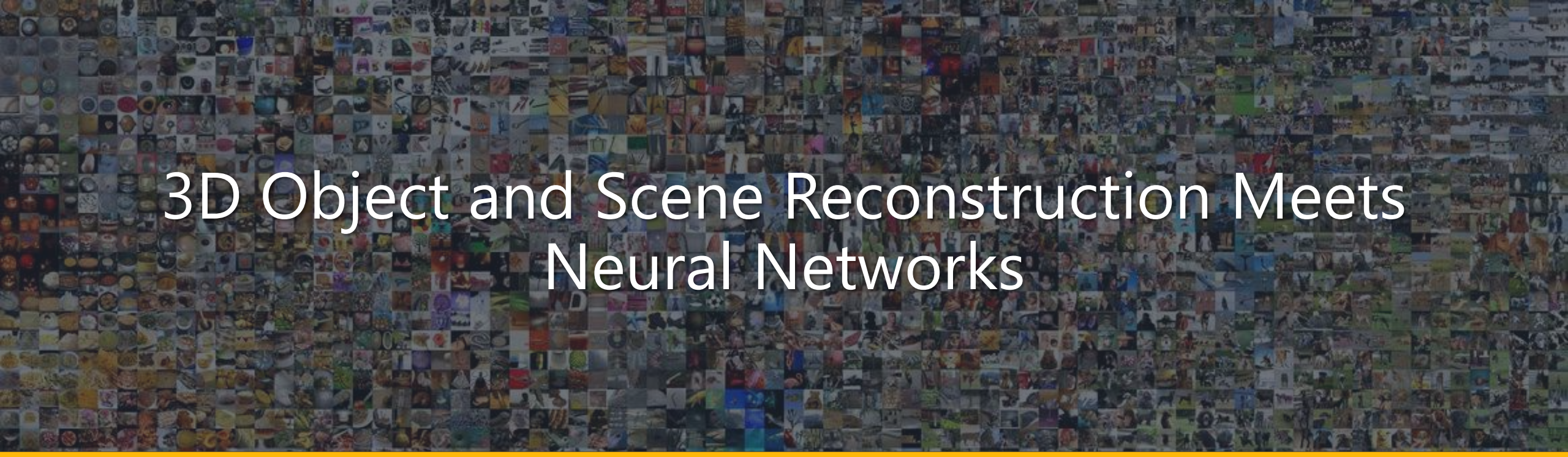


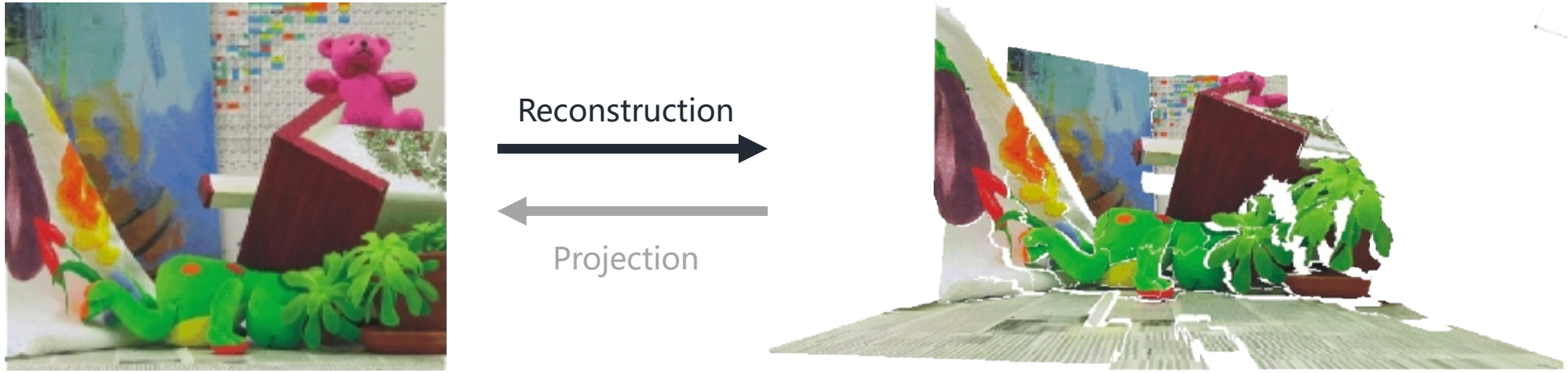
3D Object and Scene Reconstruction Meets Neural Networks

Haozhe Xie

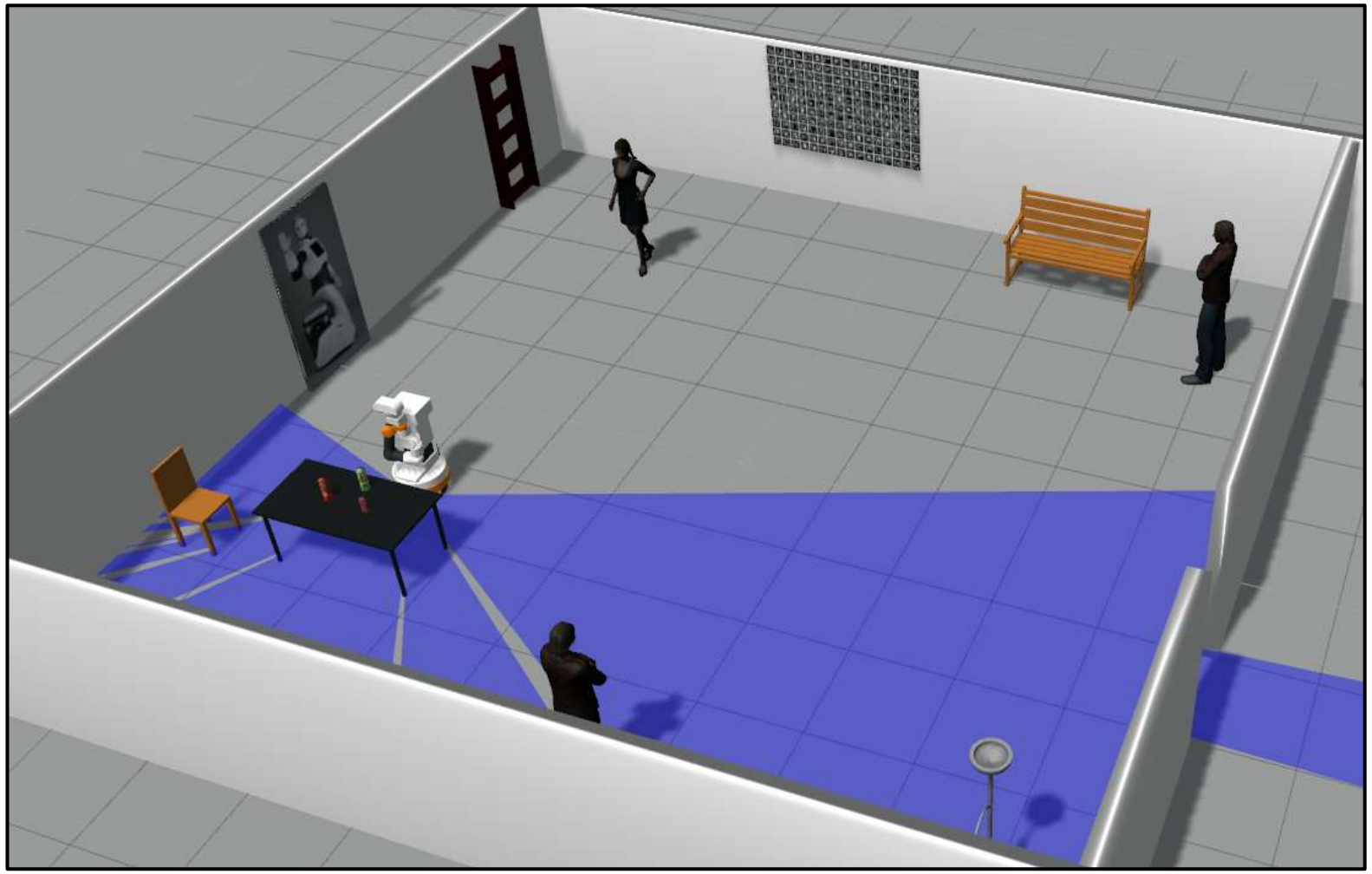
Tencent AI Lab

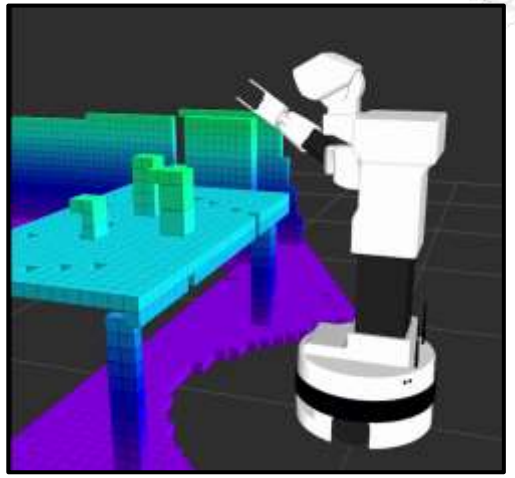
Background

- AI endow machines with the ability to perceive the real-world
  - Rapid development in 2D image understanding
  - Requires 3D information during the interaction with the 3D objects



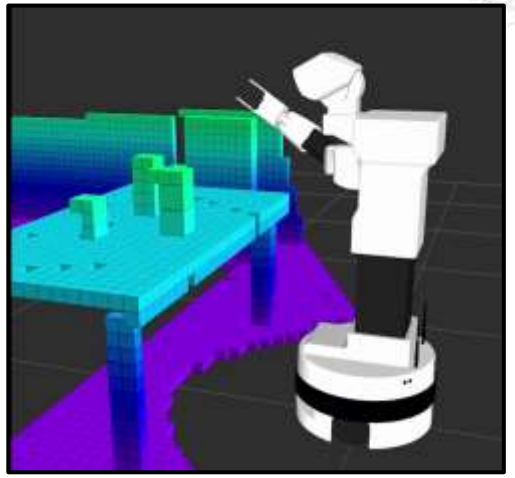
Background





Background

- Challenges
  - Ill-Posed Problem
  - Occlusion
- Opportunities
  - Rapid development in neural networks
  - Large-scale 3D datasets available



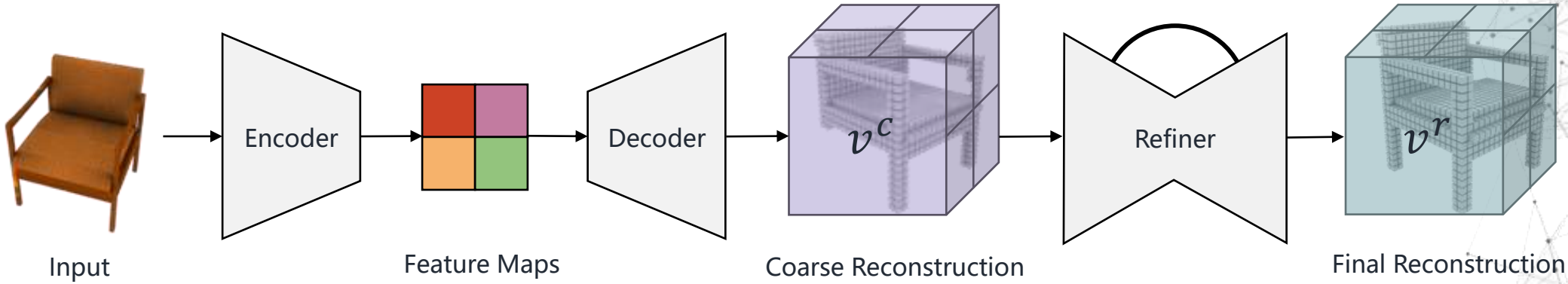
Outline

- Single-view 3D Object Reconstruction
  - From an RGB image
  - From stereo RGB images
  - From a depth image
- Multi-view 3D Object Reconstruction
- Multi-view 3D Scene Reconstruction

Single-view 3D Object Reconstruction

Single-view Reconstruction from a single RGB Image

- Overview



Xie et al. Pix2Vox: Context-aware 3D Reconstruction from Single and Multi-view Images. ICCV 2019.

