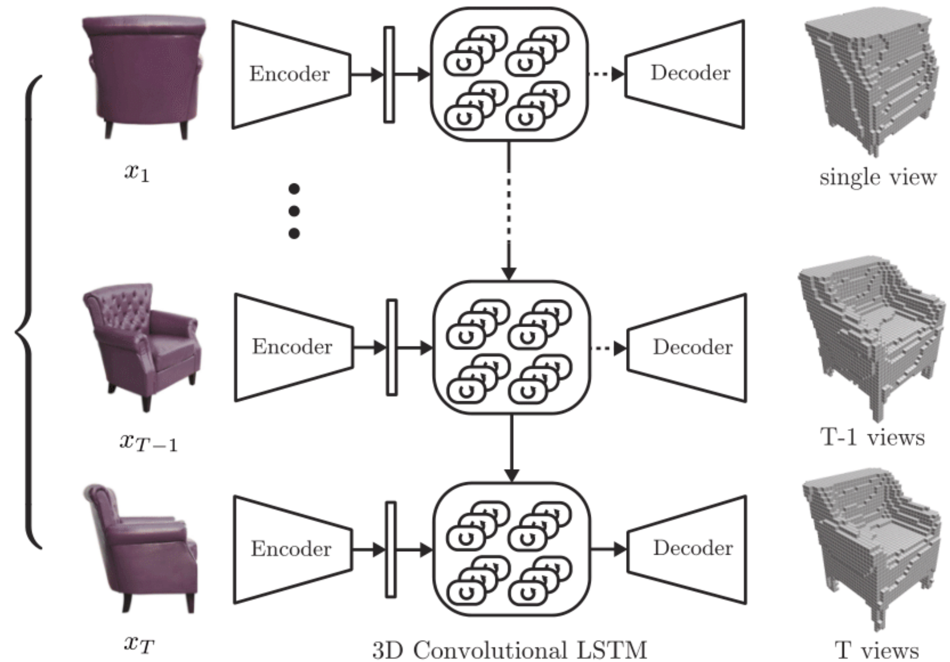
[ECCV 2016] Stanford University

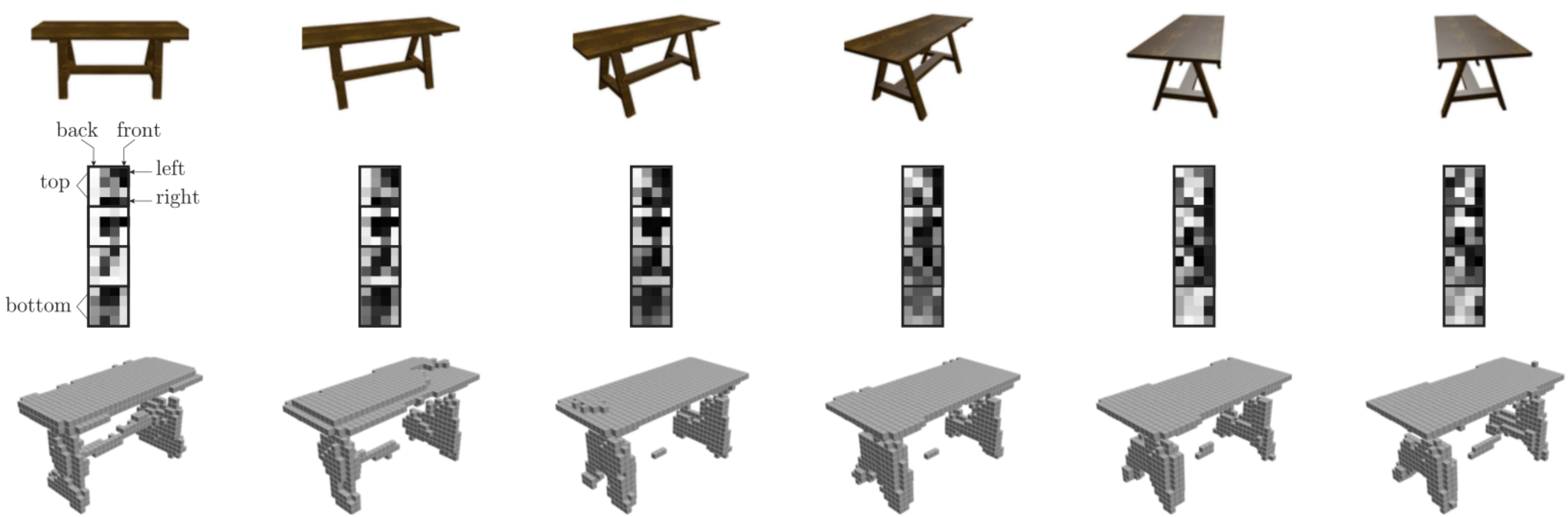
# 3D-R2N2: A Unified Approach for Single and Multi-view 3D Object Reconstruction

Abstract.

