Deep Single-View 3D Object Reconstruction with Visual Hull Embedding

Hanqing Wang[1,2], Jiaolong Yang[2], Wei Liang[1], Xin Tong[2]

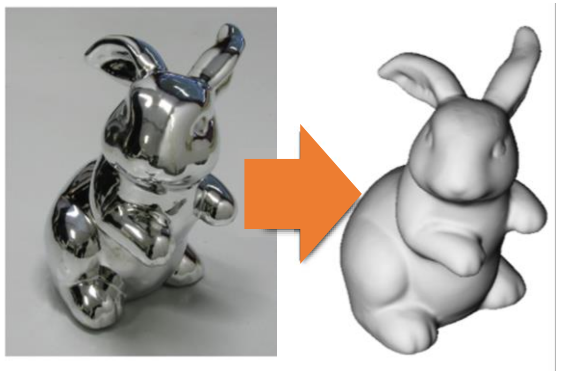
[1] Beijing Institute of Technology [2] Microsoft Research Asia

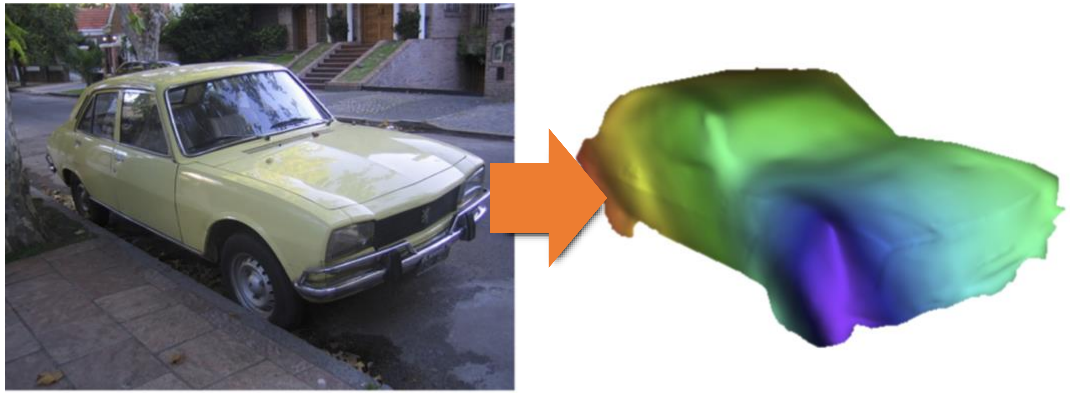
Beijing, China Beijing, China

AAAI 2019

Single-View 3D Reconstruction

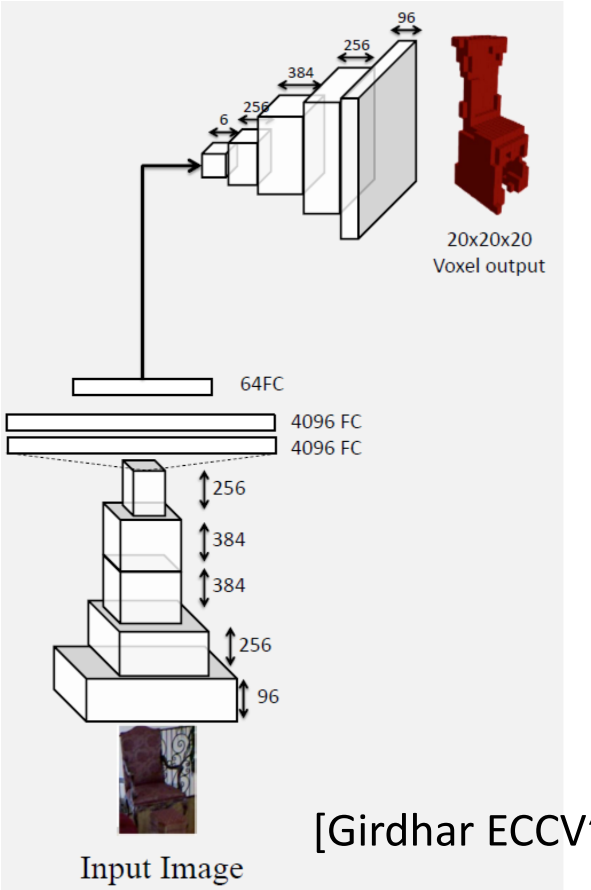
- Input: a single RGB(D) Image
- Output: the corresponding 3D representation



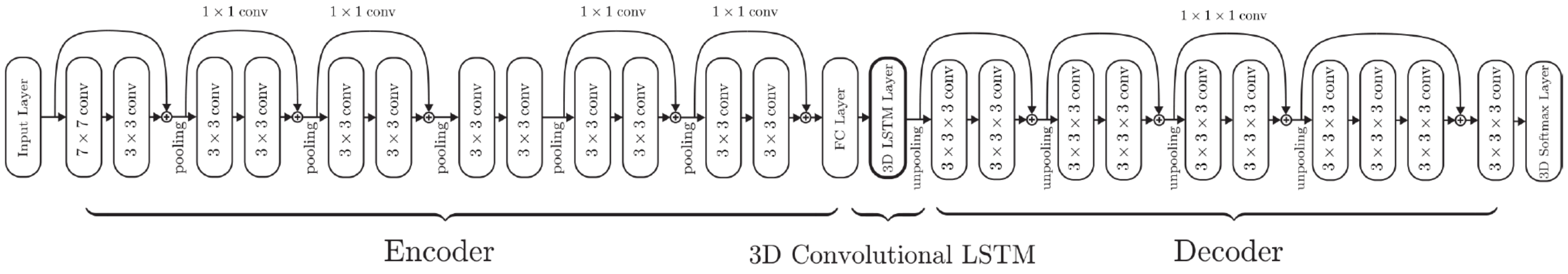


Previous Works

- Deep Learning based Methods:



[Girdhar ECCV'16]

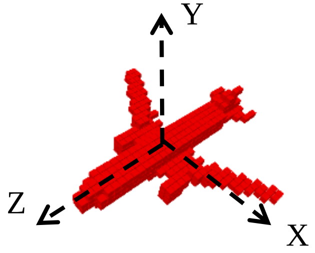


[Choy ECCV'16]

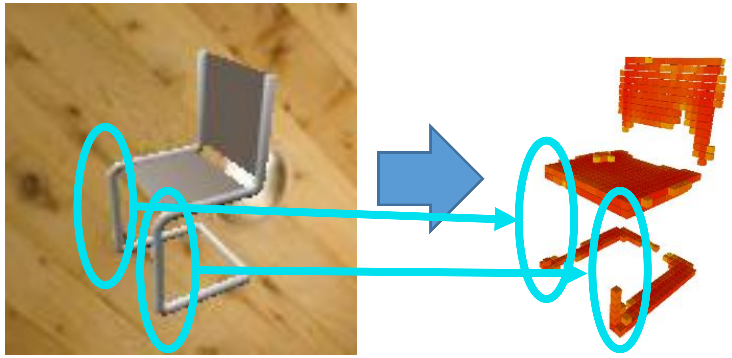
Other works:
[Yan NIPS'16][Wu NIPS'16][Tulsiani CVPR'17][Zhu ICCV'17]

