Aggregate 2D features to refine object queries

Predict a reference point using a sub-network

Transform this 3D reference point into image space

Use the transformed point to index image features

$c_{\ell m i}$ ← $c_{\ell i}$

Multi-view Images with
Camera Extrinsics &
intrinsics

ResNet+FPN

Image Feature
Extraction

Object
Queries

2D-to-3D Feature
Transformation

$Pred_1$

$Pred_2$

$Pred_i$

$Pred_{M*}$

None

Box_1

Box_2

Box_j

Box_M

Bipartite Matching
& Loss Computation