

*Input*

Camera Images

Convolutional  
Feature Extractor

State Vector

Action

$\oplus$

LSTM

Fully Connected  
Network

Output Features

*Shared*

Position  
(3D)

Velocity  
(3D)

Orientation  
(Quater-  
nion 4D)

*Output Heads*

