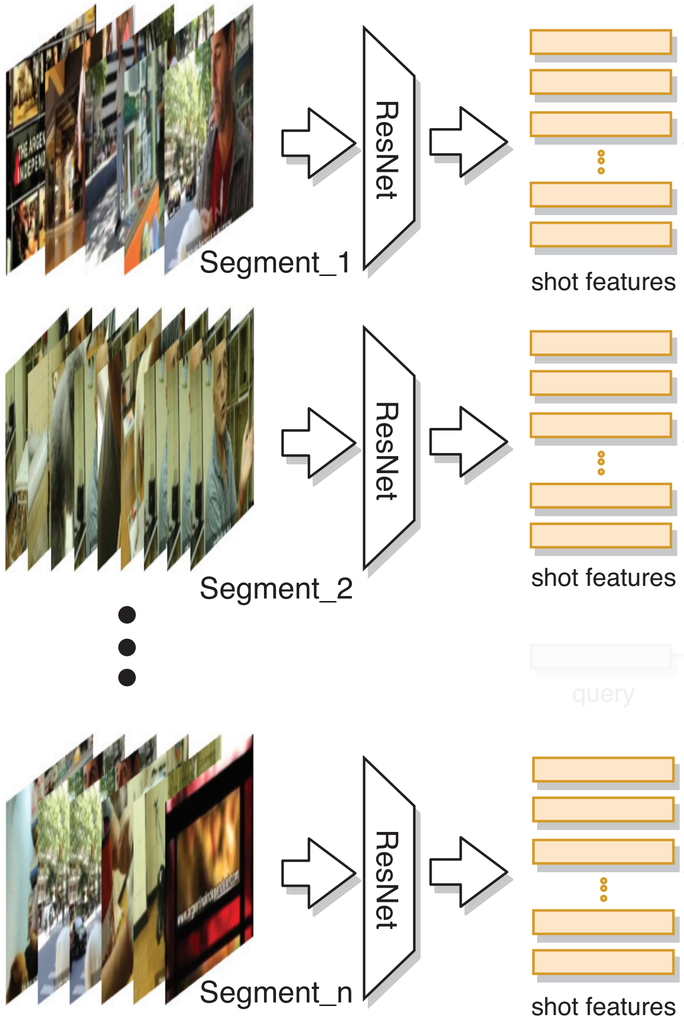
Proposed Method

Feature Encoding Network

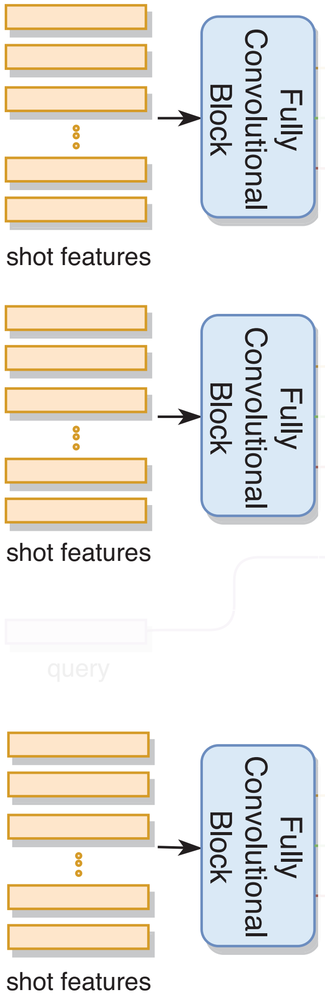


- Use the pertained ResNet to extract the visual feature of each shot after the segmentation process
- Denote the shot-level features in a specific video segment as $(v_1, v_2, \ldots, v_{s_k})$
  - s_k : the number of shots in the video segment
- Then use convolutional structure to further encode the visual feature

Proposed Method

Fully Convolutional Block

- Use 1D fully-convolutional network architecture to encode the shot-level visual feature
- Utilize **dilated convolutions** to obtain larger receptive field for handling long distance among the video segment



- First propose convolutional networks with different filter size and then concatenate their outputs which enables the model to receive more information
- The dilated convolution operation on i-th shot in a video segment: $o_i = \sum_{t=-k}^{k} f(t) \cdot v_{i+d \cdot t}$, where $2k+1$ is the filter size, f is the filter and d is dilation factor
- Then employ a pooling layer on the temporal axis of the video, can reduce computing time and also decrease the running memory of the model
- Connect different fully convolutional block and construct a multi-layer block to extracted representative features