Single-view Reconstruction from a single RGB Image

- Evaluation Metrics
  - Intersection over Union
    - $\mathrm{IoU} = \frac{\sum_{i,j,k} \mathrm{I}(p_{(i,j,k)} > t)\mathrm{I}(gt_{(i,j,k)} > t)}{\sum_{i,j,k} \mathrm{I}\left(\mathrm{I}(p_{(i,j,k)} > t) + \mathrm{I}(gt_{(i,j,k)} > t)\right)}$
    - where
      - $\mathrm{I}(\cdot)$ is the indicator
      - t is the threshold for binarize
      - $p_{(i,j,k)}$ is the prediction at (i, j, k)
      - $gt_{(i,j,k)}$ is the GT at (i, j, k)

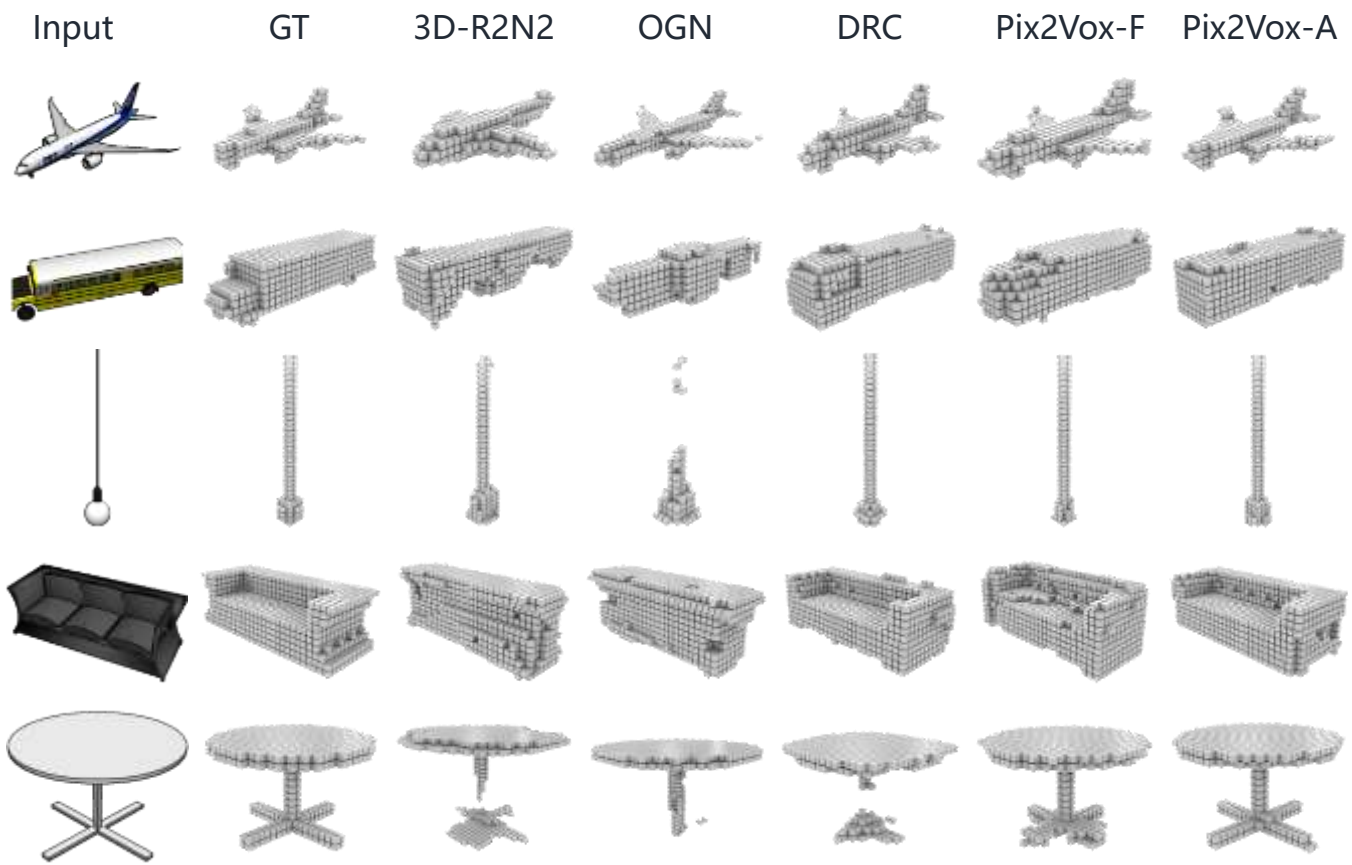
Single-view Reconstruction from a single RGB Image

- Experimental Results on ShapeNet

Category	3D-R2N2	OGN	DRC	PSGN	Pix2Vox-F	Pix2Vox-A
airplane	0.513	0.587	0.571	0.601	0.600	0.684
bench	0.421	0.481	0.453	0.550	0.538	0.616
cabinet	0.716	0.729	0.635	0.771	0.765	**0.792**
car	0.798	0.828	0.755	0.831	0.837	**0.854**
chair	0.466	0.483	0.469	0.544	0.535	**0.567**
display	0.468	0.503	0.419	**0.552**	0.511	0.537
lamp	0.381	0.398	0.415	**0.462**	0.435	0.443
speaker	0.662	0.637	0.609	**0.737**	0.707	0.714
rifle	0.544	0.593	0.608	0.604	0.598	**0.615**
sofa	0.628	0.646	0.606	0.708	0.687	**0.709**
table	0.513	0.536	0.424	**0.606**	0.587	0.601
telephone	0.661	0.702	0.413	0.749	0.770	**0.776**
watercraft	0.513	**0.632**	0.556	0.611	0.582	0.594
Avg. IoU	0.560	0.596	0.545	0.640	0.634	**0.661**

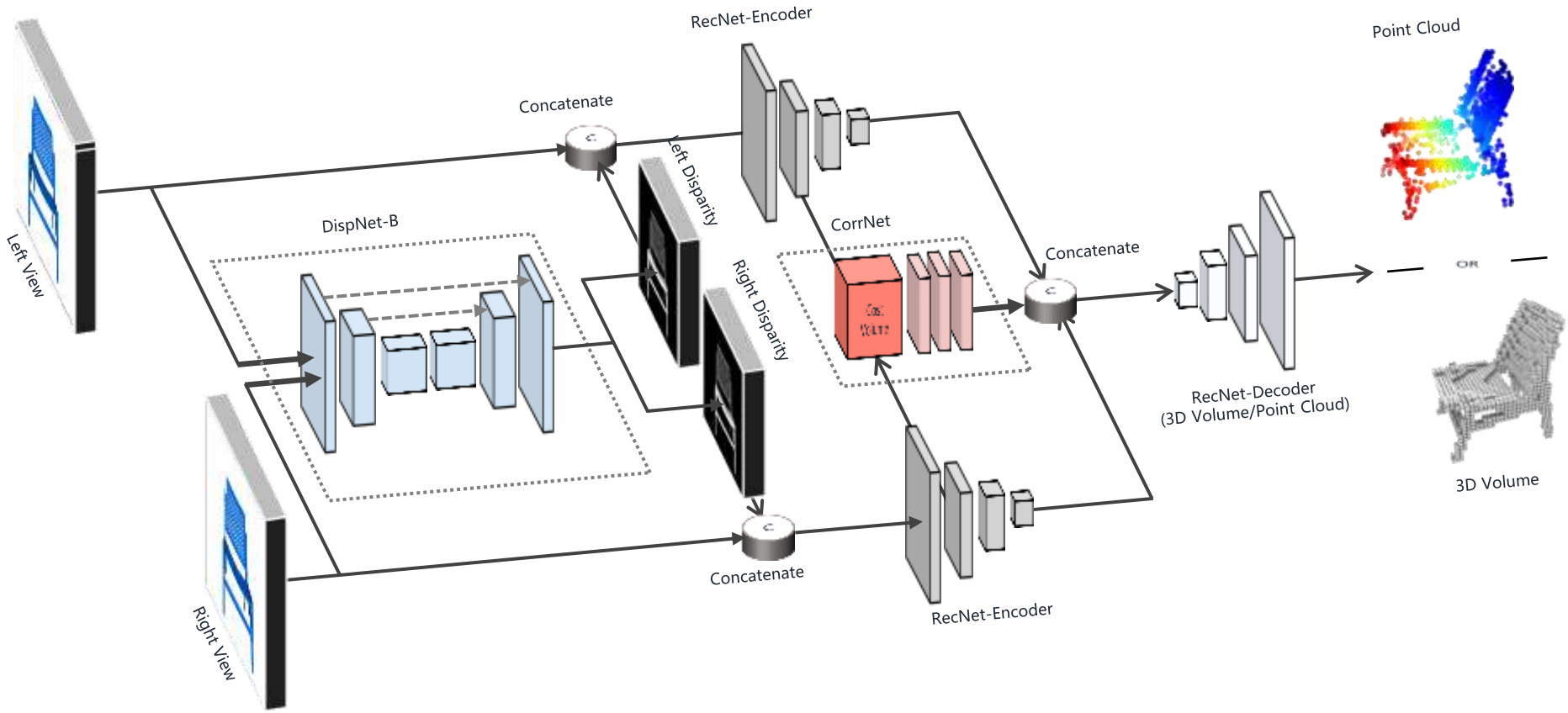
Single-view Reconstruction from a single RGB Image

- Experimental Results on ShapeNet



Single-view Reconstruction from stereo RGB Images

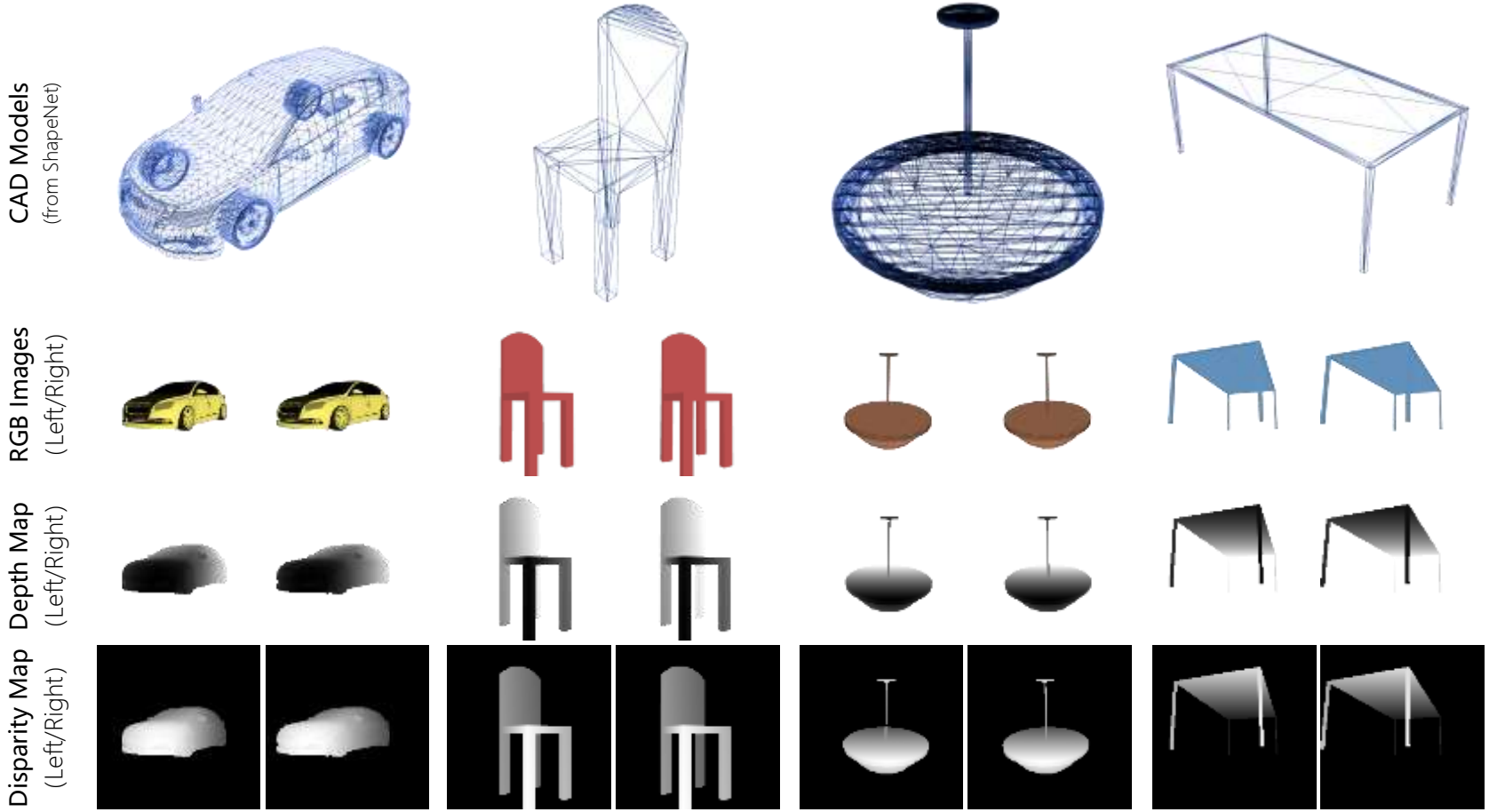
- Overview



Xie et al. Toward 3D Object Reconstruction from Stereo Images. Neurocomputing (463) 444-453, 2021.

