# Paper Reading

2022.06.23

Guan Yunyi

# Geometry-aware Recurrent Neural Networks for Active Visual Recognition

Cheng R, Wang Z, Fragkiadaki K.

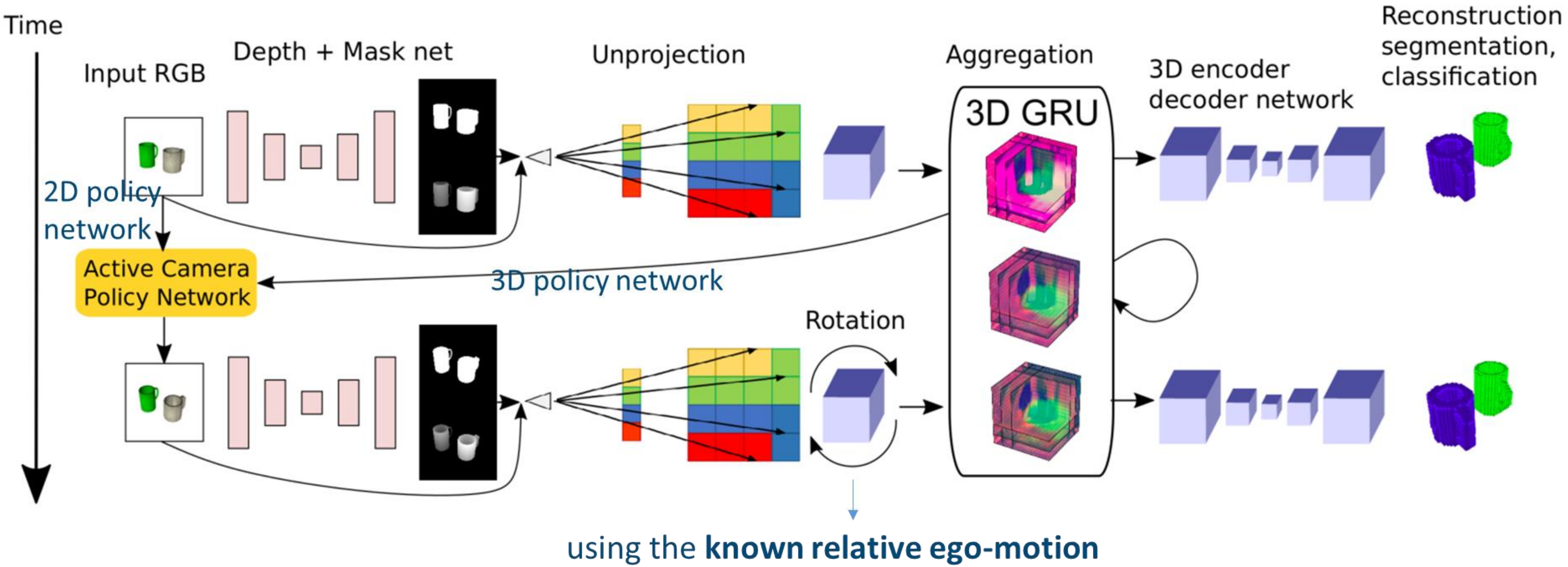
Key words: active vision, 3D reconstruction, RL

01

# Introduction

# 1. Introduction

- **Geometry-aware RNN:** trained end-to-end in a differentiable manner
- **Active view selection network:** trained with reinforcement rewarded by state estimation accuracy at each timestep.

