1. Architecture

The architectures of ϕ, ψ, and ζ are provided in Figure 1.

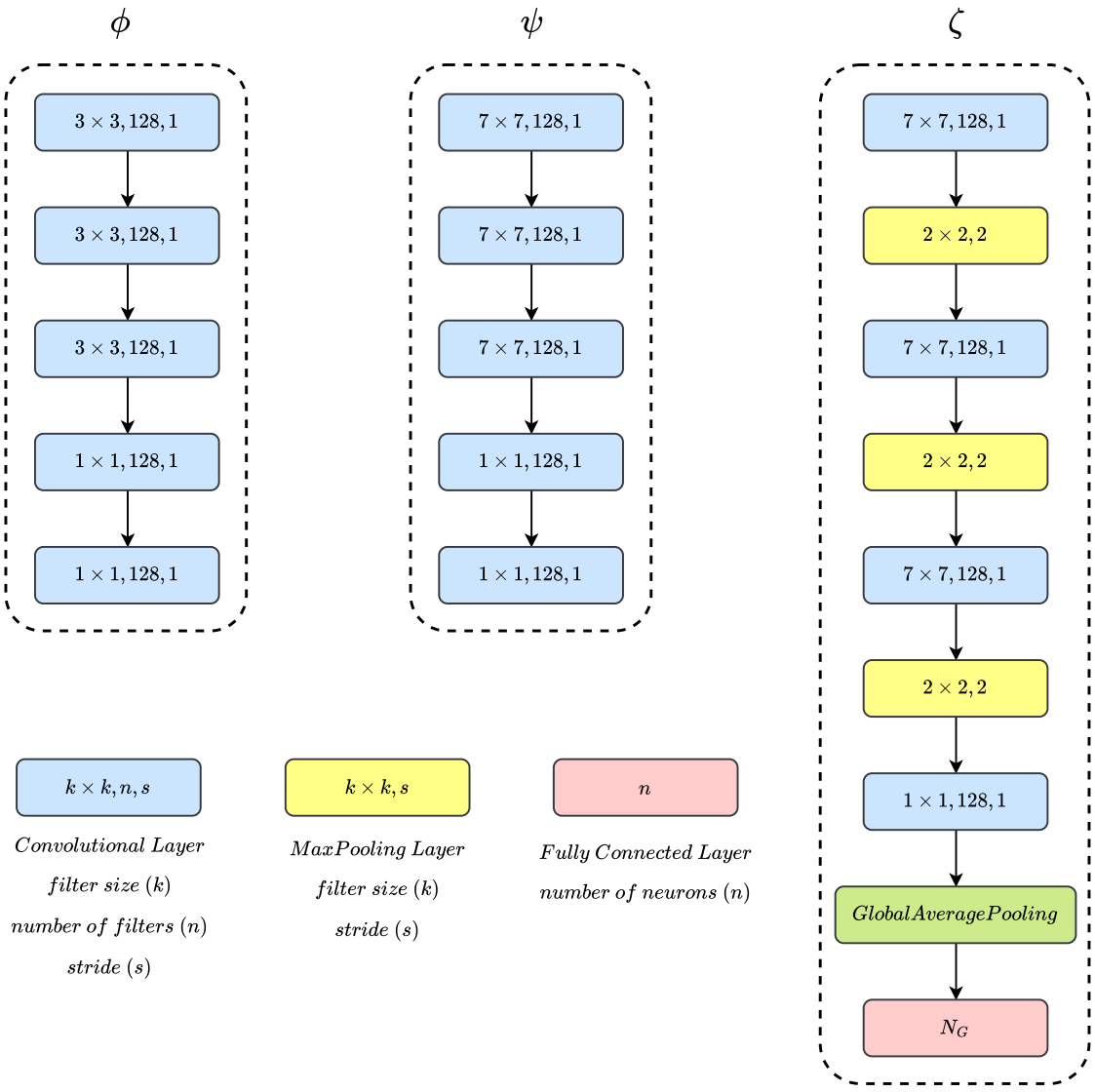

Figure 1. Details of the layers of ϕ, ψ, and ζ.

2. Inference Time

Inference times for the components of our model are given in Table 1. Depth of the input feature map for the refinement and aggregation components have direct impact on their inference times. In the InceptionV3 case the feature map of layer *Mixed_7c* has a depth of 2048. This depth is 832 for the layer *Mixed_4f* in I3D. This is why the refinement and aggregation components are slower when using Inception-V3. When using I3D on multiple frames, the inference time of the backbone is much higher and the inference time of the other parts of the model become negligible. Therefore, we get high improvements

Dataset	Backbone-CNN	Backbone	Refinement	Aggregation
Volleyball	Inception-V3	38	33	11
	I3D RGB/Flow	182/151	13/13	5/4
Collective	I3D RGB/Flow	19/57	6/7	3/3

Table 1. Inference time (milliseconds) for various components of our model in different settings on a GTX 1080 Ti. The shapes of inputs for Volleyball, Collective (RGB), and Collective (Flow), are 720×1280, 240×360, and 480×720, respectively. Feature maps in Volleyball and Collective are resized to 43×78, and 30×45, respectively.

at a relatively low cost.