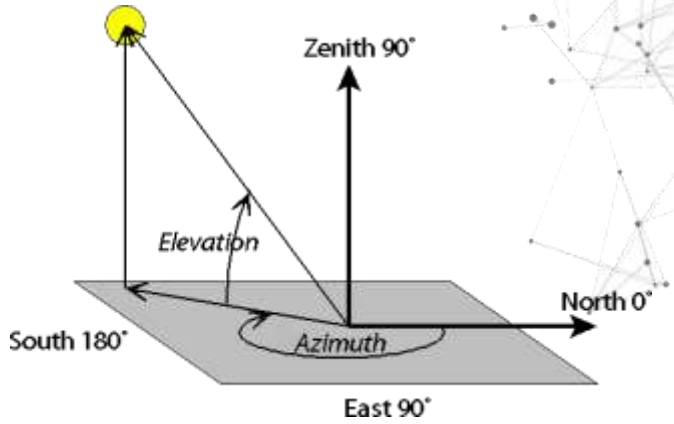
Single-view Reconstruction from stereo RGB Images

- The StereoShapeNet Dataset



Single-view Reconstruction from stereo RGB Images

- The StereoShapeNet Dataset
  - Including 1,052,976 stereo image pairs, with depth and disparity maps
  - Viewpoint Parameters
    - Azimuth: $\theta_{az} \in [0°, 360°)$
    - Elevation: $\theta_{el} \in [-30°, 30°]$
  - Camera Parameters
    - Focal Length 35 mm
    - Baseline 130 mm
    - Image Resolution 224×224



Single-view Reconstruction from stereo RGB Images

- Evaluation Metrics
  - Intersection over Union
    - $\mathrm{IoU} = \frac{\sum_{i,j,k} \mathrm{I}(p_{(i,j,k)} > t)\mathrm{I}(gt_{(i,j,k)} > t)}{\sum_{i,j,k} \mathrm{I}\left(\mathrm{I}(p_{(i,j,k)} > t) + \mathrm{I}(gt_{(i,j,k)} > t)\right)}$
  - Chamfer Distance
    - $\mathrm{CD} = \frac{1}{n_{\mathrm{gt}}} \sum_{p \in \mathbf{P}} \min_{q \in \widehat{\mathbf{P}}} \|p - q\|_2^2 + \frac{1}{n_{\mathrm{p}}} \sum_{p \in \widehat{\mathbf{P}}} \min_{q \in \mathbf{P}} \|p - q\|_2^2$
    - where
      - $\mathbf{P} = \{(x_i, y_i, z_i)\}_{\mathrm{i}=1}^{\mathrm{n_{gt}}}$ denotes the prediction
      - $\widehat{\mathbf{P}} = \{(x_i, y_i, z_i)\}_{\mathrm{i}=1}^{\mathrm{n_p}}$ denotes the GT

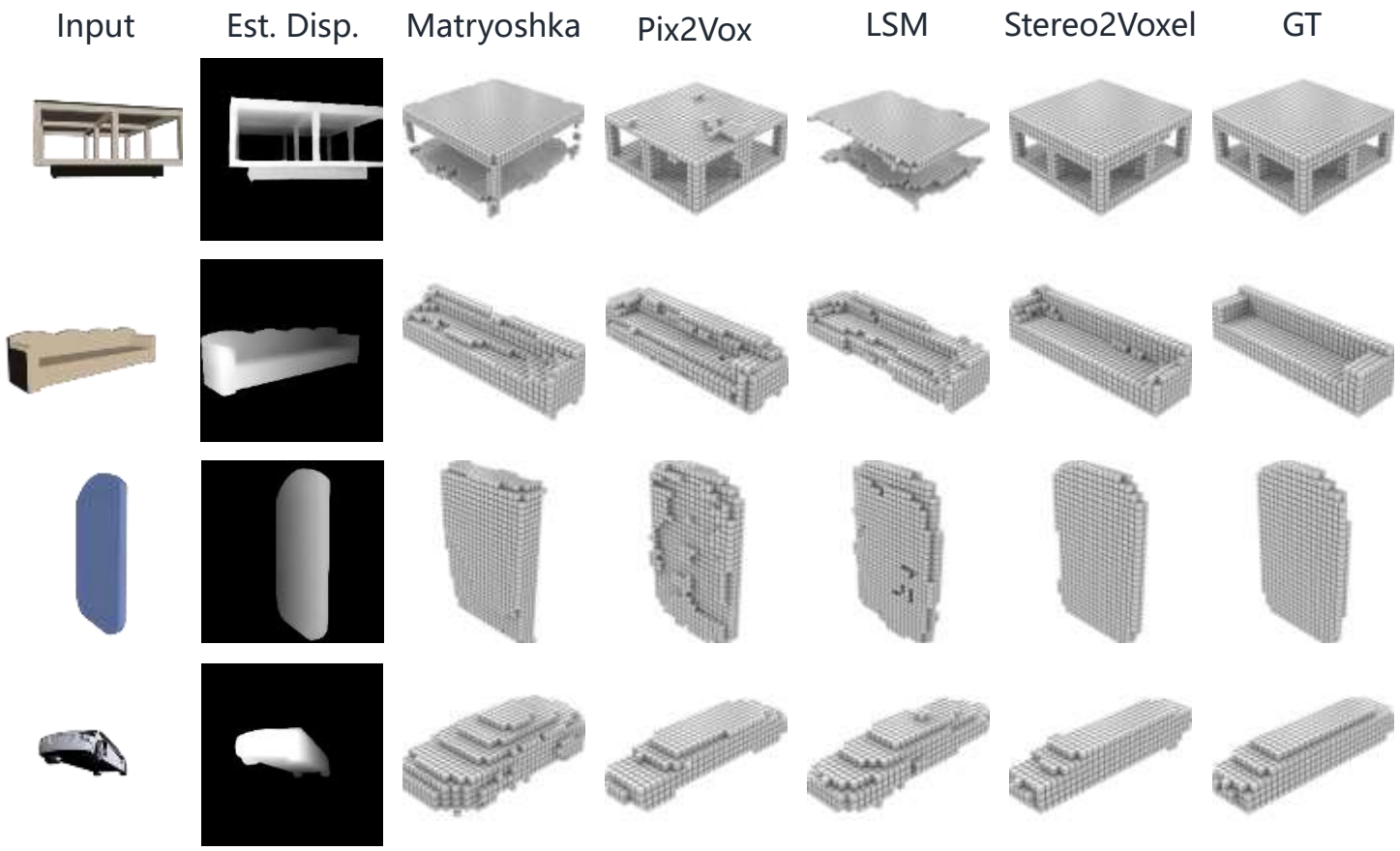
Single-view Reconstruction from stereo RGB Images

- Experimental Results on StereoShapeNet

Category	Matryoshka	Matryoshka*	Pix2Vox	LSM	Stereo2Voxel
airplane	0.557	0.535	0.686	0.621	**0.709**
bench	0.524	0.473	0.566	0.517	**0.622**
cabinet	0.766	0.763	0.754	0.691	**0.784**
car	0.827	0.810	0.811	0.796	**0.830**
chair	0.559	0.514	0.604	0.595	**0.669**
display	0.635	0.614	0.586	0.547	**0.692**
lamp	0.424	0.411	0.449	0.469	**0.521**
speaker	0.697	0.727	0.658	0.670	**0.701**
rifle	0.540	0.557	0.652	0.682	**0.690**
sofa	0.702	0.679	0.714	0.651	**0.770**
table	0.559	0.503	0.570	0.566	**0.635**
telephone	0.759	0.847	0.831	0.694	**0.866**
watercraft	0.587	**0.595**	0.558	0.592	**0.645**
Avg. IoU	0.626	0.603	0.652	0.632	**0.702**

Single-view Reconstruction from stereo RGB Images

- Experimental Results on StereoShapeNet



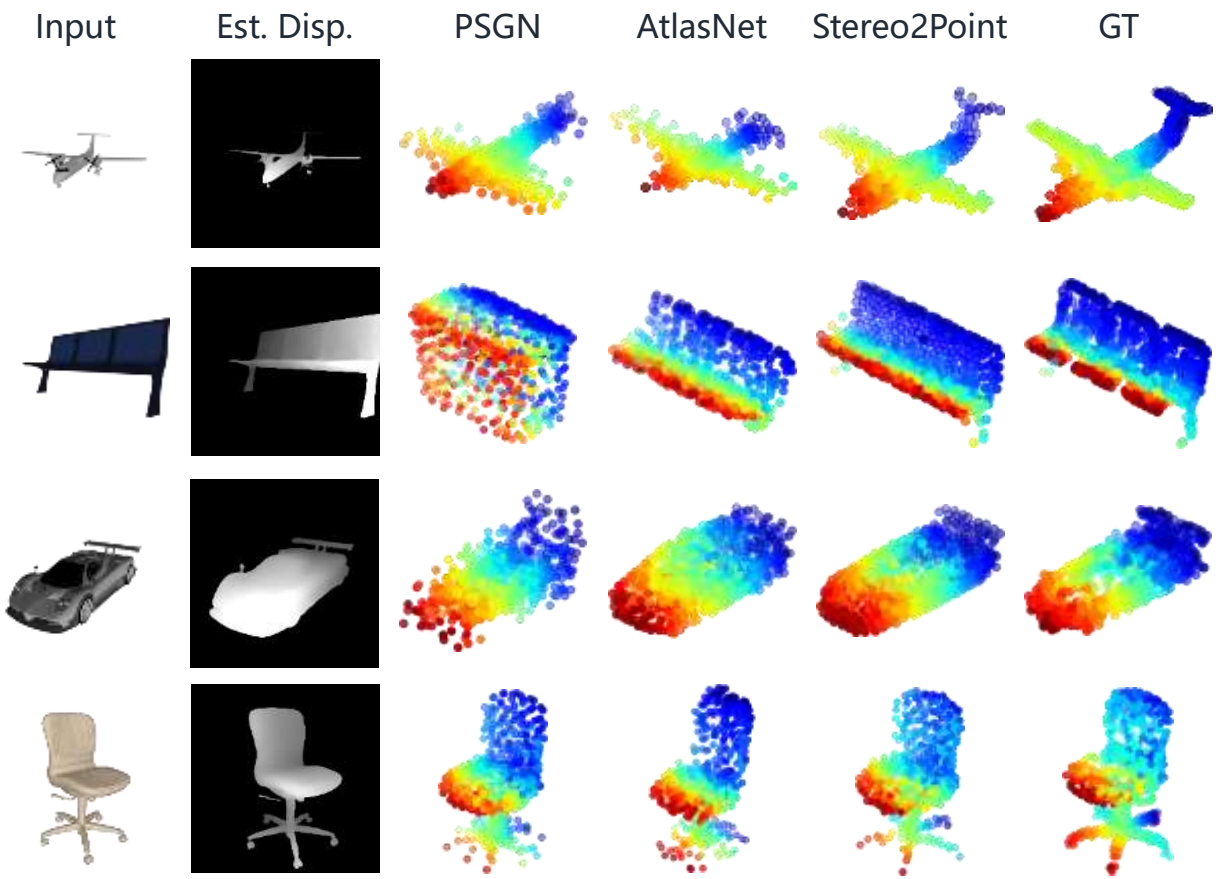
Single-view Reconstruction from stereo RGB Images