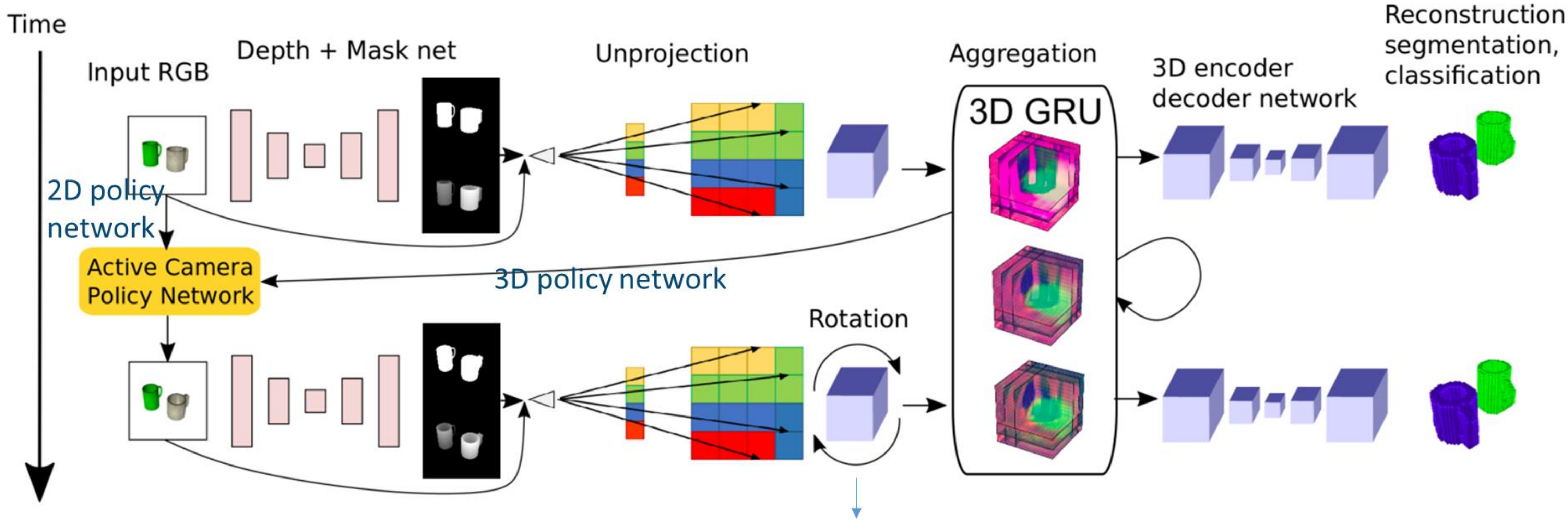


02

# Methods

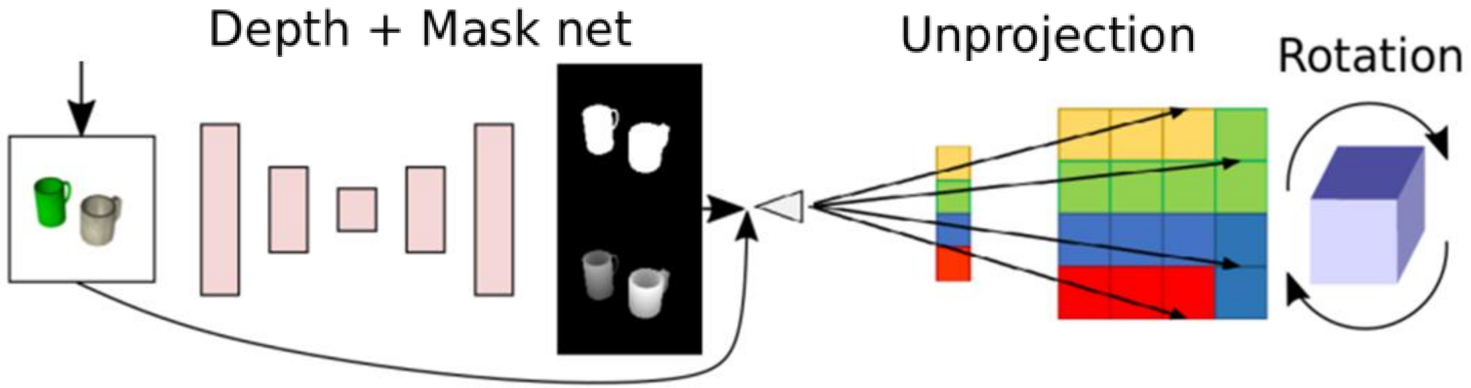
## 2.1 Geometry-aware RNN

- Integrate scene information across views into **3D latent feature tensors**
  -> information regarding the same 3D physical point is placed nearby in the tensor
- Tasks are **directly solved later from the output of the 3D latent feature**