Limitations of previous works

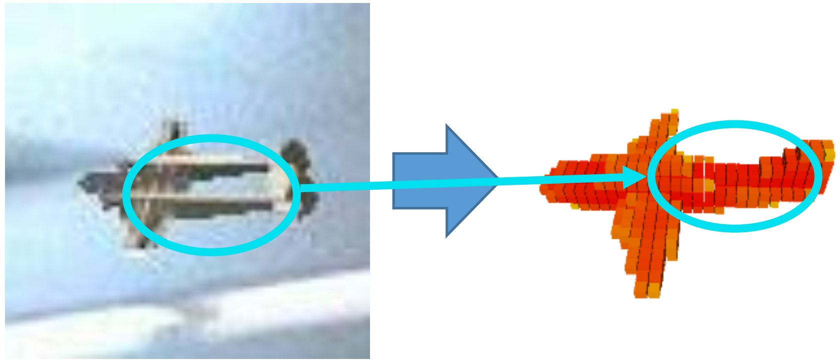
- Problems of Existing Deep Learning based Methods:
  - 1. *Arbitrary-view* images vs. *Canonical-view* aligned 3D shapes



  - 2. Unsatisfactory results



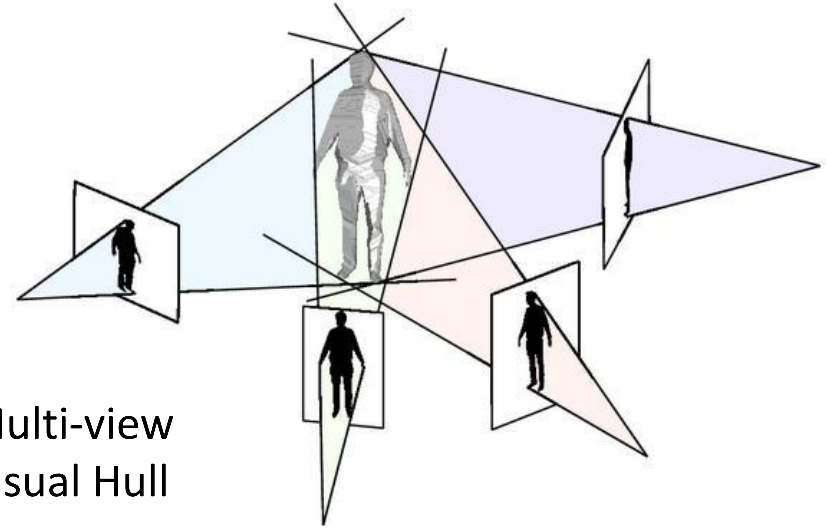
Missing shape details



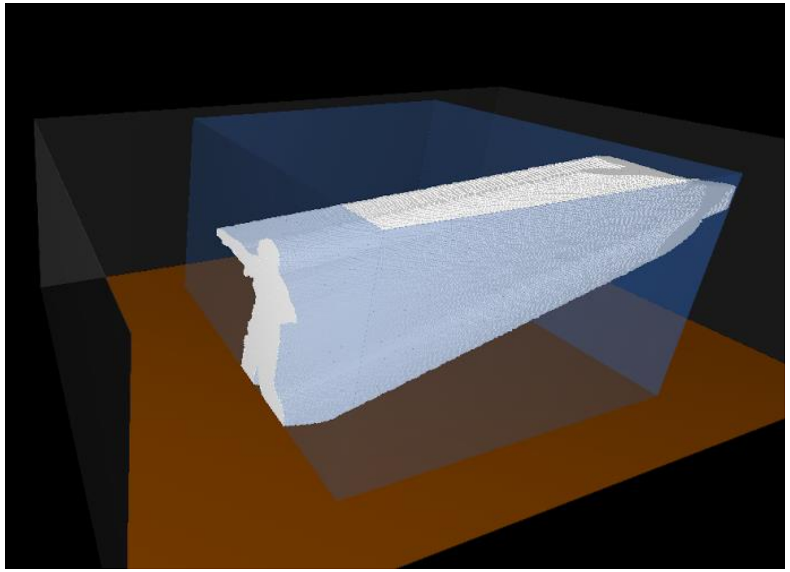
Inconsistency with input

Core Idea

- Goal: **Reconstruct** the object **precisely** with the given image
- Idea: Embed explicitly the **3D-2D projection geometry** into a network
- Approach: Estimating a **single-view visual hull** inside of the network

