Review:

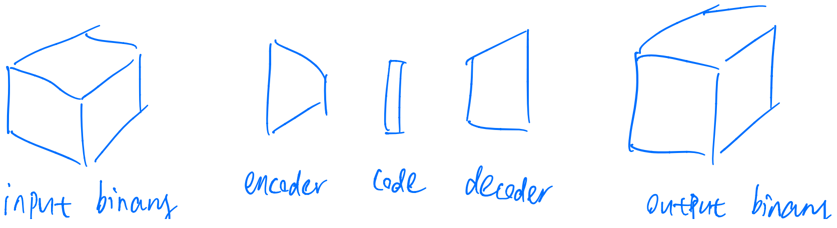
Most prior work focus on a single object seen from a limited number of views, and do not evaluate the confidence in the generated model using more traditional measurement residuals, e.g. point to surface errors.

[10] embeds modern higher order object shape priors into classical iterative residual minimization objectives.

- simplest approach of including higher-level knowledge is by explicitly representing the 1D or 2D geometric structure in man-made scene.
- We use a compact way to describe the object's shape given by the n-d shape descriptor λ, and a function to map from the latent shape space to full 3D geometry.

Let $P(\lambda)$ be the set of points describing the object's shape, the objective is:

$$\{T_{i,opt}, \lambda_{opt}\} = \underset{T_i, \lambda}{\operatorname{argmin}} \sum_i \sum_j 1_{ij} \eta (m_{ij} - \pi (T_i P(\lambda) [j])), \quad (4)$$



voxel grid   voxel grid!

1. apply object segmentation in each RGBD frame. Take 3D points measured on the object's surface and transform to world frame.

2. We use occlusion mask M, which indicates which points have not been observed by any previous measurements.

Cost function is:

$$L_{mapping}(\lambda) = \sum_i \sum_j \sum_k \{(1-M_{ijk}) f(G_{ijk}(\lambda), F_{ijk}) + \alpha M_{ijk} f(G_{ijk}(\lambda), G_{ijk}(\lambda_0))\}$$

$f(\cdot)$ is binary cross entropy function.

α is trade-off factor, governs the overall amount we enforce the prior.

(Back to DeepSLAM paper). However, it is difficult to balance the influence of the network priors against the traditional measurement residuals,

Our main insight: it is better to install a greedy search strategy in which we generate many 3D model predictions from many views, and use measurement fidelity to simply perform a discrete selection.

Notations.

For each frame i, T_{wi} is transformation from camera to world frame.

O is the set of all models, transformation from a camera to a model is T_{oi}.

Pose of each object is T_{wo}.

Observing frames of model $k \in O$ are recorded in L_k.

Candidate models generated from each RGB frame in L_k are C_k.

Our solution to produce more confident prediction: generate multiple proposals C_k from distinct views L_k, and performing a discrete selection by consulting the agreement with an accumulation of the actual depth measurements of model k.

Graph optimization:

For each edge, we then have a measurement obtained from the point-to-plane based ICP.

Parameterizing the ground plane lets us further constrain the orientation of upright objects to align with vertical residual.