Picking Prediction module Q_T1 Affordance map Tool_id Argmax **Pick position Q_T2** Affordance map **Q_T3** Affordance map

Conv2D RGBD Img Orthographic Project to 360*720*6 1920*1080*4 **General Visual** UnSampling 2D **Feature** Resnet Conv_Block (B,256,90,90) **Backbone** To match input (360,720,C) Crop img to a small pitch centered at Pick position Resnet Resnet Backbone **Backbone Rotate output Placing Visual** features **Feature** (360,720) Feature Feature with Ang 0 with Ang 10 with Ang 20 Query Place position Cross- Argmax Key₩ &orientation correlation **Placing Prediction module**

Tool Heads

Kernal 2

Learnable Learnable

Kernal 3

Learnable