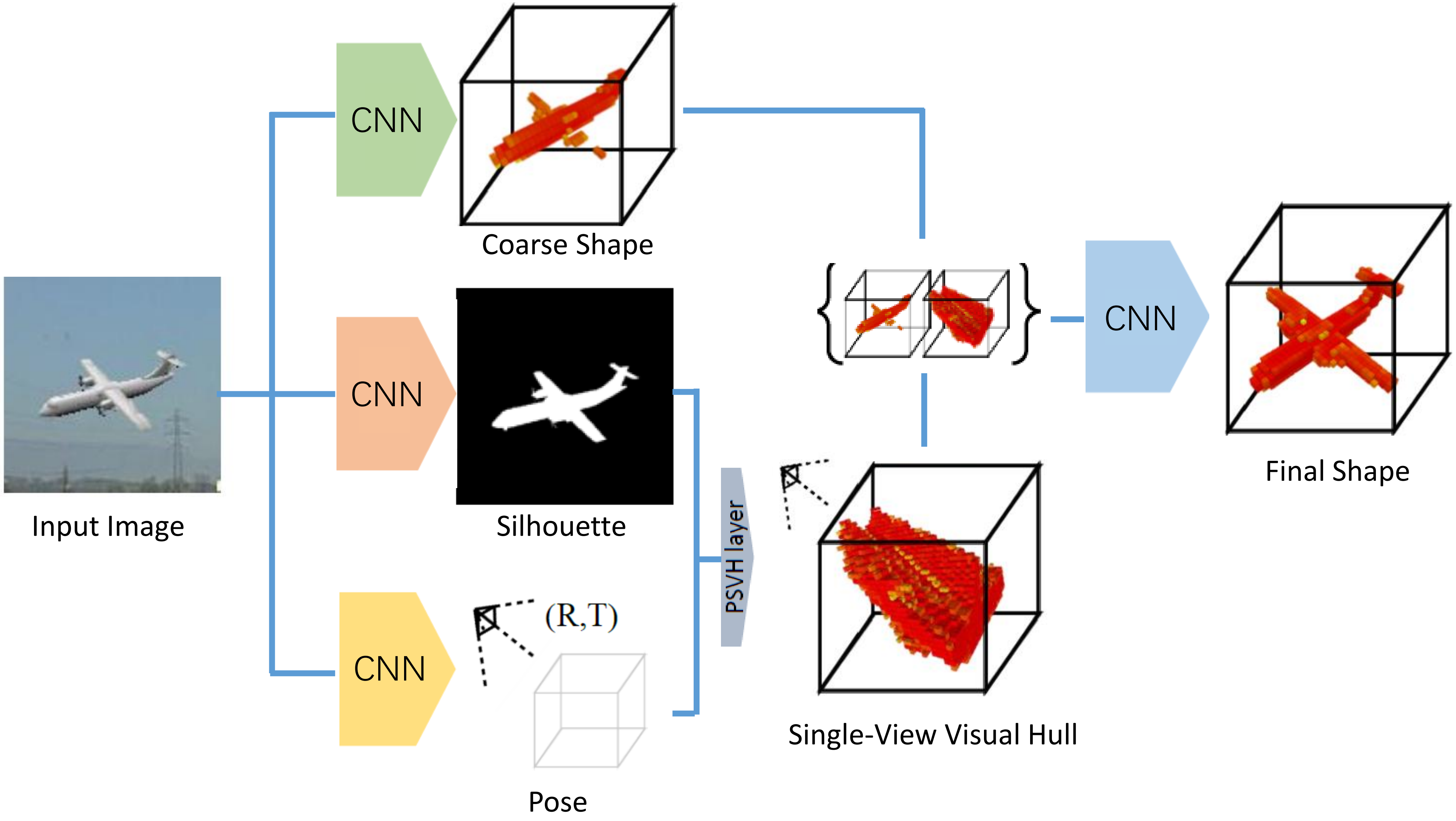


Multi-view
Visual Hull

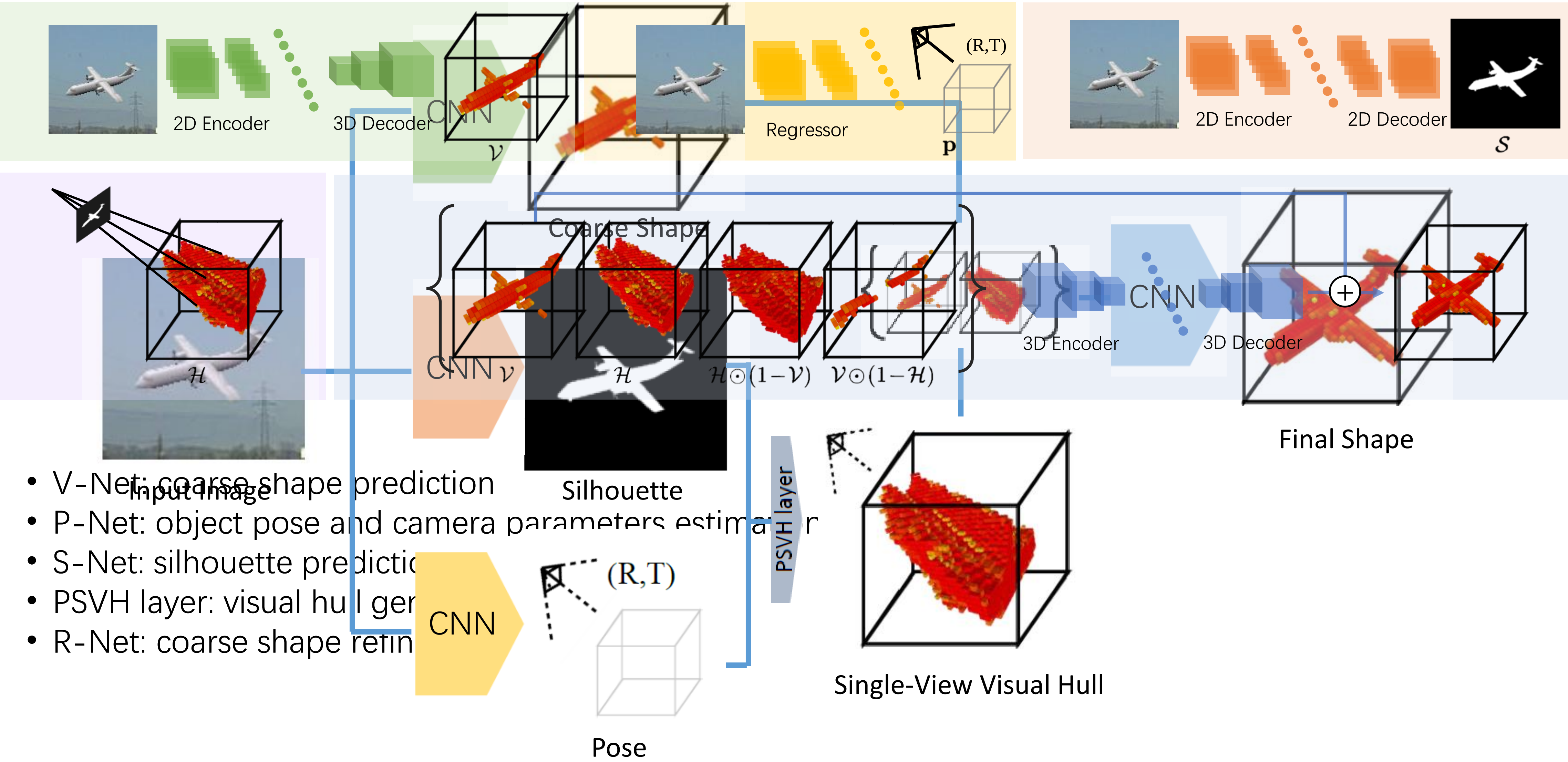


Single-view
Visual Hull

Method Overview

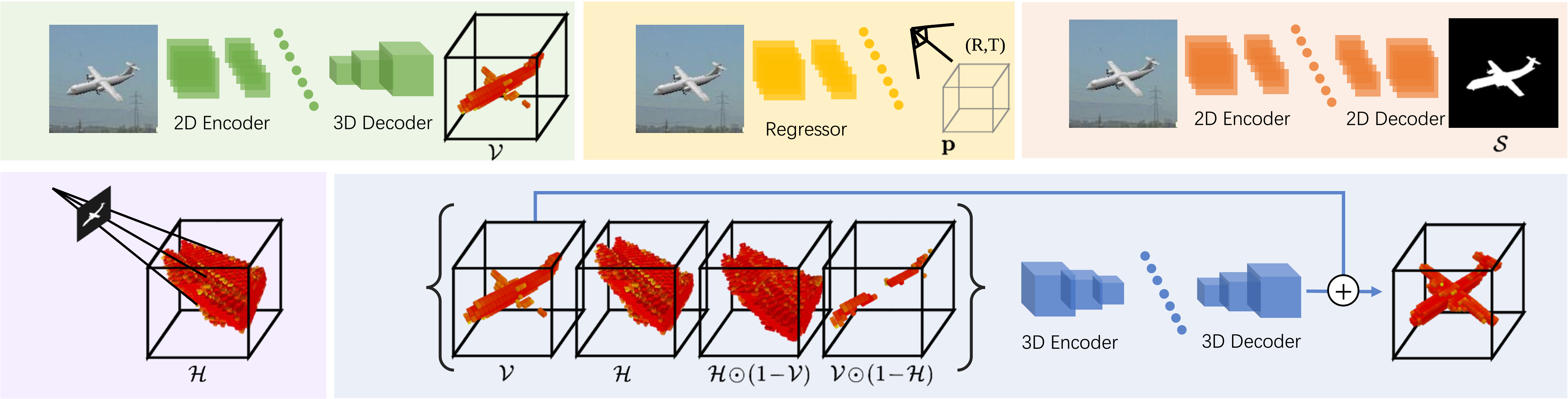


Components



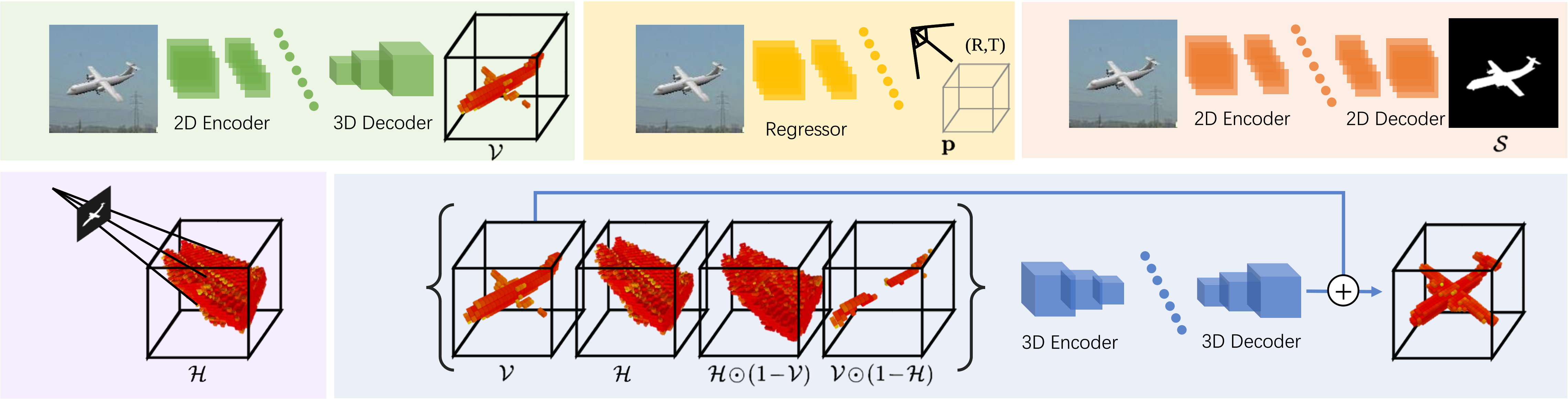
- V-Net: coarse shape prediction
- P-Net: object pose and camera parameters estimat r
- S-Net: silhouette predicti
- PSVH layer: visual hu l ger
- R-Net: coarse shape refin