- Experimental Results on StereoShapeNet

Category	PSGN	PSGN*	AtlasNet	AtlasNet*	Stereo2Point
airplane	0.826	0.699	0.807	0.796	**0.534**
bench	1.789	1.695	1.996	1.796	**1.182**
cabinet	2.360	1.853	1.756	1.692	**1.229**
car	1.295	0.882	1.045	1.036	**0.779**
chair	2.004	1.594	1.837	1.858	**1.267**
display	2.815	2.238	2.386	2.146	**1.356**
lamp	3.973	3.038	4.142	4.118	**3.001**
speaker	3.868	2.691	2.839	2.869	**2.124**
rifle	0.790	0.763	0.818	0.874	**0.524**
sofa	2.625	2.086	1.664	1.656	**1.199**
table	1.889	1.500	1.892	1.916	**1.337**
telephone	1.445	1.158	1.156	1.250	**0.896**
watercraft	2.029	1.495	1.712	1.524	**1.027**
Avg. CD	1.916	1.493	1.704	1.689	**1.185**

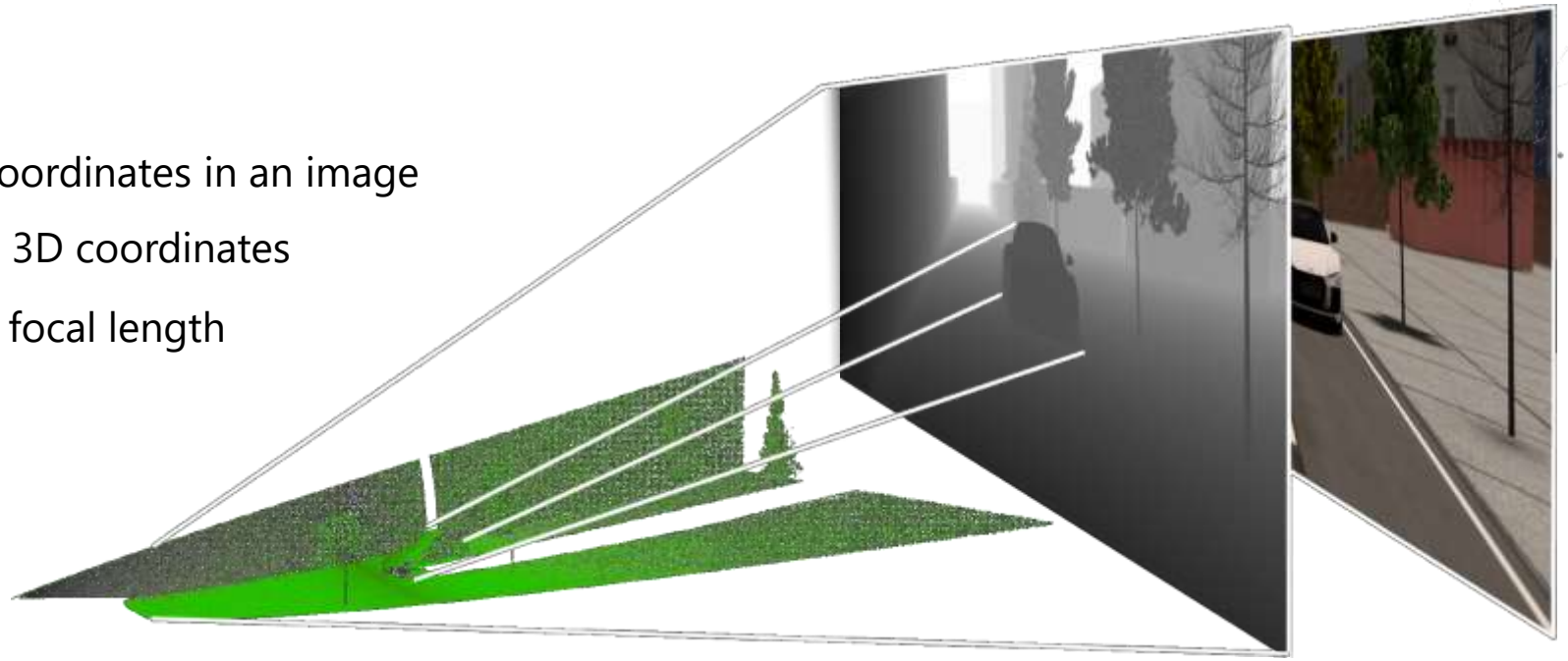
Single-view Reconstruction from stereo RGB Images

- Experimental Results on StereoShapeNet



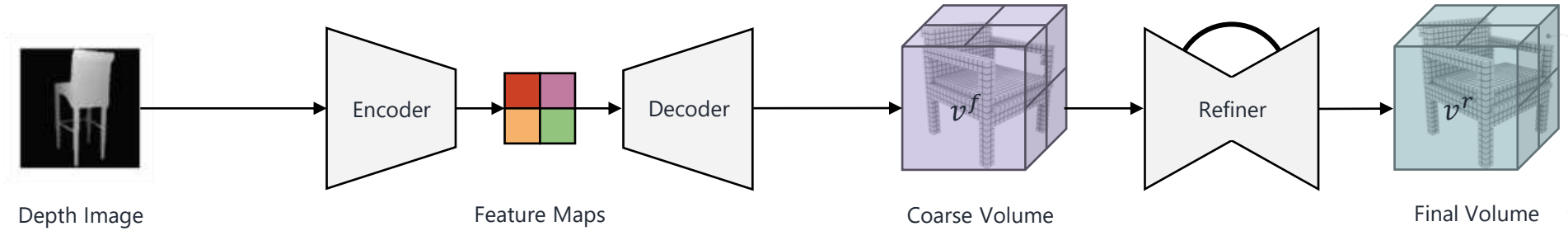
Single-view Reconstruction from a depth Image

- The difference between RGB and depth images
  - RGB Image $\mathbf{I} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 3}$: with RGB information
  - Depth Image $\mathbf{D} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 1}$: with XYZ information
    - $x = \frac{uz}{f_x}\, y = \frac{vz}{f_y}$
    - where
      - (u, v) is the coordinates in an image
      - (x, y, z) is the 3D coordinates
      - $f_x,\ f_y$ are the focal length



Single-view Reconstruction from a depth Image

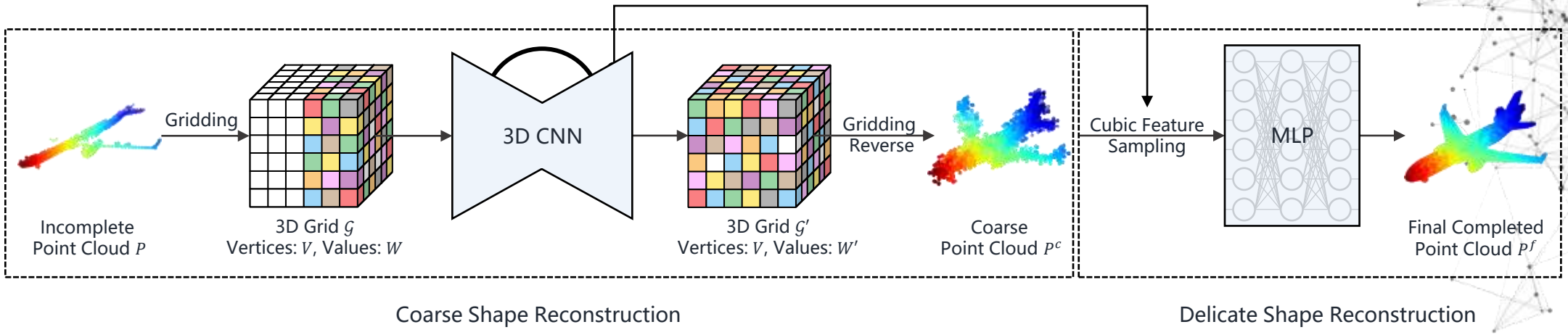
- Related work
  - 2D Convolutional Neural Networks



  - Drawbacks
    - Cannot explicitly regress the 3D coordinates of an object

Single-view Reconstruction from a depth Image

- Our Solution

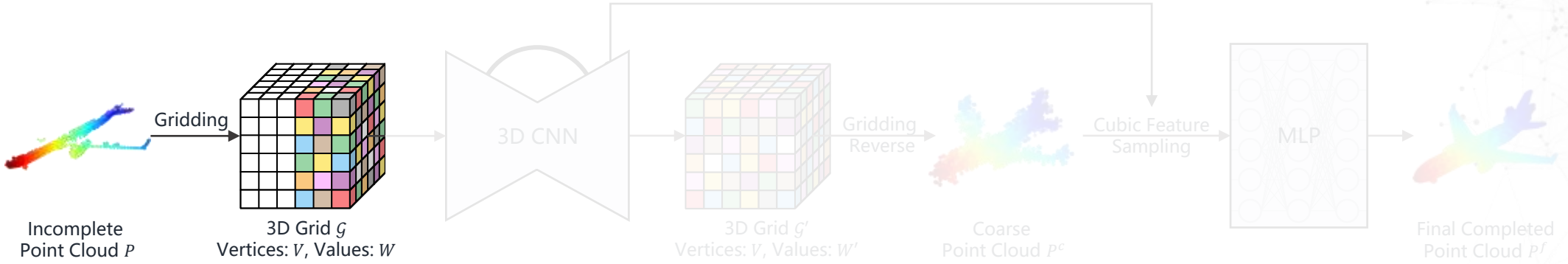


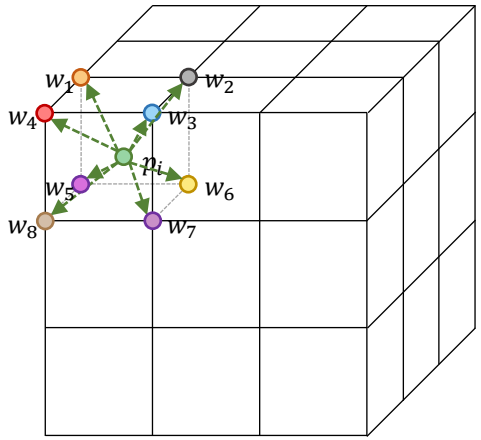
- Three differentiable operations
  - Gridding
  - Gridding Reverse
  - Cubic Feature Sampling
- Two sub-networks
  - 3D CNN
  - MLP

Xie et al. GRNet: Gridding Residual Network for Dense Point Cloud Completion. ECCV 2020.

Single-view Reconstruction from a depth Image

- Gridding





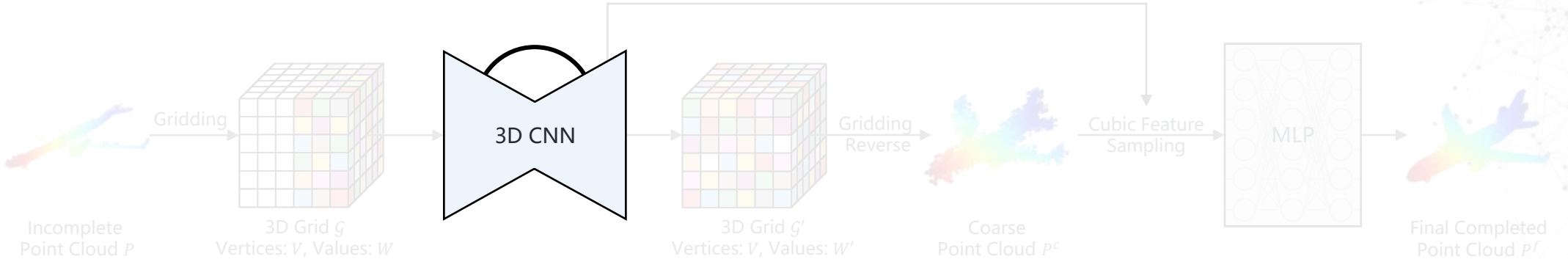
The value of the vertex w_i can be computed as

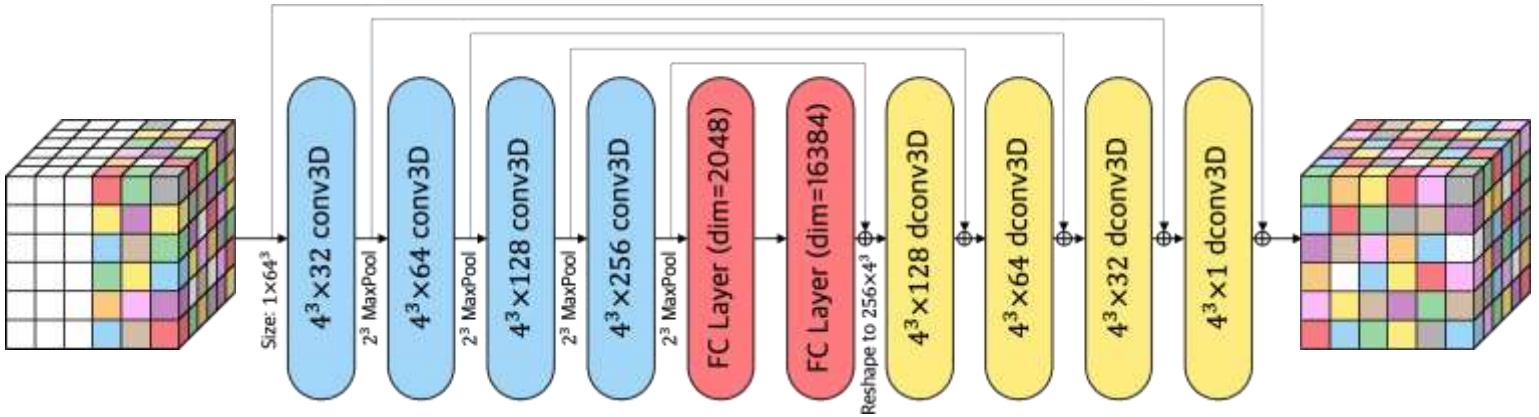
$$w_i = (1 - |x_i - x|)(1 - |y_i - y|)(1 - |z_i - z|)$$

where (x_i, y_i, z_i) and (x, y, z) are the coordinates of w_i and p_i, respectively.