- Inspired by the recent success of methods that employ shape priors to achieve robust 3D reconstructions, we propose a novel recurrent neural network architecture that we call the 3D Recurrent Reconstruction Neural Network (3D-R2N2). The network learns a mapping from images of objects to their underlying 3D shapes from a large collection of synthetic data [13]. Our network takes in one or more images of an object instance from arbitrary viewpoints and outputs a reconstruction of the object in the form of a 3D occupancy grid. Unlike most of the previous works, our network does not require any image annotations or object class labels for training or testing. Our extensive experimental analysis shows that our reconstruction framework i) outperforms the state-of-the-art methods for single view reconstruction, and ii) enables the 3D reconstruction of objects in situations when traditional SFM/SLAM methods fail (because of lack of texture and/or wide baseline).

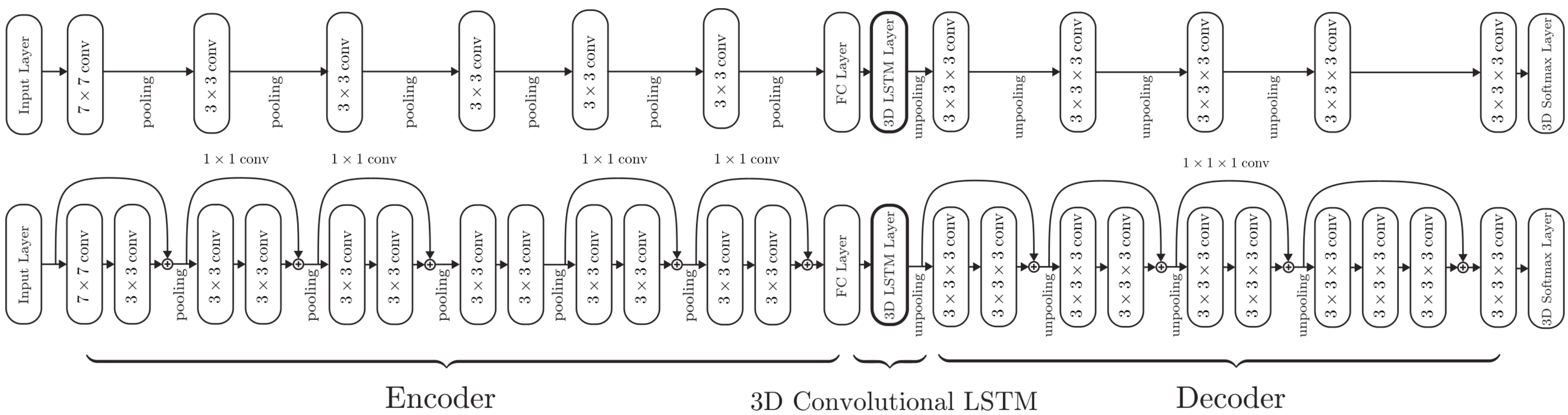


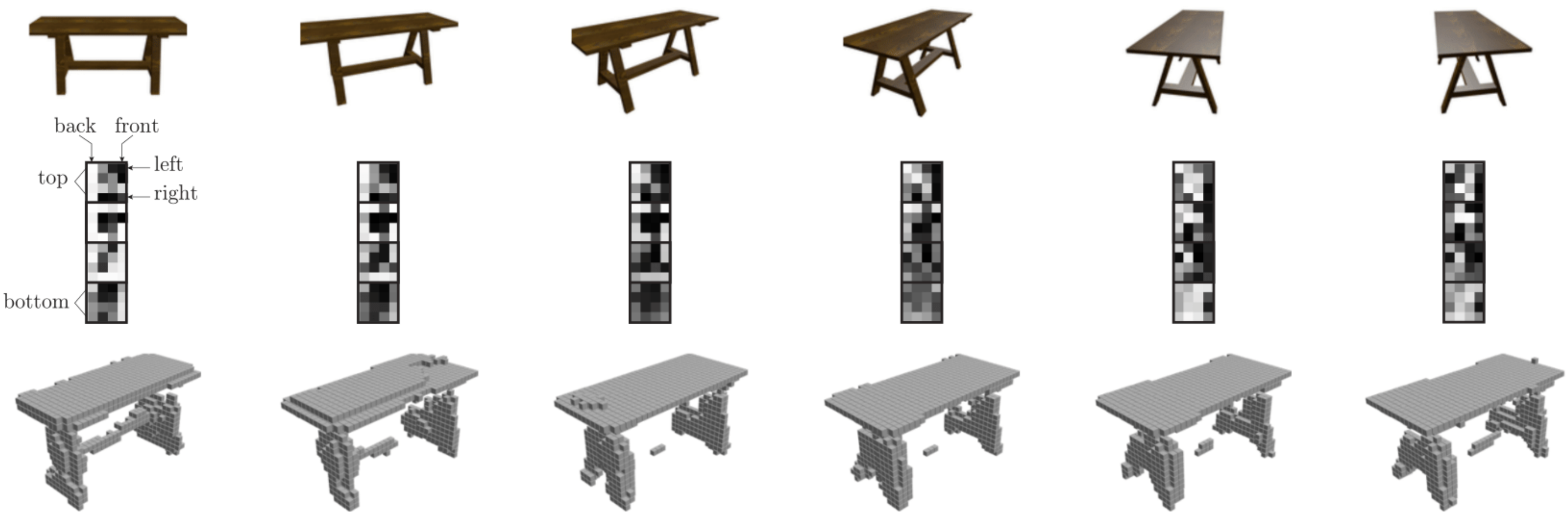
Encoder
Decoder
$x_1$
single view
Encoder
Decoder
$x_{T-1}$
T-1 views
Encoder
Decoder
$x_T$
3D Convolutional LSTM
T views