Components



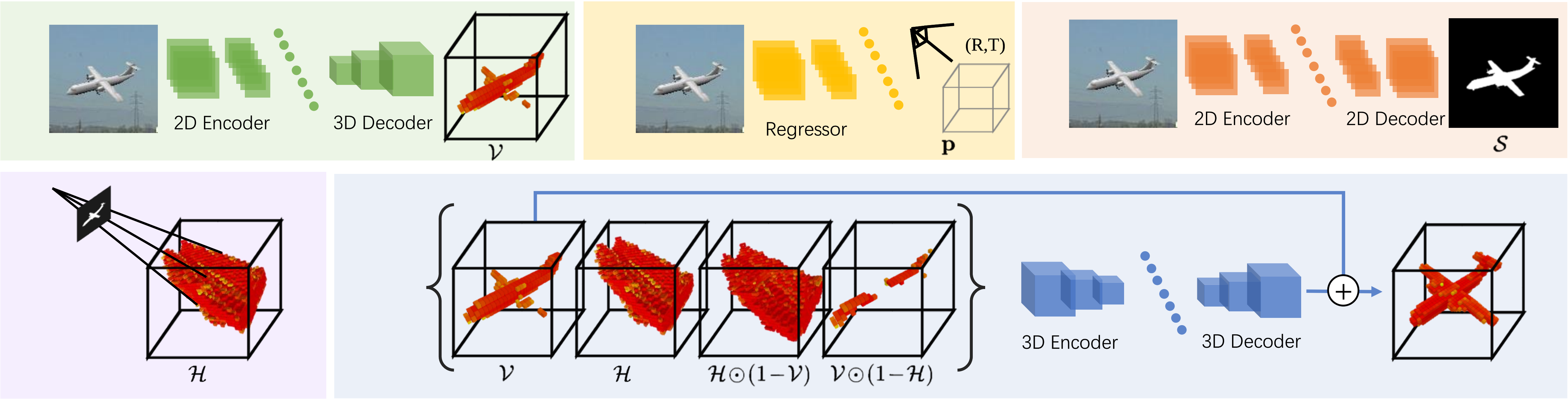
- V-Net: coarse shape prediction
- P-Net: object pose and camera parameters estimation
- S-Net: silhouette prediction
- PSVH layer: visual hull generation
- R-Net: coarse shape refinement

Components



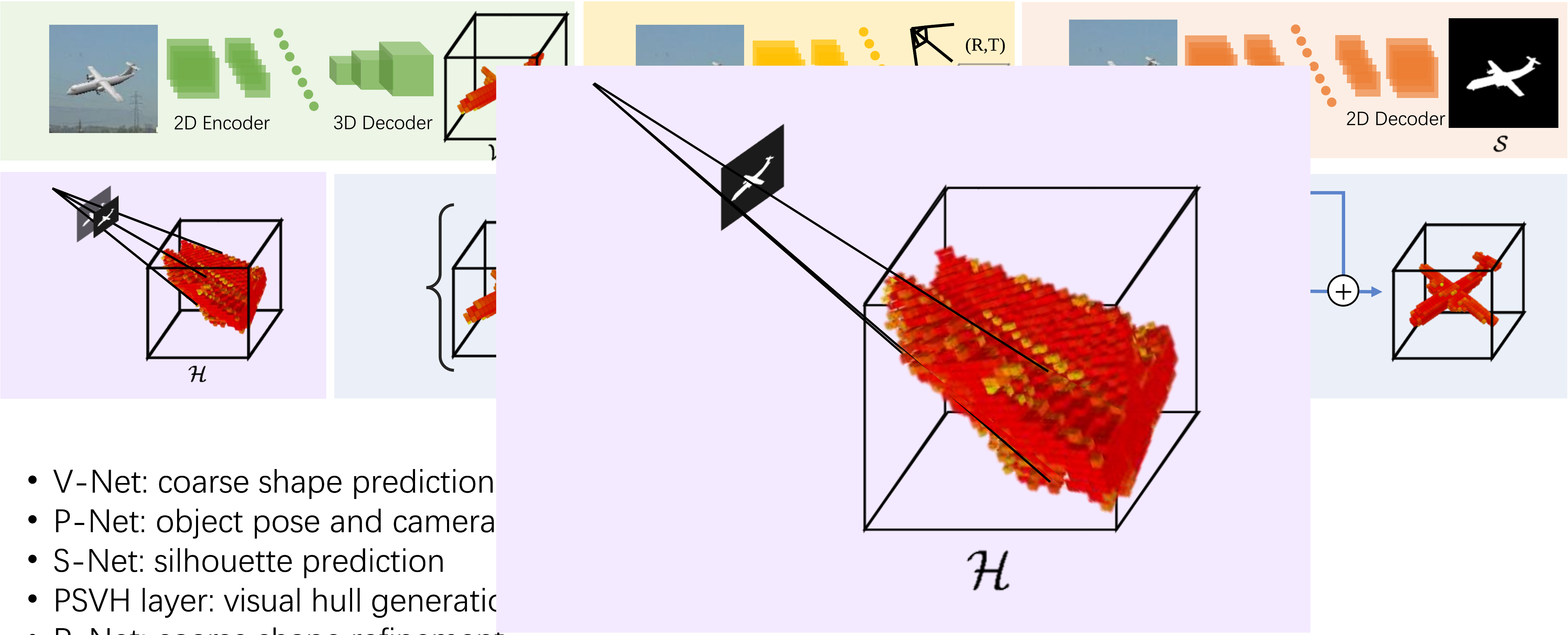
- V-Net: coarse shape prediction
- P-Net: object pose and camera parameters estimation
- S-Net: silhouette prediction
- PSVH layer: visual hull generation
- R-Net: coarse shape refinement

Components



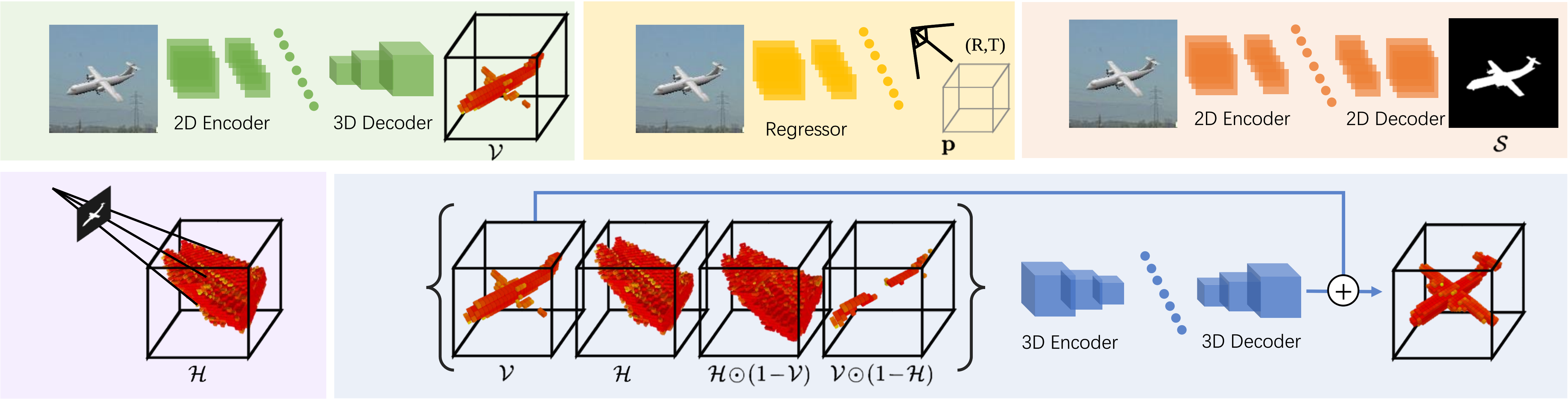
- V-Net: coarse shape prediction
- P-Net: object pose and camera parameters estimation
- S-Net: silhouette prediction
- PSVH layer: visual hull generation
- R-Net: coarse shape refinement