Single-view Reconstruction from a depth Image

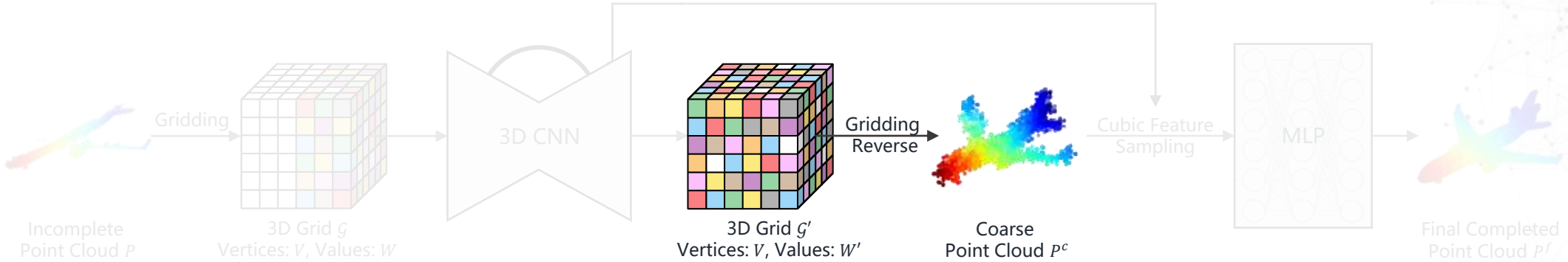
- 3D CNN

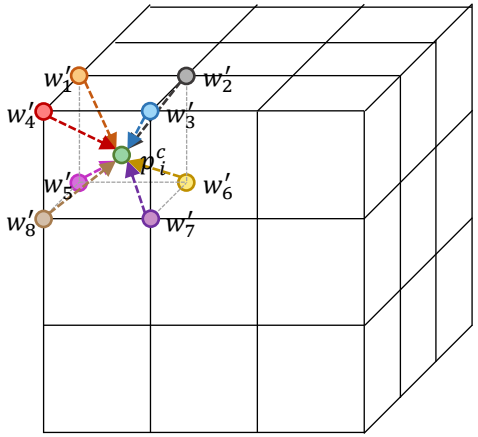




Single-view Reconstruction from a depth Image

- Gridding Reverse





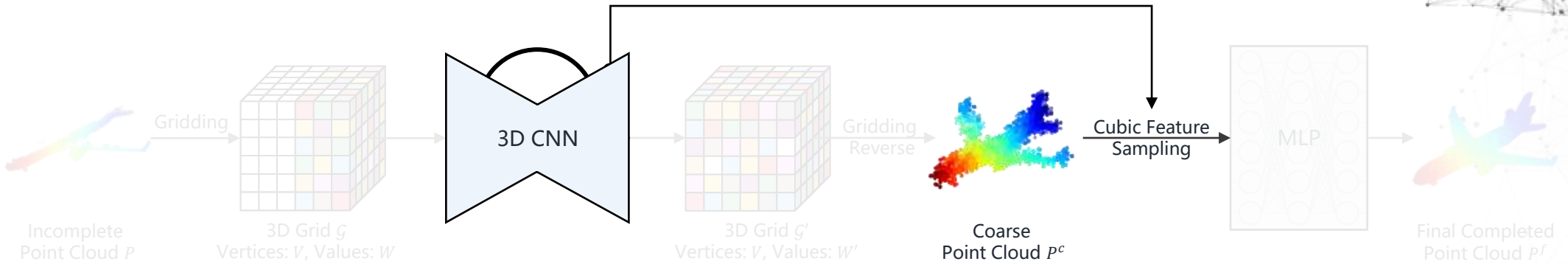
The coordinate of the generated point $\boldsymbol{p_i^c}$ can be computed as

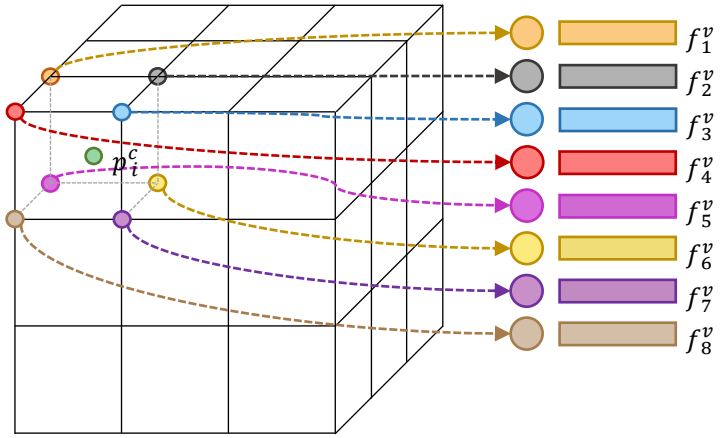
$$\boldsymbol{p_i^c} = \frac{\sum_i w_i' \boldsymbol{v_i}}{\sum_i w_i'}$$

where the $\boldsymbol{v_i}$ and w_i' be the coordinate and value of the vertex i ($i = 1, \ldots, 8$).

Single-view Reconstruction from a depth Image

- Cubic Feature Sampling



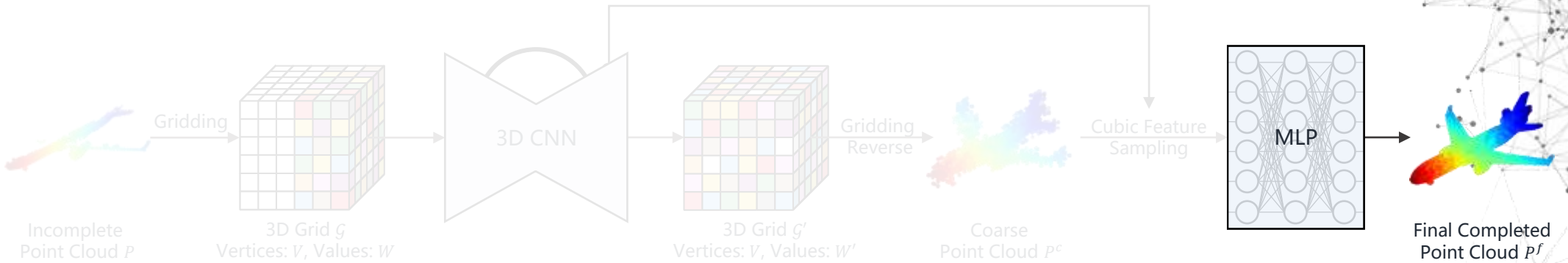


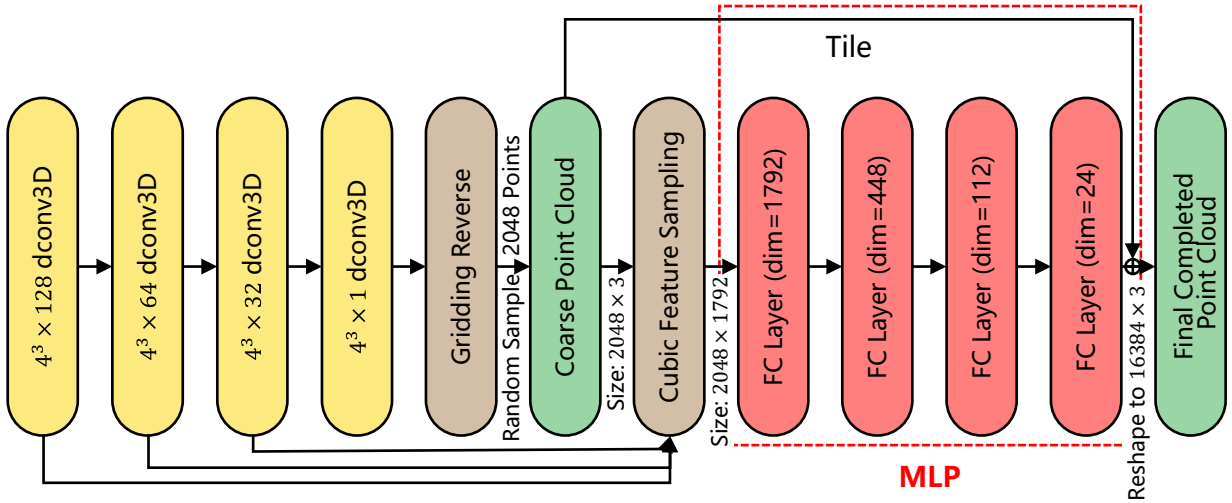
Therefore, the features f_i^c for point p_i^c can be computed as:

$$f_i^c = [f_1^v, f_2^v, \dots, f_8^v]$$

Single-view Reconstruction from a depth Image

- MLP





Single-view Reconstruction from a depth Image

- Evaluation Metrics
  - Chamfer Distance
    - $\mathrm{CD} = \frac{1}{n_{\mathrm{gt}}} \sum_{p \in \mathbf{P}} \min_{q \in \widehat{\mathbf{P}}} \|p - q\|_2^2 + \frac{1}{n_{\mathrm{p}}} \sum_{p \in \widehat{\mathbf{P}}} \min_{q \in \mathbf{P}} \|p - q\|_2^2$
    - where
      - $\mathbf{P} = \{(x_i, y_i, z_i)\}_{\mathrm{i}=1}^{\mathrm{n_{gt}}}$ denotes the prediction
      - $\widehat{\mathbf{P}} = \{(x_i, y_i, z_i)\}_{\mathrm{i}=1}^{\mathrm{n_p}}$ denotes the GT