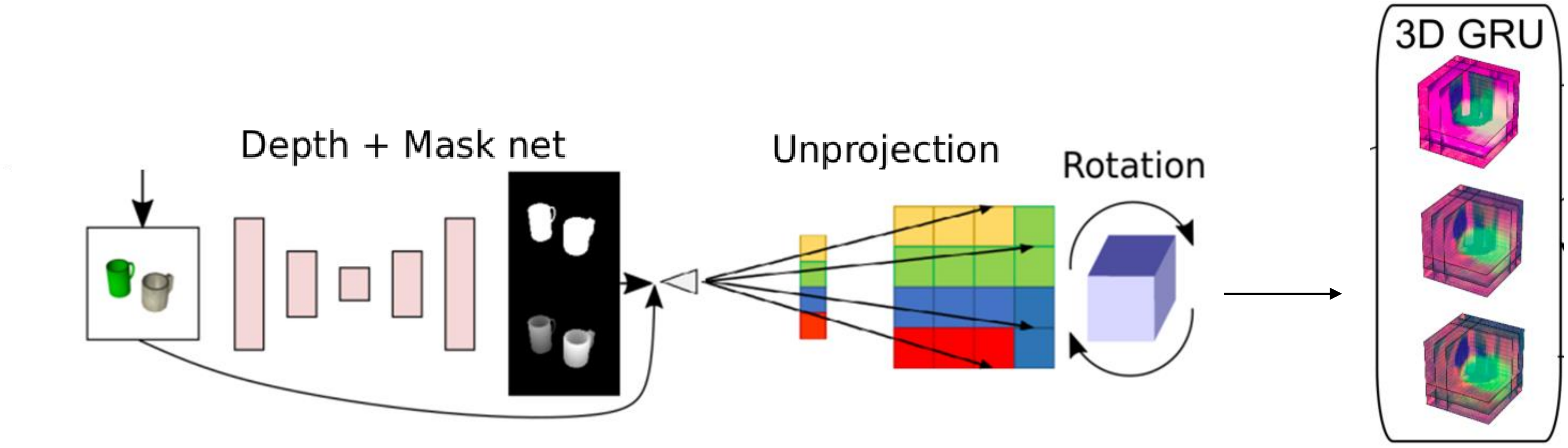

using the **known relative ego-motion**

# 2.1 Geometry-aware RNN – (1) Unprojection




- **Input**: an RGB image of the selected viewpoint
- **Depth + Mask net:** 2D convolutional encoder-decoder network, predicts 2D depth and object foreground map
- **Unprojection**: 2D -> 3D, use RGB, depth and mask information to fill in z-axis (All voxels along the same ray shot will be filled with nearly the same z)
- **Output**: an initial feature tensor

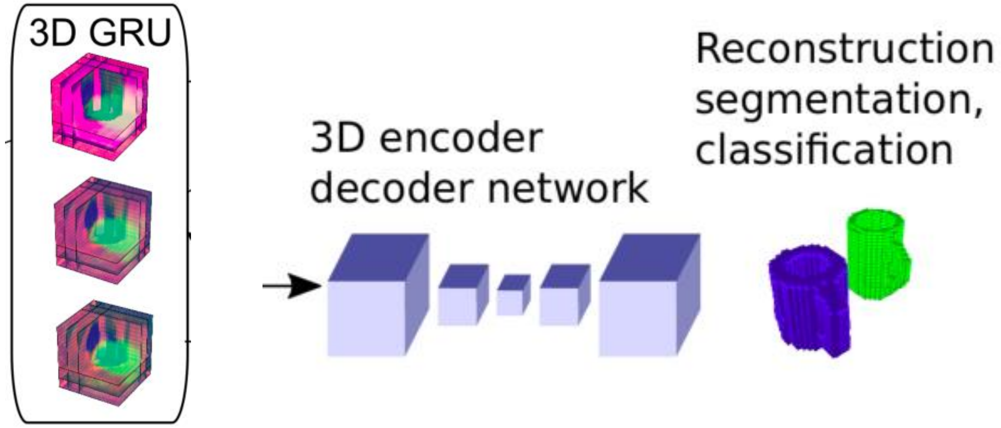
## 2.1 Geometry-aware RNN – (2) Recurrent memory update




- **Rotation:** rotate unprojected feature to the first view using known egomotion