Components



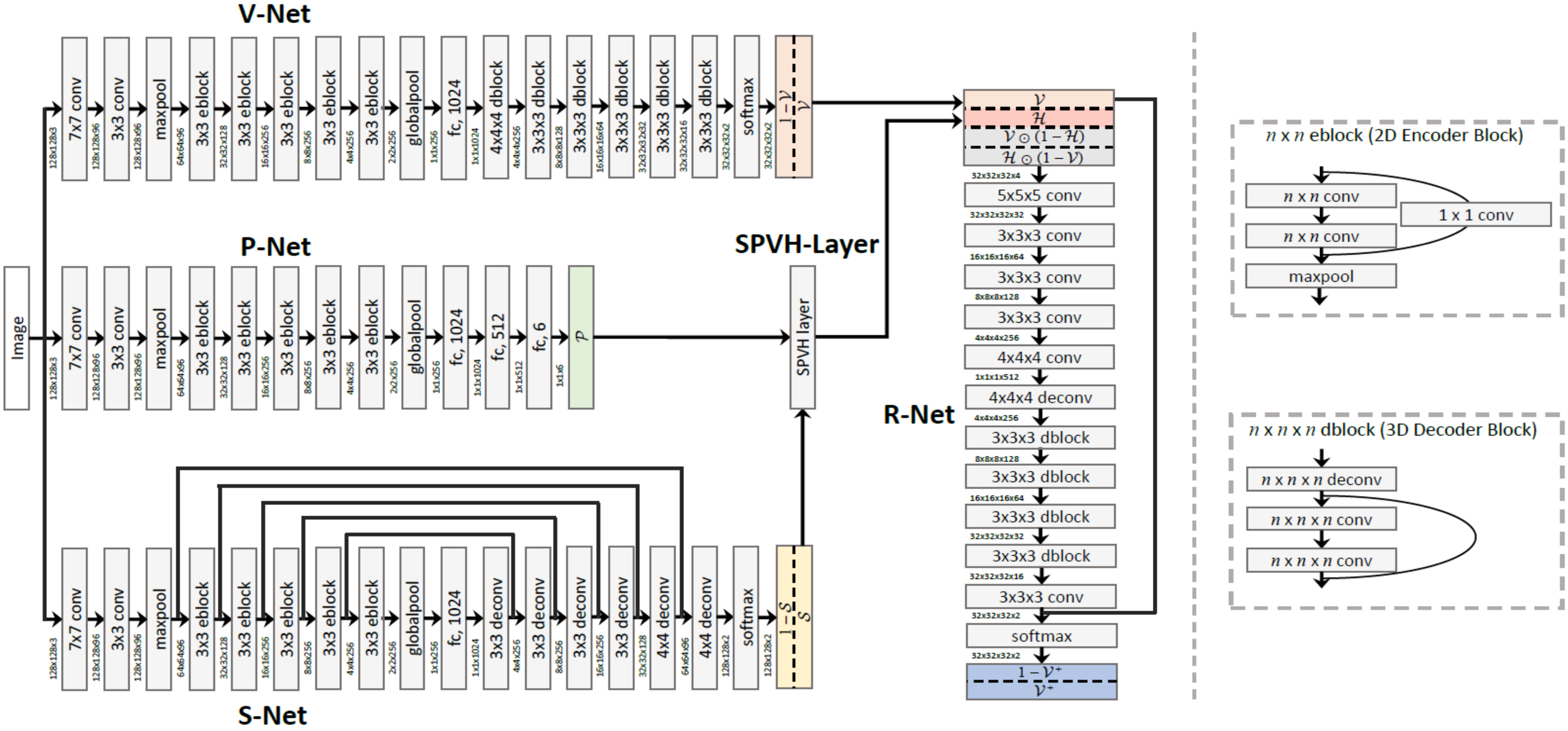
- V-Net: coarse shape prediction
- P-Net: object pose and camera
- S-Net: silhouette prediction
- PSVH layer: visual hull generatio
- R-Net: coarse shape refinement

Components



- V-Net: coarse shape prediction
- P-Net: object pose and camera parameters estimation
- S-Net: silhouette prediction
- PSVH layer: visual hull generation
- R-Net: coarse shape refinement

Network Architecture



Training Details

Loss:

We use the binary cross-entropy loss to train *V-Net*, *S-Net* and *R-Net*, let p_n be the estimated probability at location n, the loss is defined as

$$l = -\frac{1}{N}\sum_{n}(p_n^* \log p_n + (1 - p_n^*)\log(1 - p_n)) \quad \textbf{(2)}$$

Where p_n^* is the target probability

For *P-Net,* we use the L_1 regression loss to train the network:

$$l = \sum_{i=1,2,3} \alpha|\theta_i - \theta_i^*| + \sum_{j=u,v} \beta\left|t_j - t_j^*\right| + \gamma|t_Z - t_Z^*| \quad \textbf{(3)}$$

where we set $\alpha = 1, \gamma = 1, \beta = 0.01$

Training Details

Steps:

1. Train the V-Net, S-Net, P-Net independently.
2. Train the R-Net with the coarse shape predicted by V-Net and the ground truth visual hull.
3. Train the whole network end-to-end.

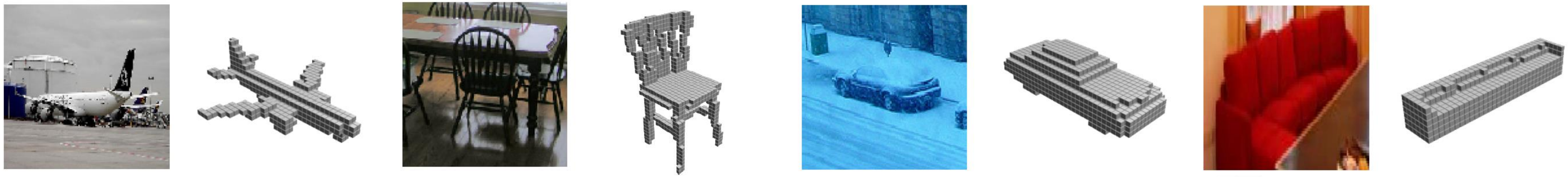
Implementation Details

- Network implemented in Tensorflow
- Input image size: 128x128x3
- Output voxel grid size: 32x32x32

Dataset

- **Object categories**: *car, airplane, chair, sofa*
- **Datasets**:
  - Rendered ShapeNet objects – (ShapeNet) dataset of tremendous CAD models