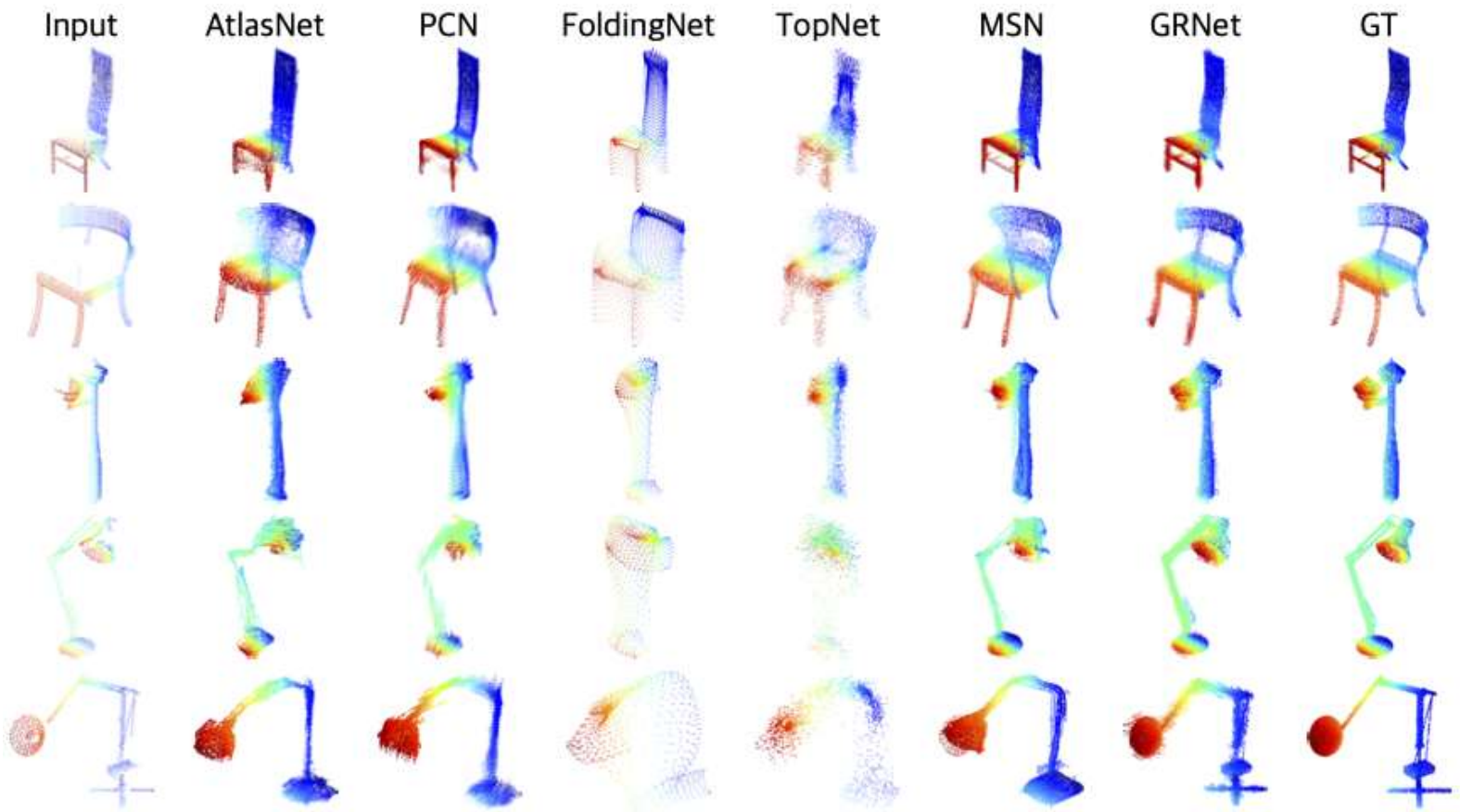
Single-view Reconstruction from a depth Image

- Experimental Results on ShapeNet

Category	AltasNet	PCN	FoldingNet	TopNet	MSN	GRNet
airplane	1.753	**1.400**	3.151	2.152	1.543	1.531
cabinet	5.101	4.450	7.943	5.623	7.249	**3.620**
car	3.237	**2.445**	4.676	3.513	4.711	2.752
chair	5.226	4.838	9.225	6.436	4.539	**2.945**
lamp	6.342	6.238	9.324	7.502	6.479	**2.649**
sofa	5.990	5.129	8.895	6.949	5.894	**3.613**
table	4.359	3.569	6.691	4.784	3.797	**2.552**
watercraft	4.177	4.062	7.325	4.359	3.853	**2.122**
Avg. CD	4.523	4.016	7.142	5.154	4.758	**2.723**

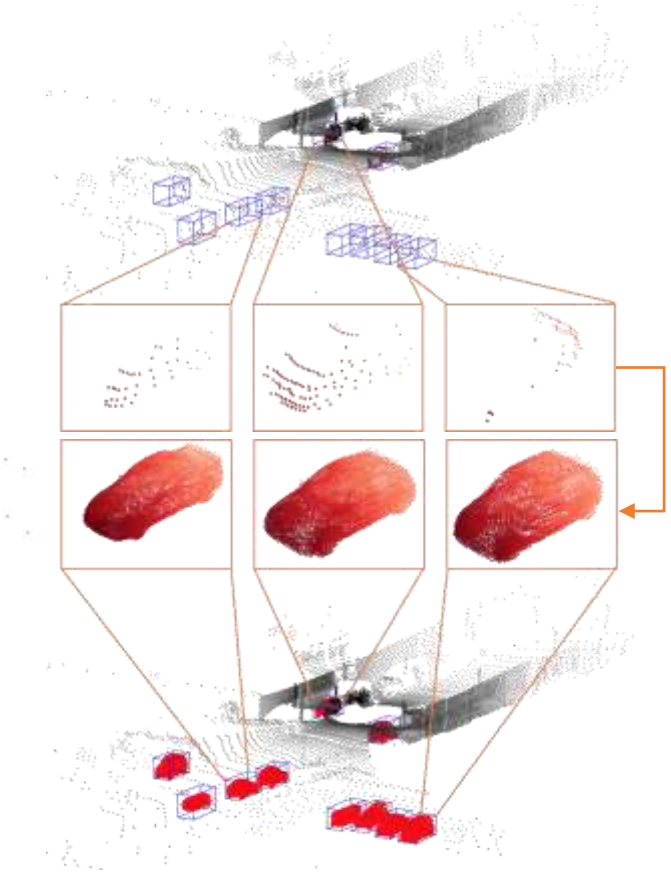
Single-view Reconstruction from a depth Image

- Experimental Results on ShapeNet



Single-view Reconstruction from a depth Image

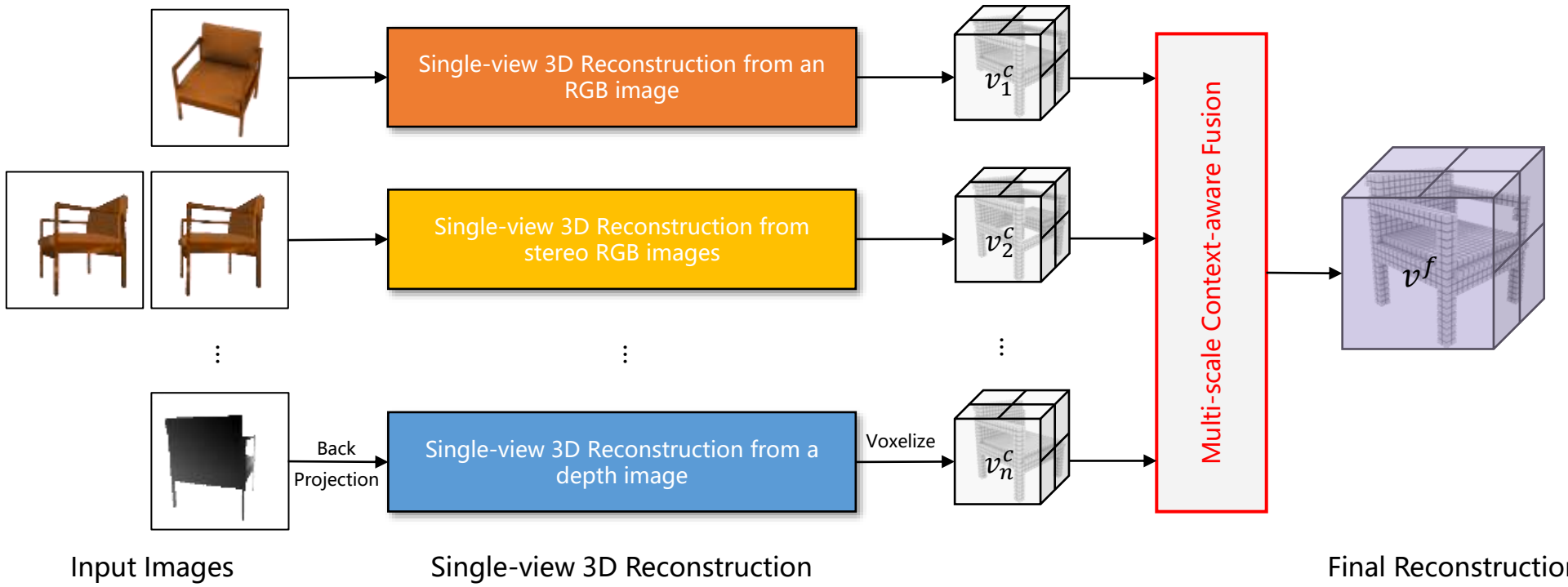
- Application: Autonomous Driving



Multi-view 3D Object Reconstruction

Multi-view 3D Object Reconstruction

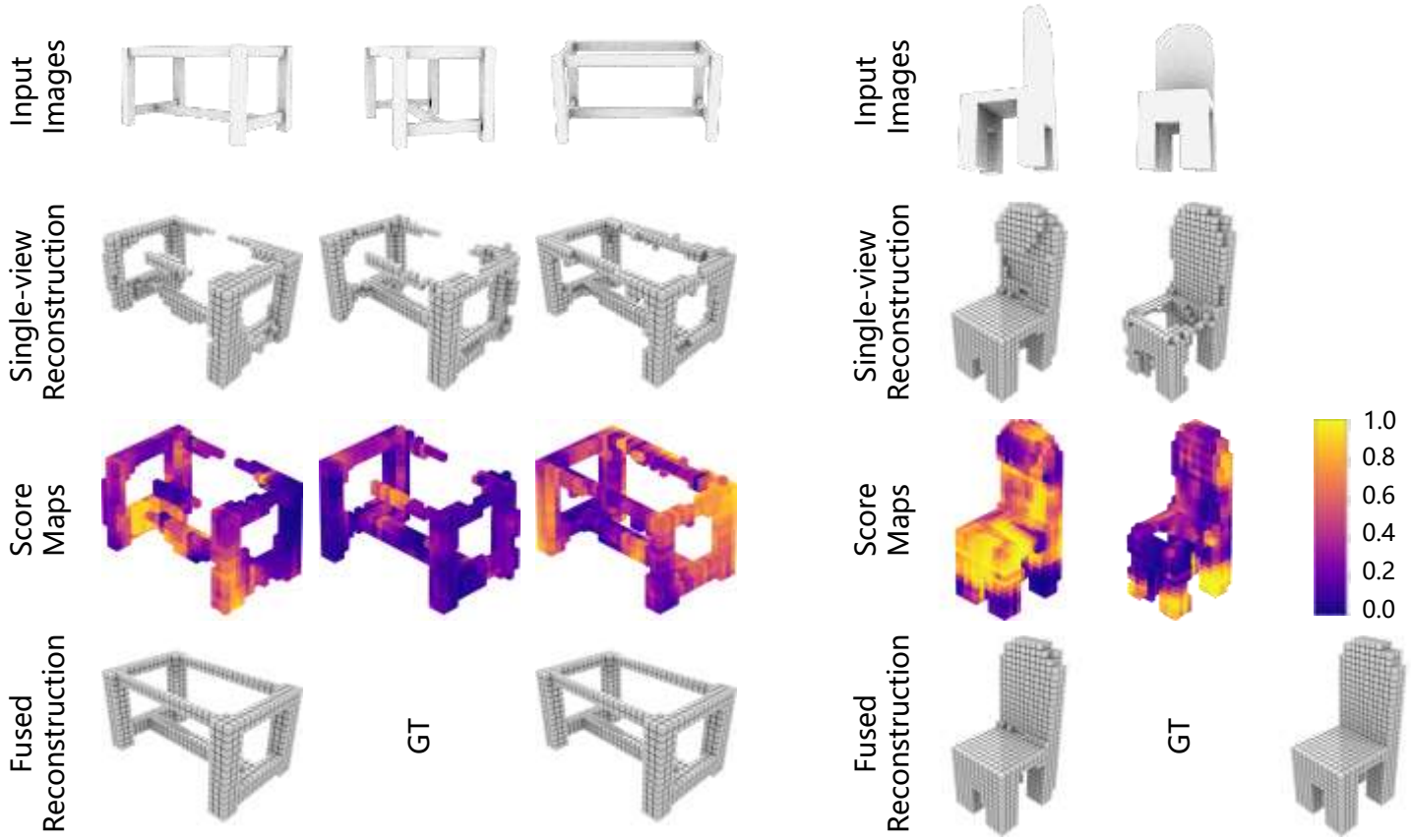
- Overview



Xie et al. Pix2Vox++: Multi-scale Context-aware 3D Object Reconstruction from Single and Multiple Images. IJCV 128 (12) 2919-2935, 2020.