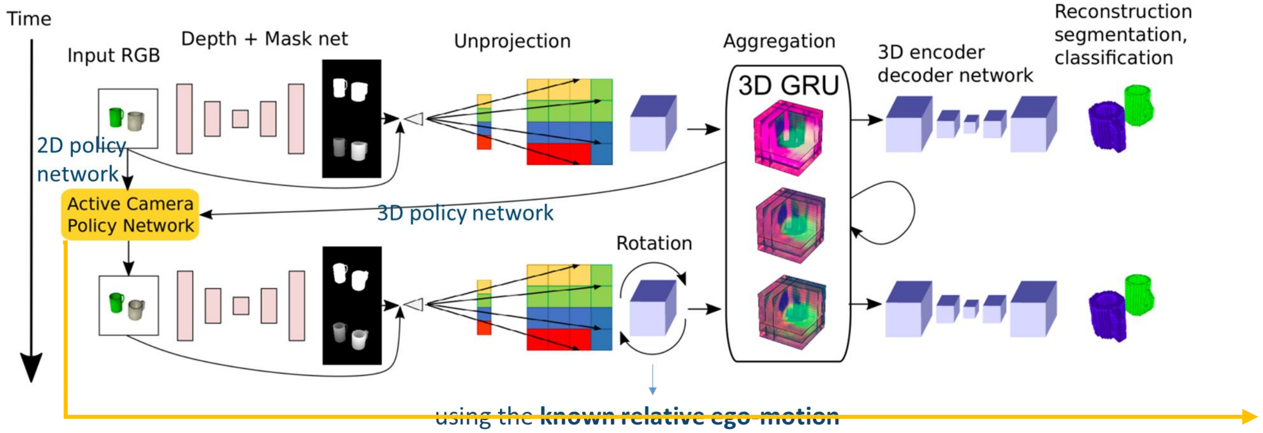
  -> 2D projections regarding the same physical point are placed nearby in the 3D memory tensor
- **3D GRU**: input the rotated features to update memory

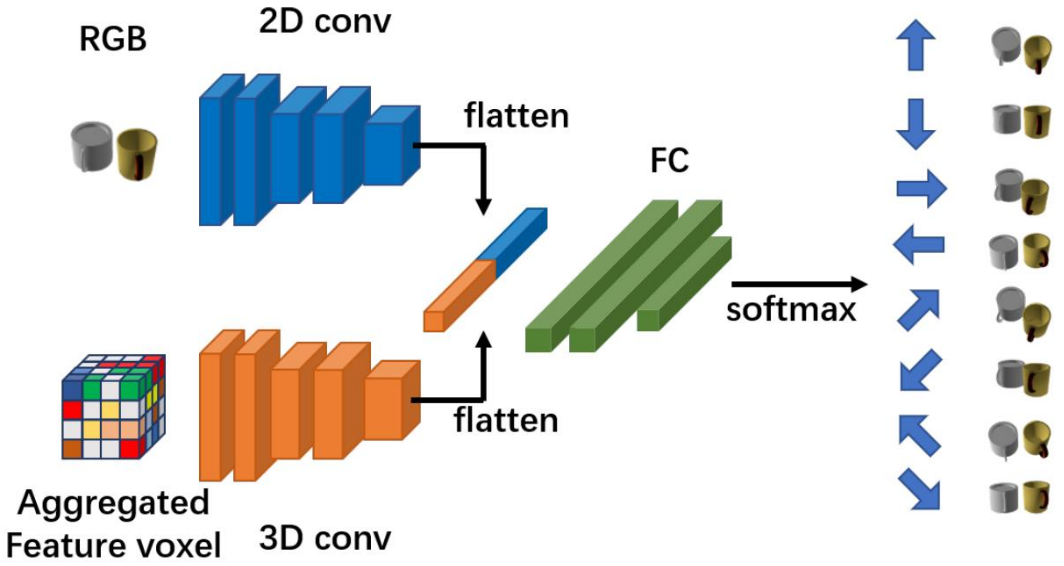
## 2.1 Geometry-aware RNN – (3) 3D tasks




- **Supervised training:** using 3D occupancy voxel grids, ground-truth bounding boxes and masks available in simulator environments
- **3D convolutional encoder-decoder network:** produce the final set of outputs
- 3D sigmoid output which predicts voxel occupancy
- 3D segmentation embedding feature
- multiclass softmax output at every voxel

# 2.2 View selection policy




- at each t, predict a distribution over **eight adjacent views in the neighborhood of the current view**
- **Policy network:** CNN with 2D and 3D branches
- Output: final categorical distribution over 8 possible directions
- Training: with REINFORCE
- Rewards: reconstruction-driven