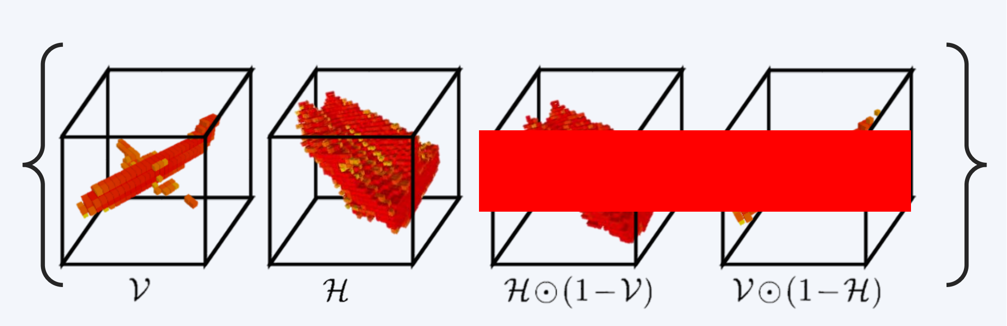
  - Real images - (PASCAL 3D+ dataset) manually associated with limited CAD models



Experiments

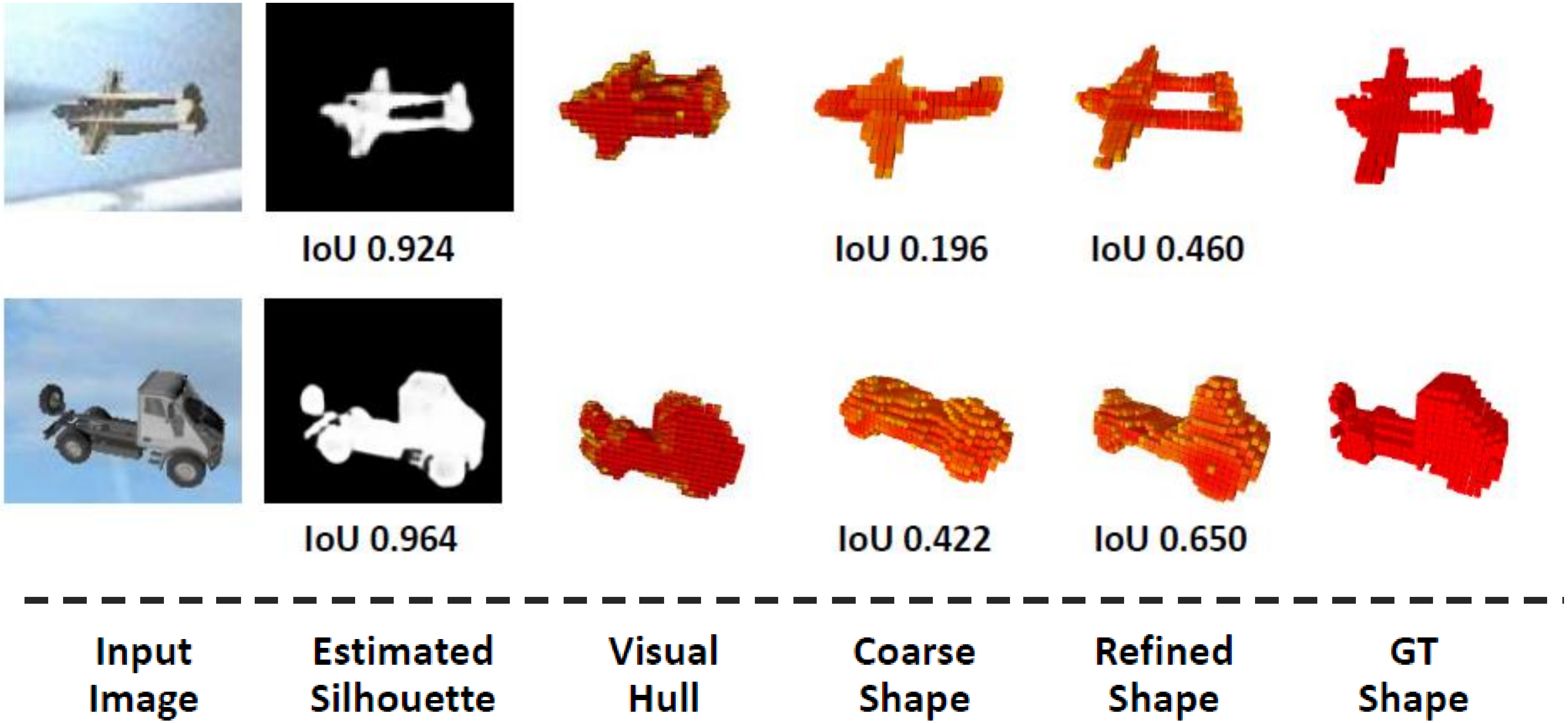
- Results on the 3D-R2N2 dataset (rendered ShapeNet objects)
  - **Ablation study:**

	car	airplane	chair	couch	Mean
Before Refine.	0.819	0.537	0.499	0.667	0.631
After Refine.	**0.839**	**0.631**	**0.552**	0.698	**0.680**
Refine. w/o $\mathcal{H}$	0.824	0.541	0.505	0.675	0.636
Refine. w. GT $\mathcal{H}$	0.869	0.701	0.592	0.741	0.726
Refine. w/o 2 prob.maps	0.840	0.610	0.549	0.701	0.675
Refine. w/o end-to-end	0.822	0.593	0.542	0.677	0.658



