

The diagram illustrates the architecture of a deep learning model for video classification, showing three parallel processing paths for different input types (video, image, and audio) and their fusion.

Top Path (Video): This path processes video input. It starts with a green block (input) and a blue block (feature map). The output is a sequence of blocks: a large blue block, a green block, a light blue block, a light blue block, a light blue block, a black block, a black block, a black block, a light green block, and a large green block (output).

Middle Path (Image): This path processes image input. It starts with a green block (input) and a blue block (feature map). The output is a sequence of blocks: a large blue block, a light green block, a light green block, a light green block, a light blue block, a light blue block, a light blue block, and a green block (output).

Bottom Path (Audio): This path processes audio input. It starts with a green block (input) and a blue block (feature map). The output is a sequence of blocks: a large green block, a green block, a green block, and a green block (output).

The three paths are connected by arrows, indicating the flow of information from the input blocks to the final output blocks. The top path is the most complex, involving multiple stages of processing and fusion. The middle and bottom paths are simpler, involving fewer stages of processing and fusion.