(as Intersection over Union (IoU) of the discretized voxel occupancy from each view to the next increases)

03

# Experiments

# 3.1 Multi-view reconstruction of single objects

| | single object | | | |
|---|---|---|---|---|
| | view-1 | view-2 | view-3 | view-4 |
| 1D-LSTM | 0.57 | 0.59 | 0.60 | 0.60 |
| LSM | 0.63 | 0.66 | 0.68 | 0.69 |
| LSM+gt depth | **0.65** | 0.68 | 0.69 | 0.70 |
| ours+gt depth | 0.55 | **0.69** | **0.72** | **0.73** |

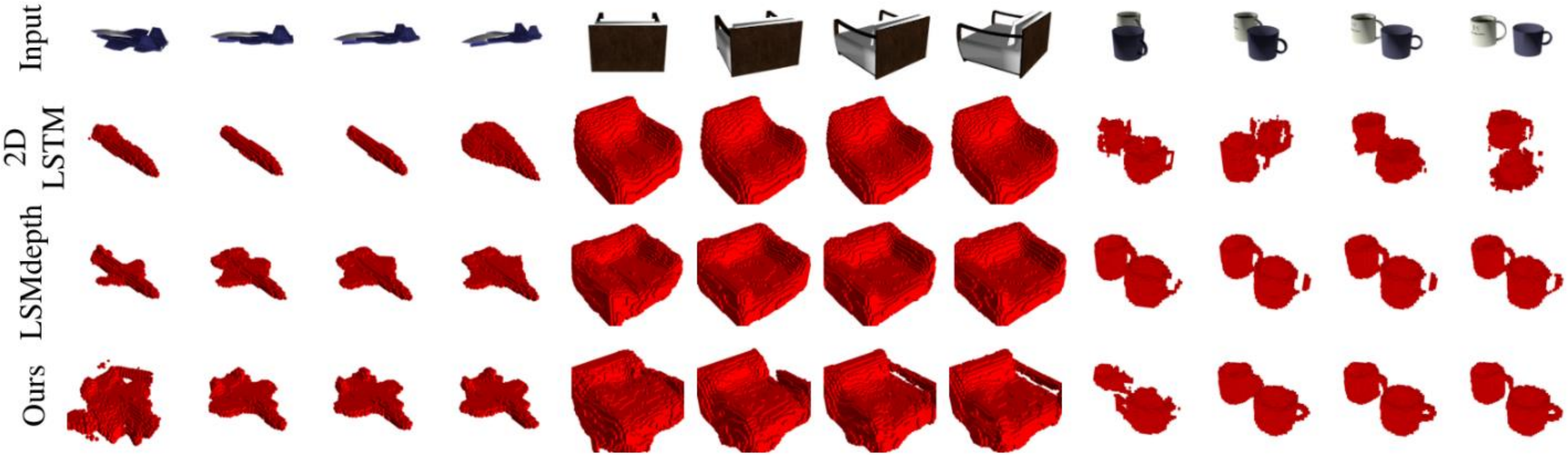
| | view-1 | view-2 | view-3 | view-4 |
|---|---|---|---|---|
| 1D-LSTM | 0.12 | 0.16 | 0.16 | 0.18 |
| ours | 0.24 | 0.28 | 0.31 | 0.32 |

IoU between the prediction and ground-truth 3D voxel grids

- **Dataset**: `SUNCG` and chairs, cars, and airplanes from `ShapeNet`
- Train a single 3D reconstruction model with ground-truth 3D voxel occupancy (on sequences of randomly selected views)

-> the proposed geometry-aware RNN outperforms the baselines, especially after aggregating information from more views.