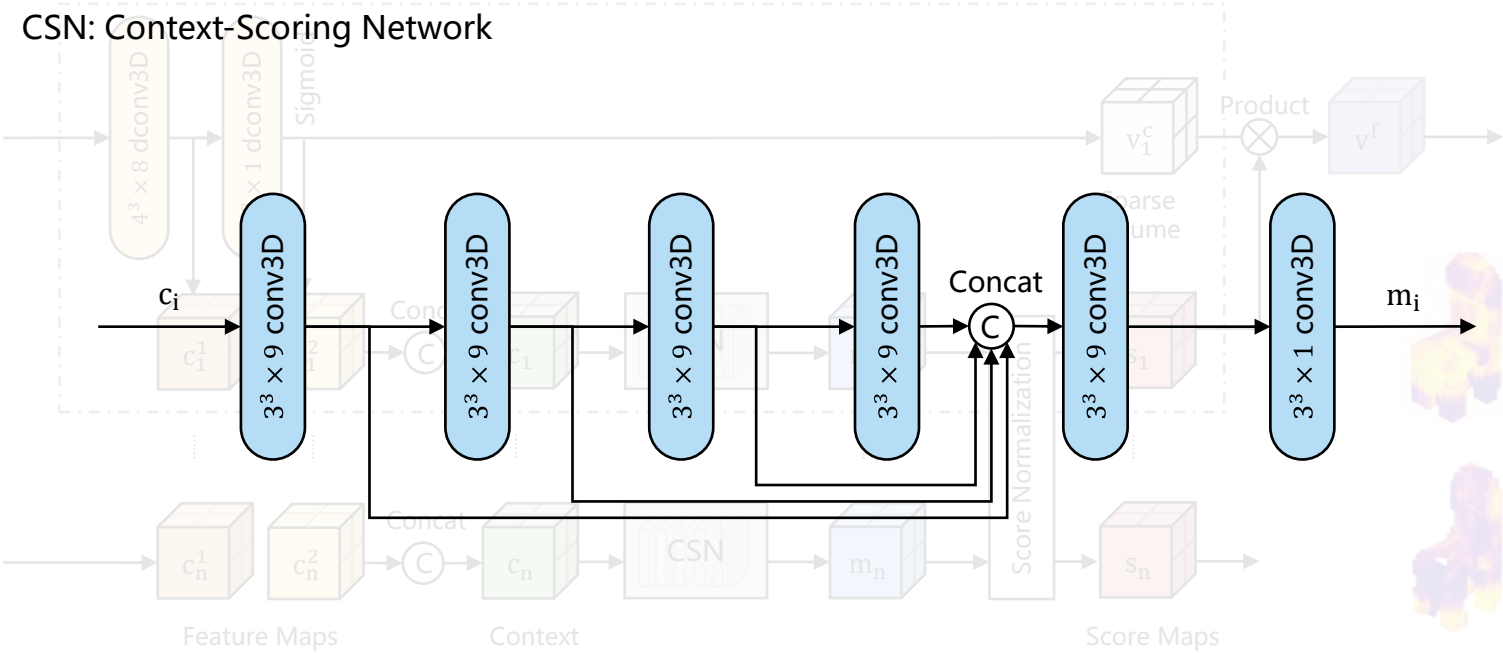
Multi-view 3D Object Reconstruction

- Multi-scale Context-aware Fusion



Multi-view 3D Object Reconstruction

- Overview



- Context: $\mathcal{C} = \{c_t | c_t \in \mathbb{R}^{9\times r\times r\times r}\}$
- Score Normalization: $s_t = \frac{\exp\left(m_t^{(i,j,k)}\right)}{\sum_{p=1}^{n}\exp\left(m_p^{(i,j,k)}\right)}$
- Scores: $\mathcal{M} = \{m_t | m_t \in \mathbb{R}^{r\times r\times r}\}$
- $\mathrm{v}^{\mathrm{f}} = \sum_{t=1}^{n} s_t v_t^c$

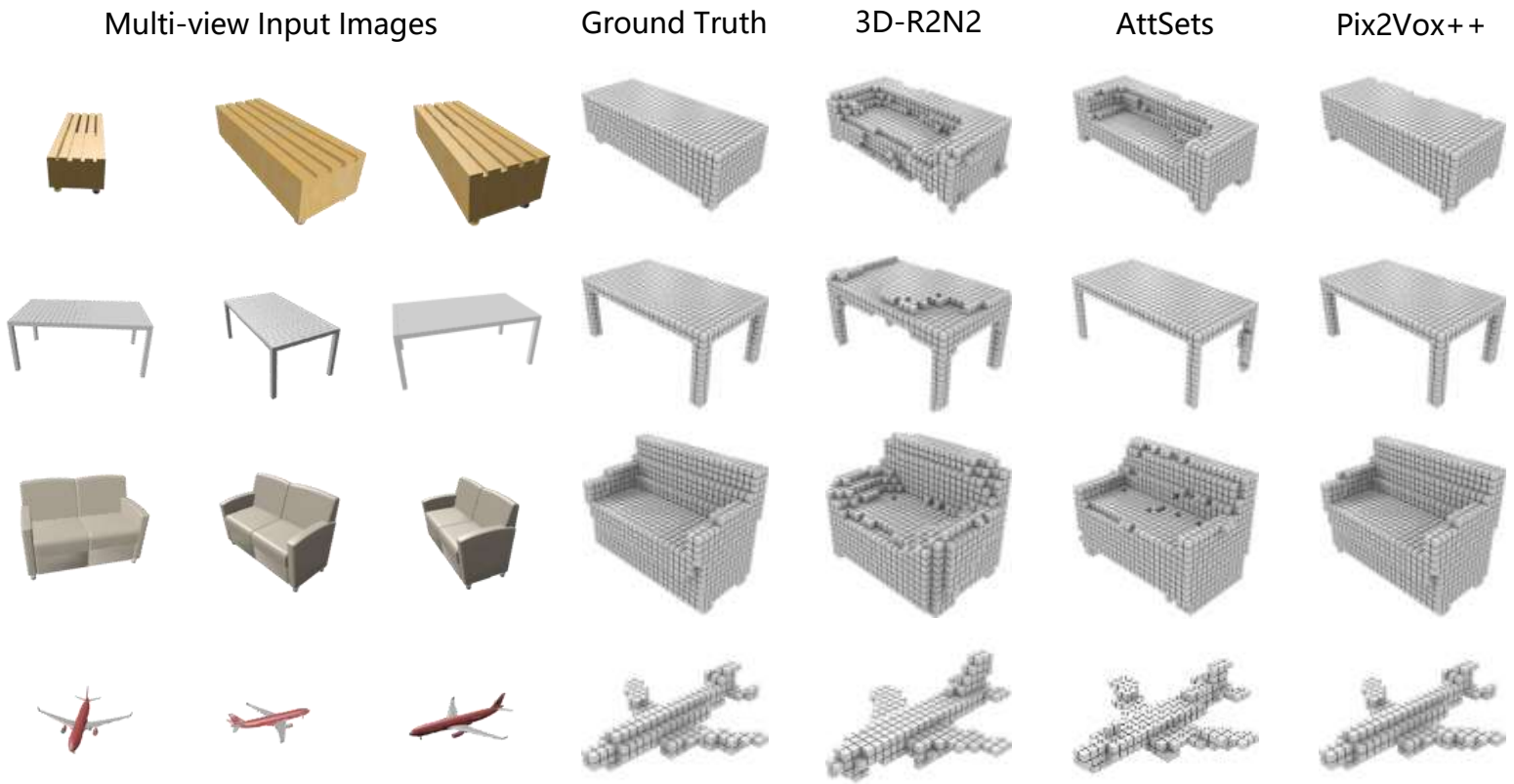
Multi-view 3D Object Reconstruction

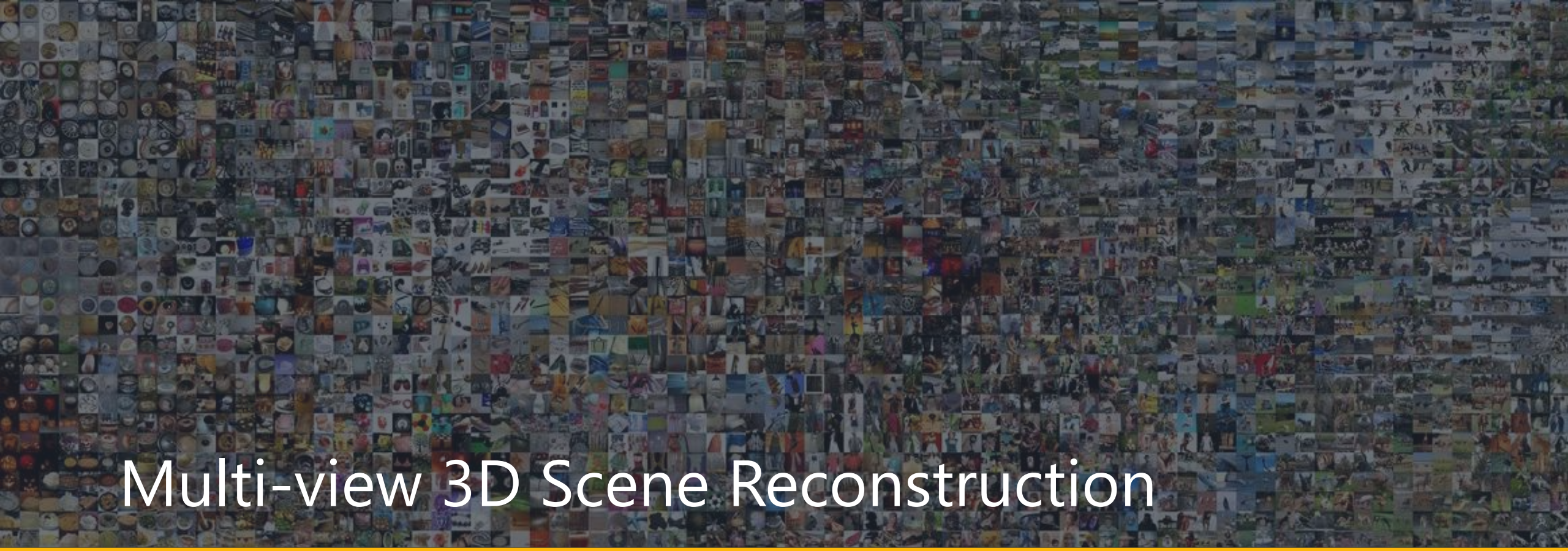
- Experimental Results on ShapeNet

# views	3D-R2N2	AttSets	Pix2Vox++
1 view	0.560	0.642	**0.670**
2 views	0.603	0.662	**0.695**
3 views	0.617	0.670	**0.704**
4 views	0.625	0.675	**0.708**
5 views	0.634	0.677	**0.711**
8 views	0.635	0.677	**0.715**
12 views	0.636	0.688	**0.717**
16 views	0.636	0.692	**0.718**
20 views	0.636	0.693	**0.719**

