# Appendix

## A Pseudo Code of the Proposed ACE-BNV Policy

We summarize the overall NBV planning policy for grasping target object in the following algorithm 1.

**Algorithm 1** Grasp Affordance Prediction and Next-Best-View Planning

**Input:** A cluttered scene, a occluded target object given its 3D bounding box $\mathbf{T}_{\text{bbox}}$
**Output:** A feasible grasp of target object $\mathbf{G}$

```
for t ≤ T_max do
    M_t ← D_t                                          ▷ Intergrate depth image into TSDF
    C_t ← 3D CNN(M_t)                                  ▷ Encode feature
    if G*_q(T_bbox, C_t, O_v,t) ≤ q_max then
        O_v,t+1 ← arg max_{G_v ∈ 𝔾_v} G*_q(T_bbox, G_v)   ▷ Evaluate candidate next views
        Move camera to O_v,t+1                         ▷ Go to the next-best-view
    else
        Execute grasp G*_q(T_bbox, C_t, O_v,t)
        Break                                          ▷ Execute grasp
    end if
end for
```

## B Network Architecture and Implementation Details

We adopt the same encoder with GIGA that takes TSDF $\mathbf{M}_t \in \mathbb{R}^{40\times40\times40}$ as input and output a feature embedding for each voxel with a 3D CNN layer. Then, the tri-plane feature grids is constructed by projecting each input voxel on a canonical feature plane via orthographic projection. Then, three feature planes are processed with a 2D U-Net that consist of a series of down-sampling and up-sampling 2D convolution layers with skip connections. The output is formulated as the shared tri-plane feature volume $\mathbf{C} \in \mathbb{R}^{3\times40\times40\times32}$, where 32 is the dimension of the feature embedding.

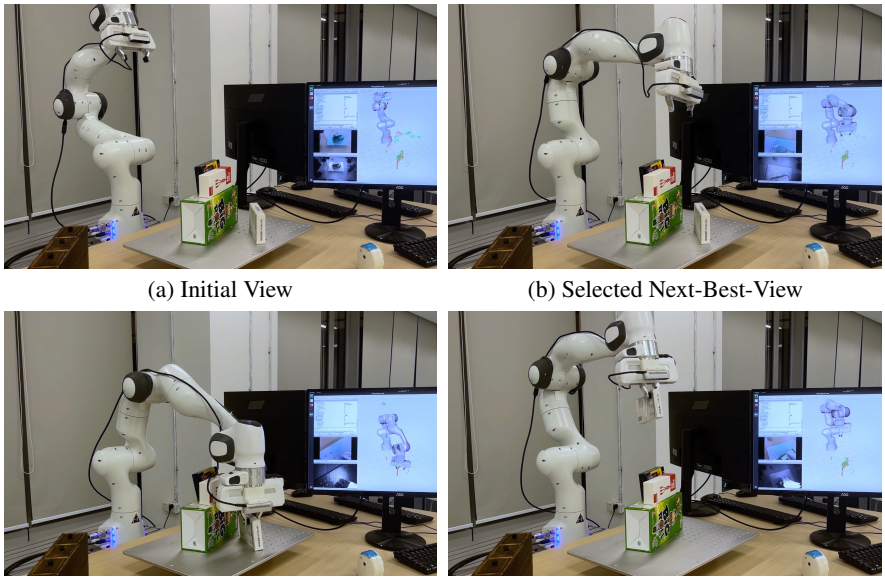
Based on the shared tri-plane feature volume, the local feature $\mathbf{C}_\mathbf{p}$ of a 3D point $\mathbf{p} = (x, y, z)$ is obtained by projecting it to each feature plane and query three features $\mathbf{C}_{\mathbf{p}_x}, \mathbf{C}_{\mathbf{p}_y}, \mathbf{C}_{\mathbf{p}_z}$ at the projected locations using bilinear interpolation, and the local feature $\mathbf{C}_\mathbf{p}$ is the concatenated feature of these queried features, i.e., $\mathbf{C}_\mathbf{p} = \text{concat}(\mathbf{C}_{\mathbf{p}_x}, \mathbf{C}_{\mathbf{p}_y}, \mathbf{C}_{\mathbf{p}_z})$. We implement our grasp affordance prediction network with a five layer fully-connected network with residual connections. The input dimension of this MLP network is $3 + 96 + 9 \times 96 = 963$ which is composed of view direction unit vector $\mathbf{v} \in \mathbb{R}^3$, the ray feature $\mathbf{C}_{\text{ray}} \in \mathbb{R}^{96}$, and the local geometry feature $\mathbf{C}_{\text{geo}} \in \mathbb{R}^{9\times96}$. The output dimension for grasp affordance prediction is $1 + 1 + 1 = 3$ that consist of grasp quality $\mathbf{G}_q$, in-plane rotation $\mathbf{G}_r$, and gripper width $\mathbf{G}_w$.