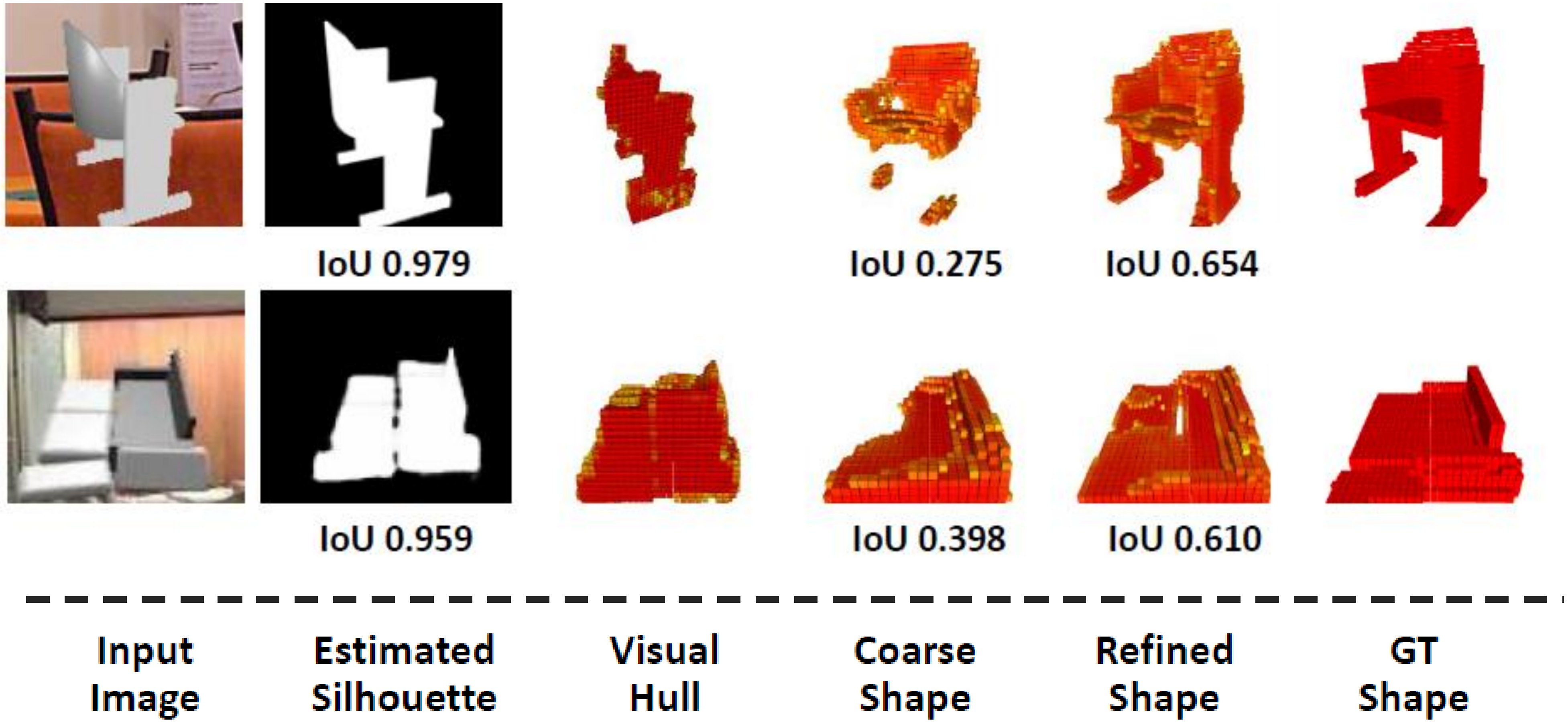
Experiments

- Results on the rendered ShapeNet objects



Experiments

- Results on the rendered ShapeNet objects



Experiments

- Results on the synthetic dataset (rendered ShapeNet objects)
  - **Ablation study:**

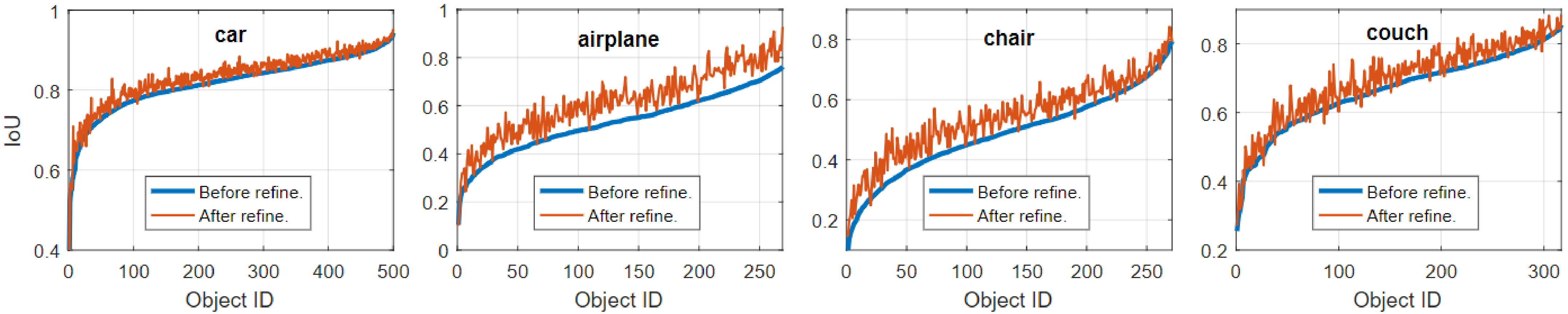
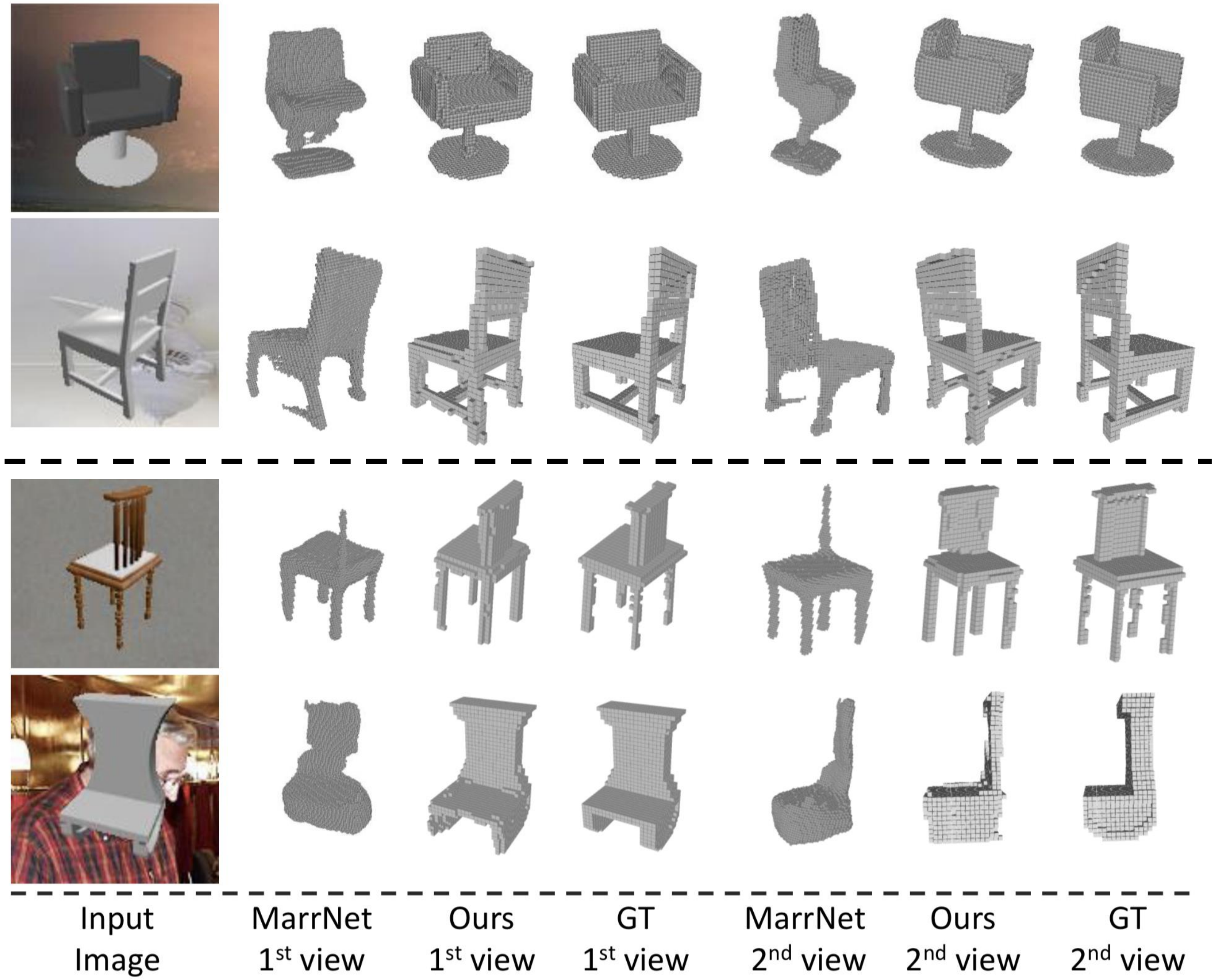


Figure 5: Comparison of the results before and after refinement on rendered ShapeNet objects.