Multi-view 3D Object Reconstruction

- Experimental Results on ShapeNet

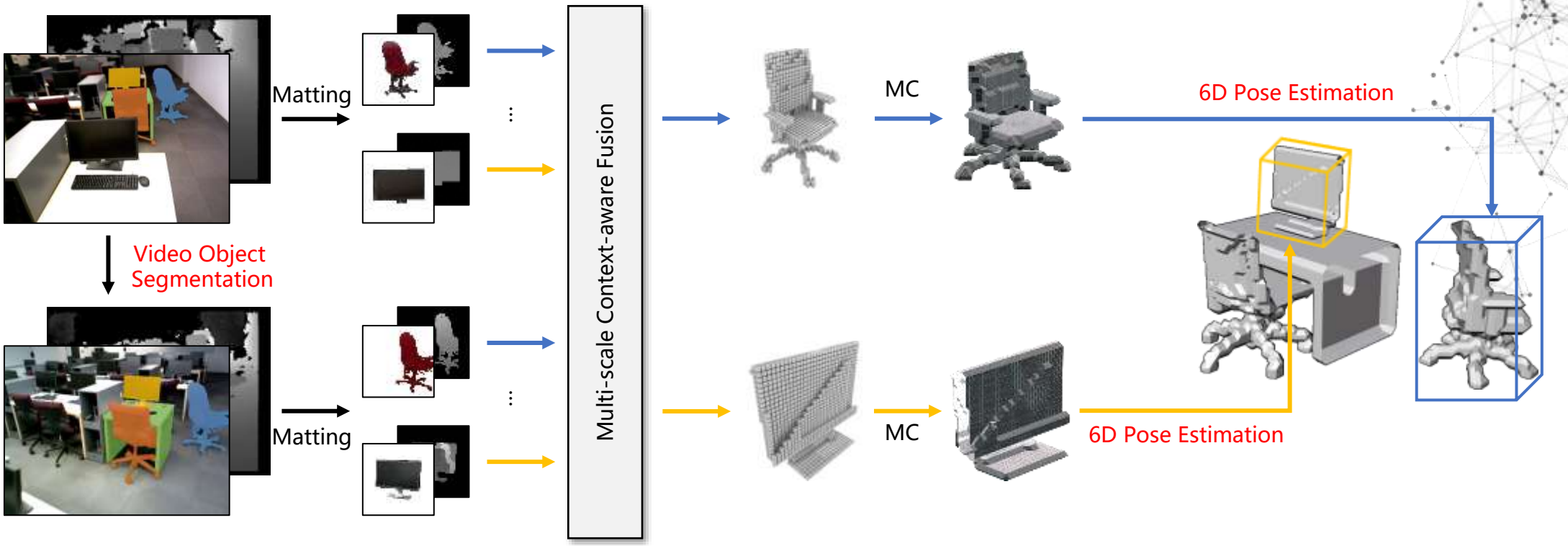




Multi-view 3D Scene Reconstruction

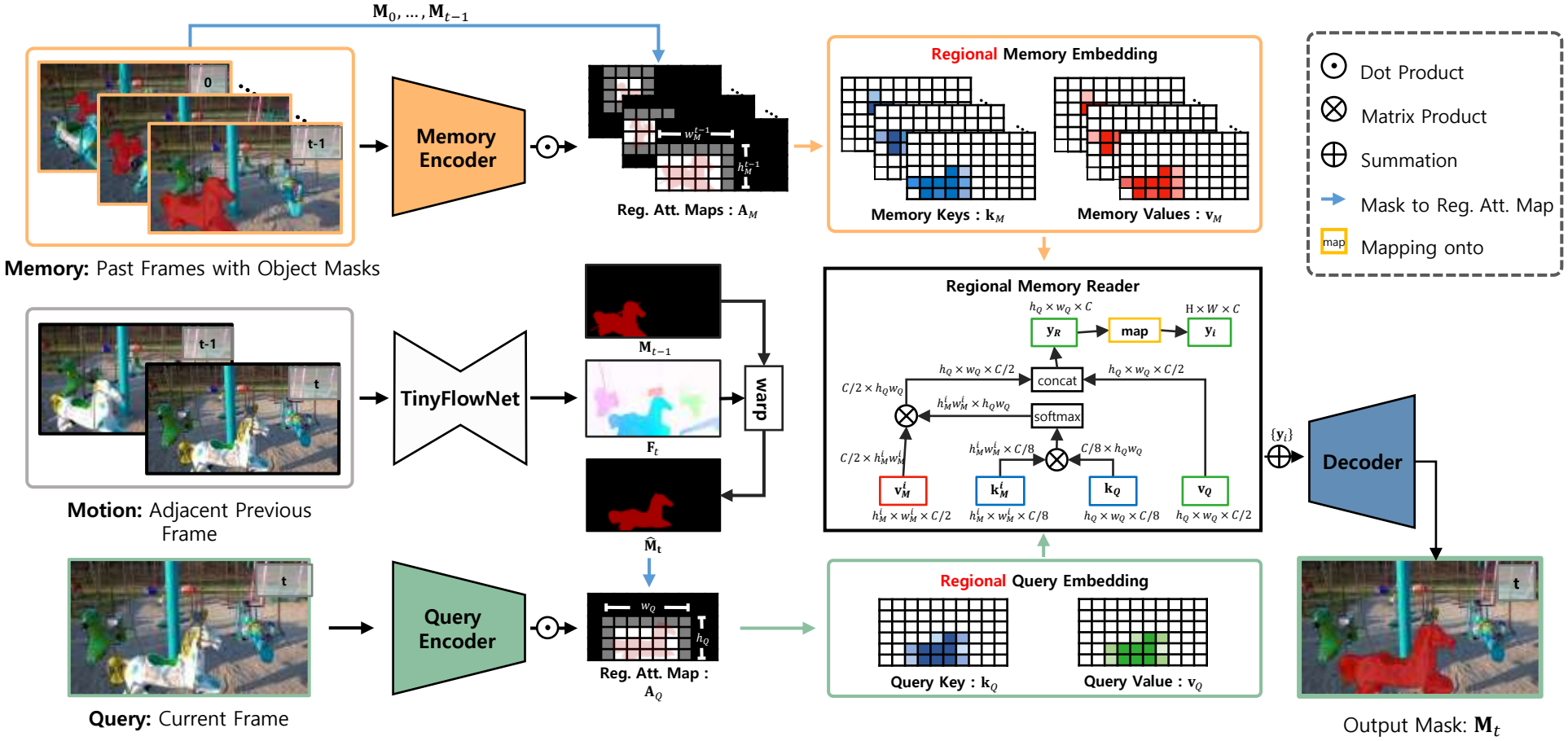
Multi-view 3D Scene Reconstruction

- Overview



Multi-view 3D Scene Reconstruction

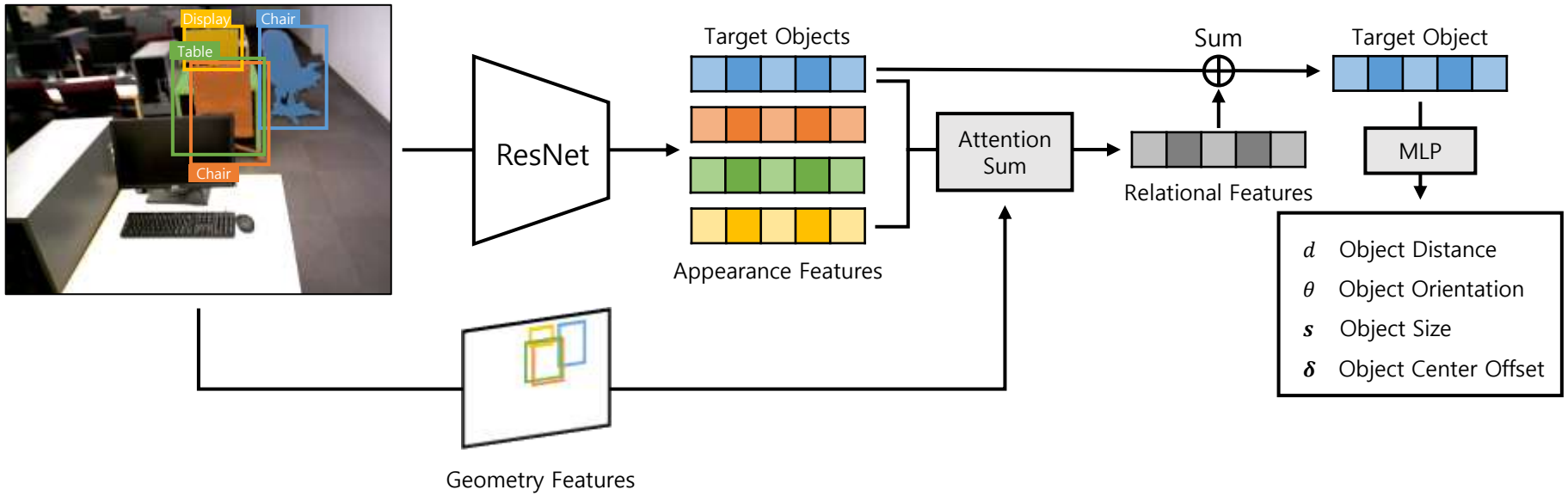
- Video Object Segmentation



Xie et al. Efficient Regional Memory Network for Video Object Segmentation. CVPR 2021.

Multi-view 3D Scene Reconstruction

- 6D Pose Estimation



- The center $\mathbf{C}$ of the 3D bounding box
  - $\mathbf{C} = \mathbf{R}^{-1} \mathrm{d} \frac{\mathbf{K}^{-1}[\mathbf{c}^{\mathrm{b}}+\boldsymbol{\delta},1]^{\mathrm{T}}}{\left\|\mathbf{K}^{-1}[\mathbf{c}^{\mathrm{b}}+\boldsymbol{\delta},1]^{\mathrm{T}}\right\|_2}$, where $\mathbf{c}^{\mathrm{b}}$ is the center of the 2D bounding box

Nie et al. Total3DUnderstanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes from a Single Image. CVPR 2020.

Multi-view 3D Scene Reconstruction

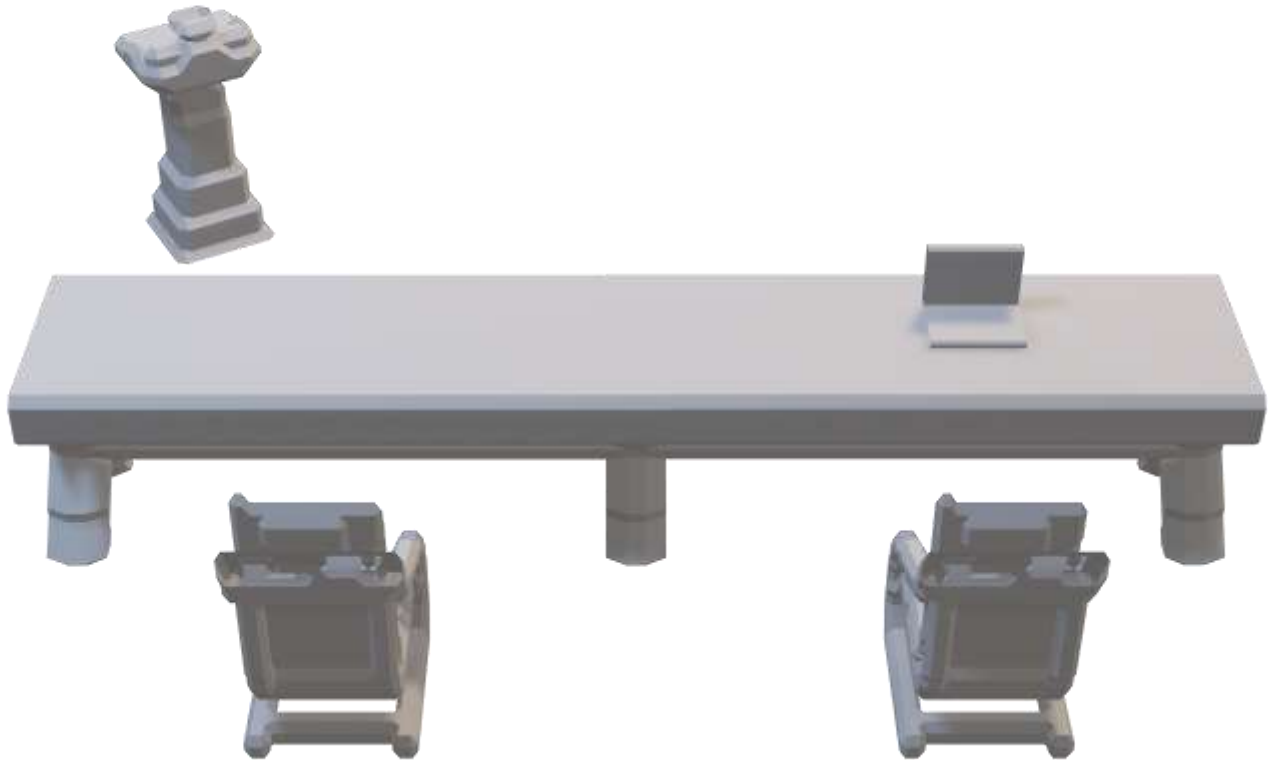
- Real-world Scene Captured with a ZED Camera

Multi-view 3D Scene Reconstruction

- Real-world Scene Captured with a ZED Camera

Multi-view 3D Scene Reconstruction

- Real-world 3D Scene Reconstruction



Thank You!