As for the novel view depth synthesis, we employ a MLP network that takes the 3D point feature $\mathbf{C}_\mathbf{p} \in \mathbb{R}^96$ as input and output the SDF value of this point, and adpot the same rendering technique with NeuS to synthesize depth images($\eta = 12, \gamma = 5$). We sample 128 rays in depth image in each batch, 64 points on a ray, 4 times importance sampling with 32 points each time. We set the near and far range close to the ground truth depth at the beginning of training, and then gradually relax the range to the maximum range of the implicit feature volume.

For experiments in simulation and real word, the size of cubic workspace $L = 30$cm. The size of the cubic for $\mathbf{C}_\mathbf{vert}$ is $0.25$, which is $7.5$cm corresponding to the real world. The points $\mathbf{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$ for $\mathbf{C}_\mathbf{ray}$ is uniformly sampled with a step of $0.1$. The sizes of the three datasets front-observe-front-grasp, front-observe-side-grasp and multi-observe-front-grasp are 1M, 1M and 2M grasps, respectively. Each scene contains 240 grasps and $\eta = 12$ ground-truth depth images with the resolution of 480×640. After data cleaning and balancing, there are about $40\%$ data left. We separate the datasets randomly into 90% training and 10% validation. We train the models with the Adam optimizer and a learning rate of $2 \times 10^{-4}$ and batch sizes of 128.

## C Qualitative Results of Real Robot Experiments

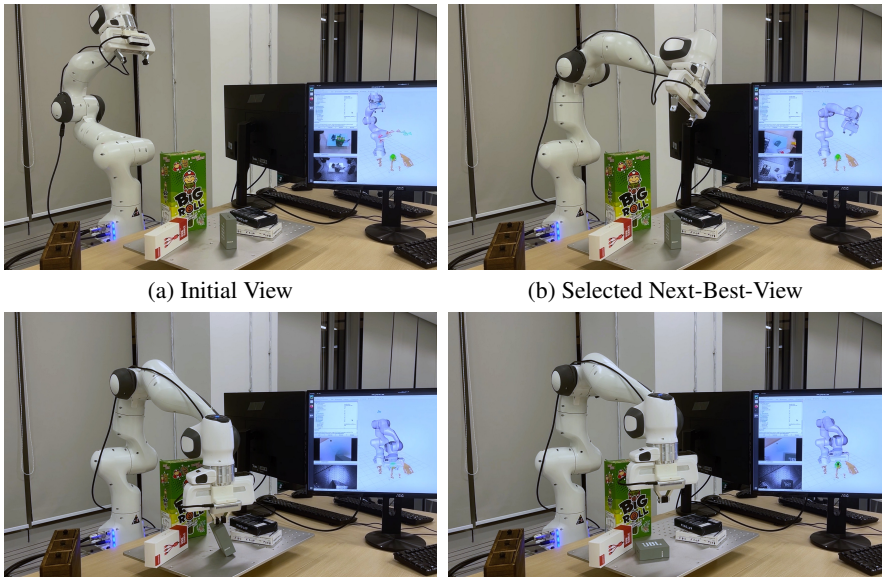
We present qualitative results in Fig. 1 and 2 and recommend readers watch the supplementary video for more comprehensive real robot experimental results. Note that our model can select reasonable next-best-view to observe the occluded target object. We show a representative failure case in Fig. 2, where small errors in grasp affordance prediction leads to a unsuccessful grasp. This small prediction inaccuracy occurs in most failure experiments. Therefore, in the future, we plan to exploit a better grasp affordance prediction module to improve the success rate of our method.



(a) Initial View (b) Selected Next-Best-View

(c) Execute Grasp (d) Grasp Success

Figure 1: Success Case. The robot planned one new view to observe the target box and successfully grasped it.



(a) Initial View (b) Selected Next-Best-View

(c) Execute Grasp (d) Grasp Fail

Figure 2: Failure Case. The robot failed to predict accurate grasp affordances of the target object after obtain a new observation. As a result, the grasping is failed.