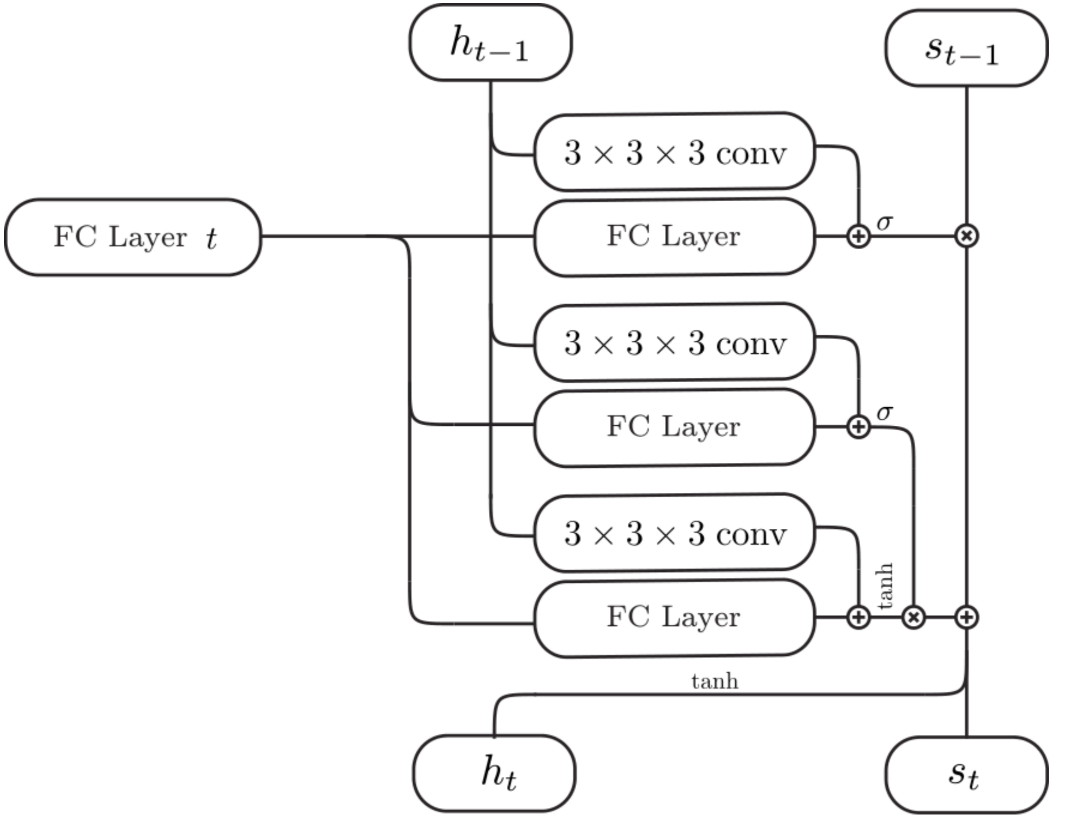



back
front
left
top
right
bottom

# Network Architecture

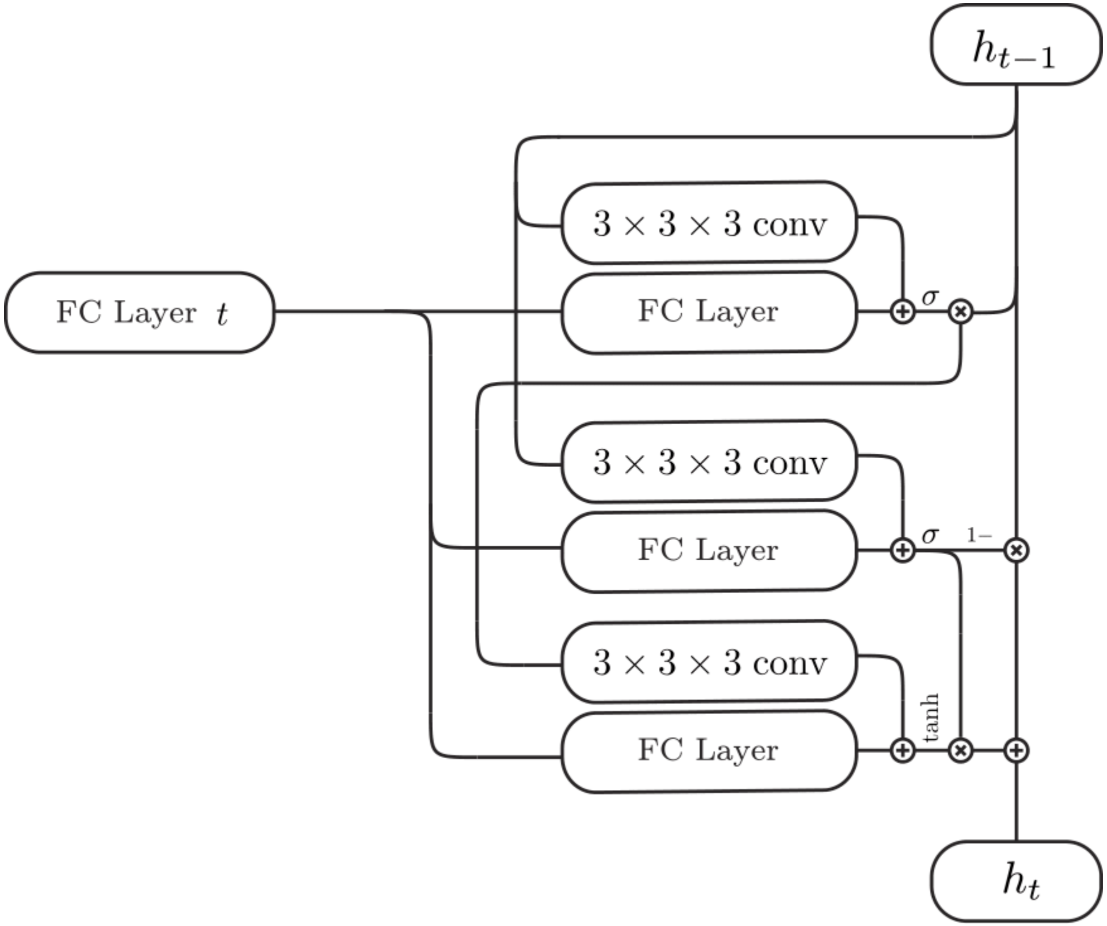



back
front
left
top
right
bottom




3D Convolutional LSTM

$$f_t = \sigma(W_f \mathcal{T}(x_t) + U_f * h_{t-1} + b_f)$$
$$i_t = \sigma(W_i \mathcal{T}(x_t) + U_i * h_{t-1} + b_i)$$
$$s_t = f_t \odot s_{t-1} + i_t \odot \tanh(W_s \mathcal{T}(x_t) + U_s * h_{t-1} + b_s)$$
$$h_t = \tanh(s_t)$$