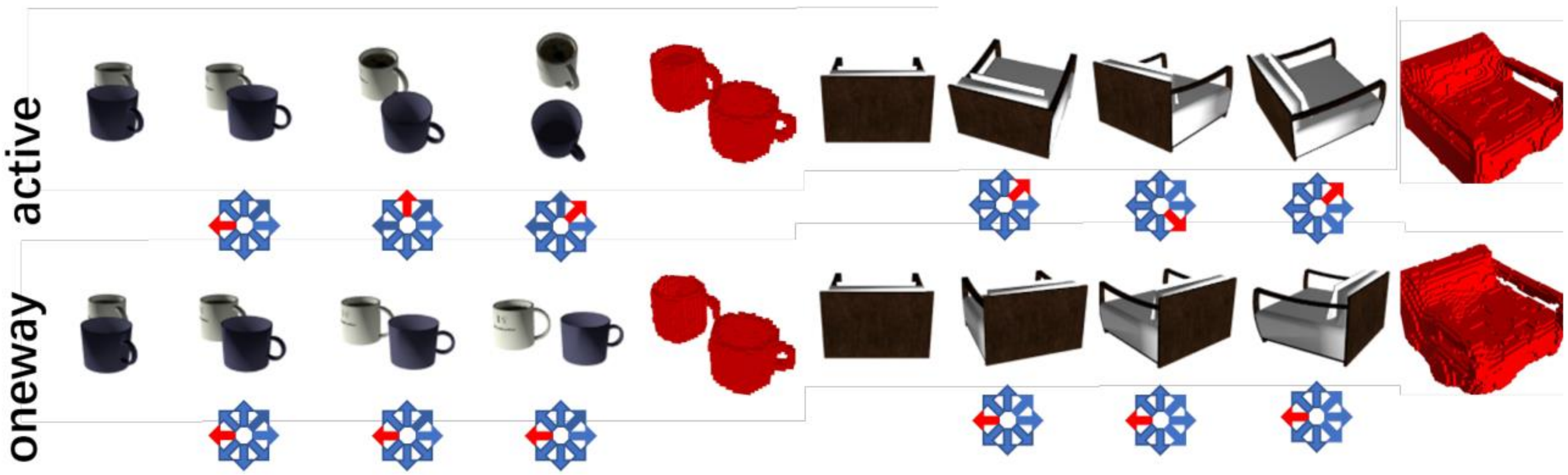
## 3.1 Multi-view reconstruction of single objects



# 3.2 Active view selection

| | view-1 | view-2 | view-3 | view-4 |
|---|---|---|---|---|
| | single object | | | |
| 1-step greedy | 0% | 29.1% | 32.7% | 33.3% |
| Active | 0% | 27.7%±3.23% | 27.7%±3.23% | 36.6%±1.17% |
| Oracle | 0% | 32.7% | 38.1% | 38.1% |
| | multiple objects | | | |
| 1-step greedy | 0% | 36.2% | 40.4% | 40.4% |
| Active | 0% | 31.3%±2.26% | 39.0%±1.99% | 41.8%±2.36% |
| Oracle | 0% | 29.8% | 42.6% | 44.7% |

Percent increase in IoU between 3D reconstructions and ground-truth over the single-view reconstruction



## 3.3 multi-object tasks

| | multi-objects | | | |
|---|---|---|---|---|
| | view-1 | view-2 | view-3 | view-4 |
| 1D-LSTM | 0.11 | 0.15 | 0.17 | 0.20 |
| LSM | 0.43 | 0.47 | 0.51 | 0.53 |
| LSM+gt depth | **0.48** | 0.51 | 0.54 | 0.56 |
| ours+gt depth | 0.47 | **0.58** | **0.62** | **0.64** |
| ours+learnt depth | 0.45 | 0.56 | 0.60 | 0.62 |

IoU between the prediction and ground-truth 3D voxel grids

| | view-1 | view-2 | view-3 | view-4 |
|---|---|---|---|---|
| 3D voxel occupancy IoU | 0.54 | 0.63 | 0.64 | 0.65 |
| 3D segmentation IoU | 0.60 | 0.69 | 0.70 | 0.71 |
| Classification accuracy | 0.56 | 0.83 | 0.83 | 0.83 |

- For 3D reconstruction, learned depth results in lower IoU than ground-truth depth

04

# Conclusion

## 4.1 Innovation Points

- Selecting views for **jointly optimizing** 3D reconstruction, object instance segmentation, and classification
- Proposing a geometry-aware RNN that accumulates feature information **directly in 3D**

## 4.2 Limitations

- it assumes ground-truth ego-motion
- it consumes lots of GPU memory for maintaining the full scene tensor
- it cannot handle moving objects
- it requires 3D ground-truth for object detection, which is very expensive to obtain
- My opinion:

- focus on active 3D reconstruction but not active recognition
- can only predict eight adjacent views in the neighborhood of the current view