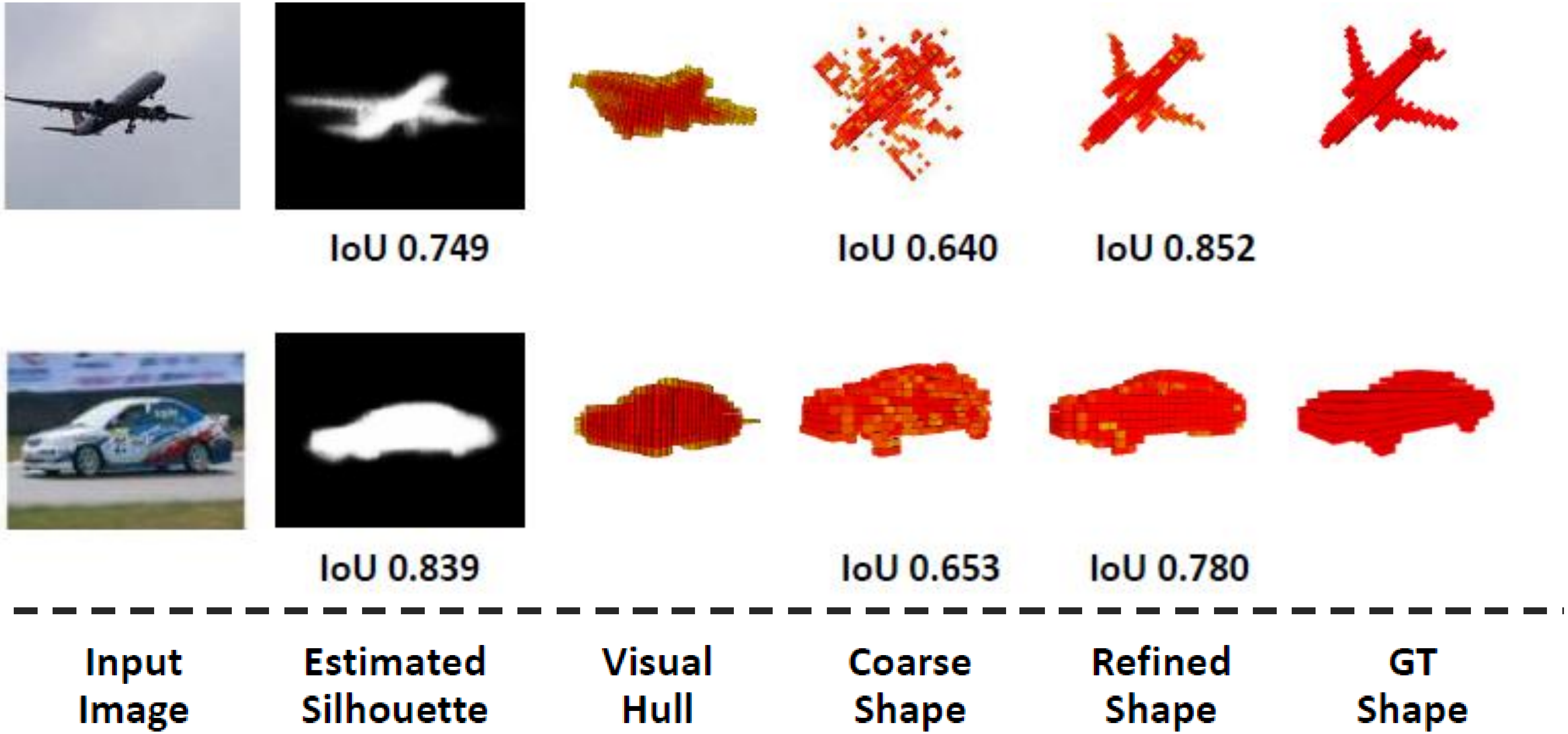
Experiments

- Comparison with MarrNet[Wu et al. 2017] on the synthetic dataset



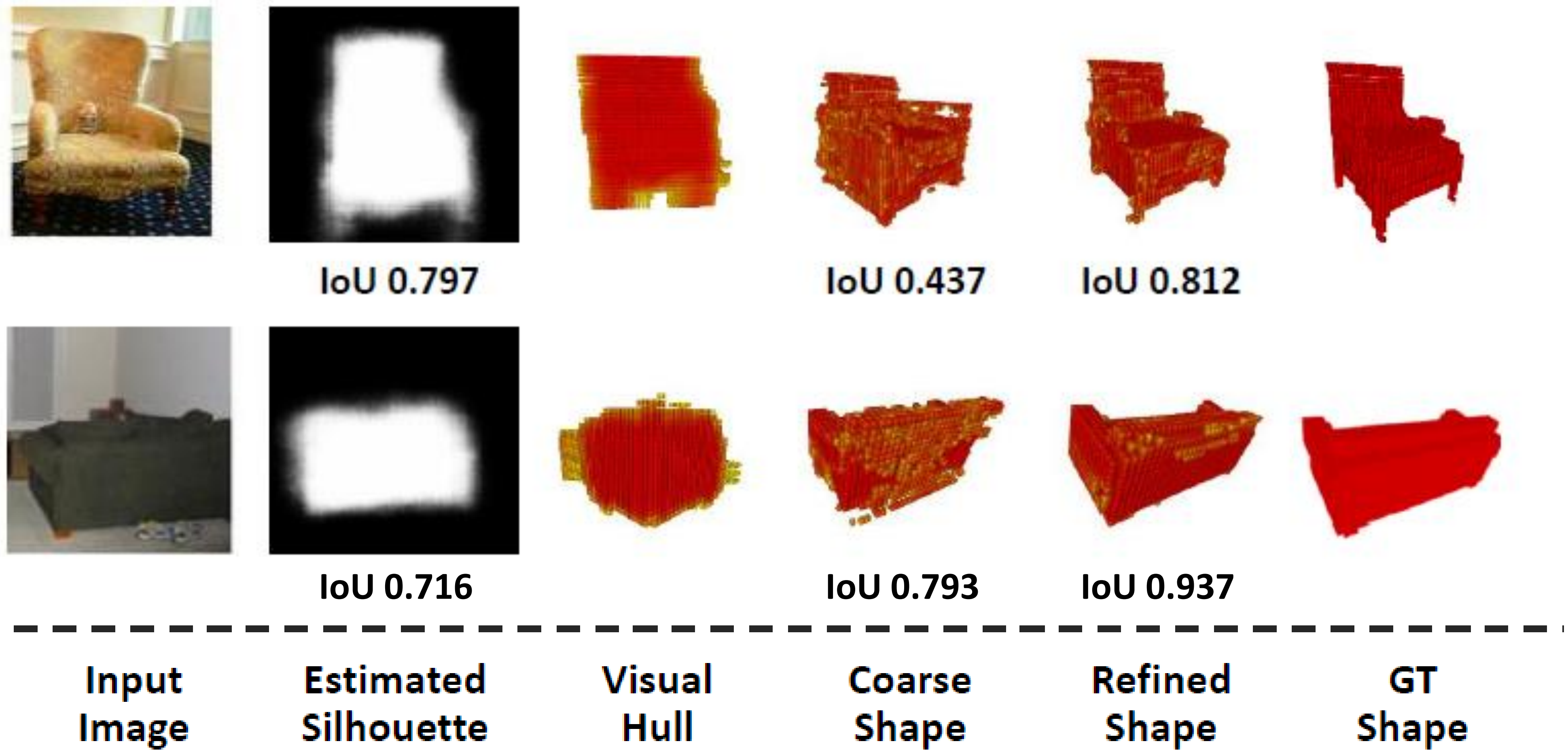
Experiments

- Results on the PASCAL 3D+ dataset (real images)



Experiments

- Results on the PASCAL 3D+ dataset (real images)



Running Time

- ~18ms for one image (**55 fps!**)
- (Tested with a batch of 24 images on a NVIDIA Tesla M40 GPU)

Contributions

- Embedding **Domain knowledge** (3D-2D perspective geometry) into a DNN
- Performing reconstruction jointly with **segmentation** and **pose estimation**
- A novel, **GPU-friendly PSVH** (Probabilistic Single-view Visual Hull) layer

Thanks for listening!

- Welcome to ask any problem!
- Email: hanqingwang@bit.edu.cn