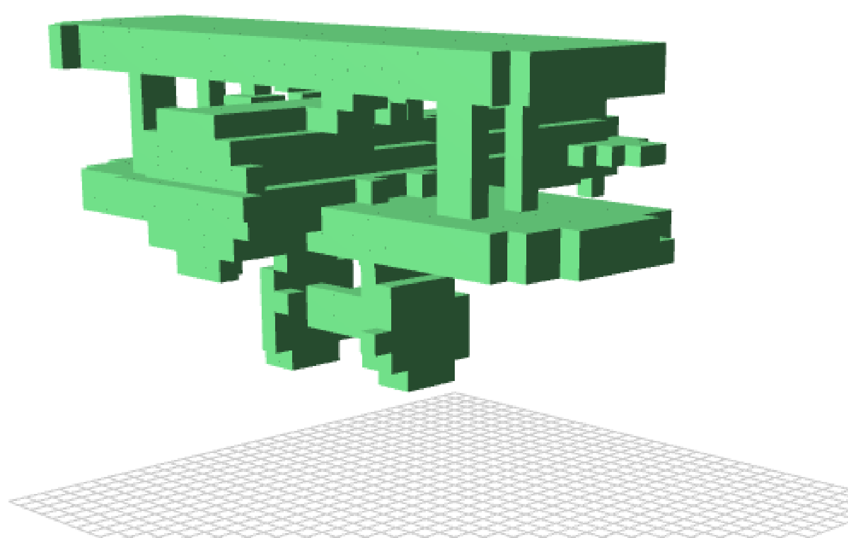

3D Convolutional GRU

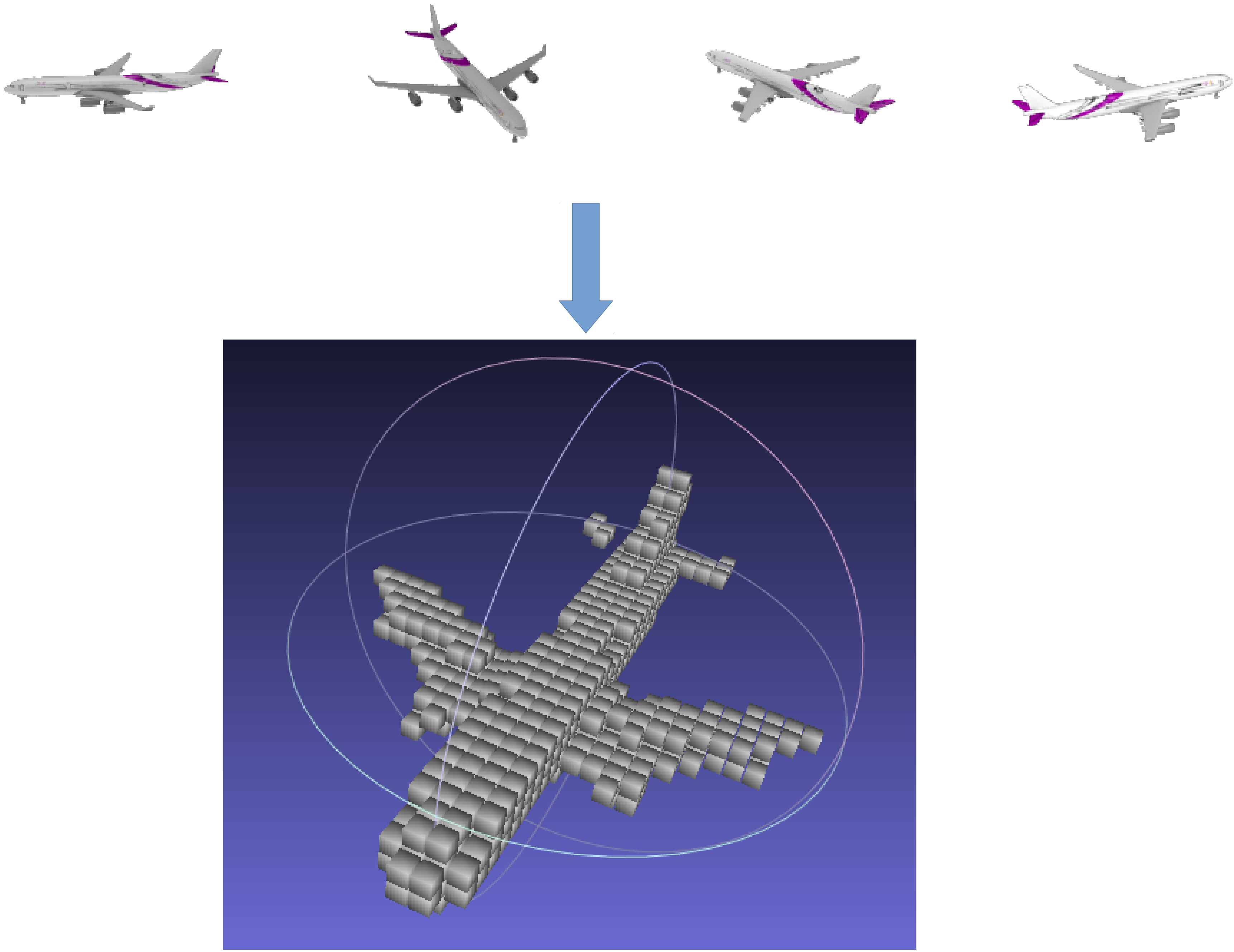
$$u_t = \sigma(W_{fx}\mathcal{T}(x_t) + U_f * h_{t-1} + b_f)$$
$$r_t = \sigma(W_{ix}\mathcal{T}(x_t) + U_i * h_{t-1} + b_i)$$
$$h_t = (1 - u_t) \odot h_{t-1} + u_t \odot \tanh(W_h\mathcal{T}(x_t) + U_h * (r_t \odot h_{t-1}) + b_h)$$

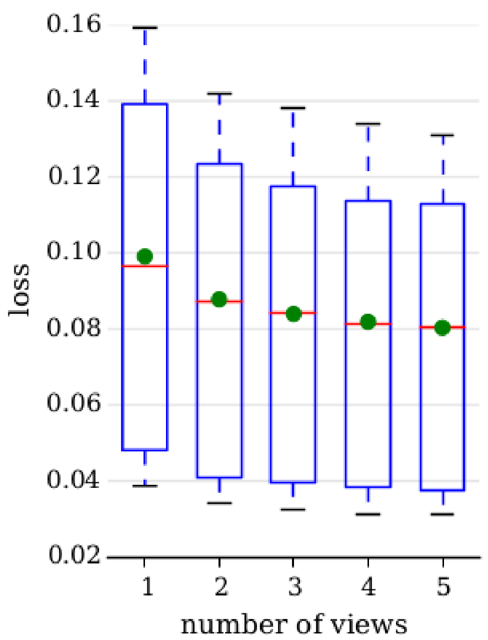
# Dataset: ShapeNet(Stanford)

Input(127*127):

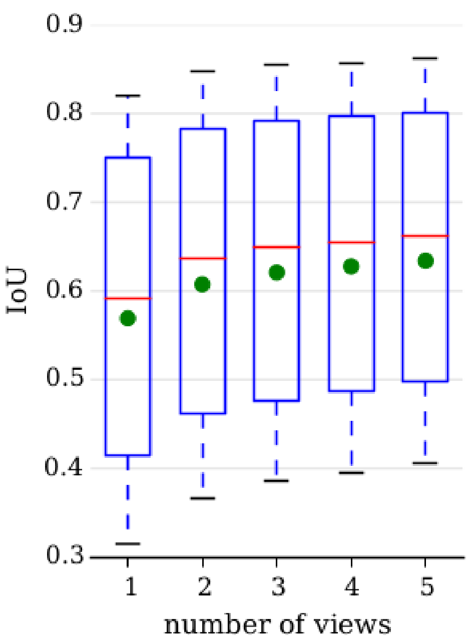
Ground truth(32*32*32):








(a) Cross entropy loss

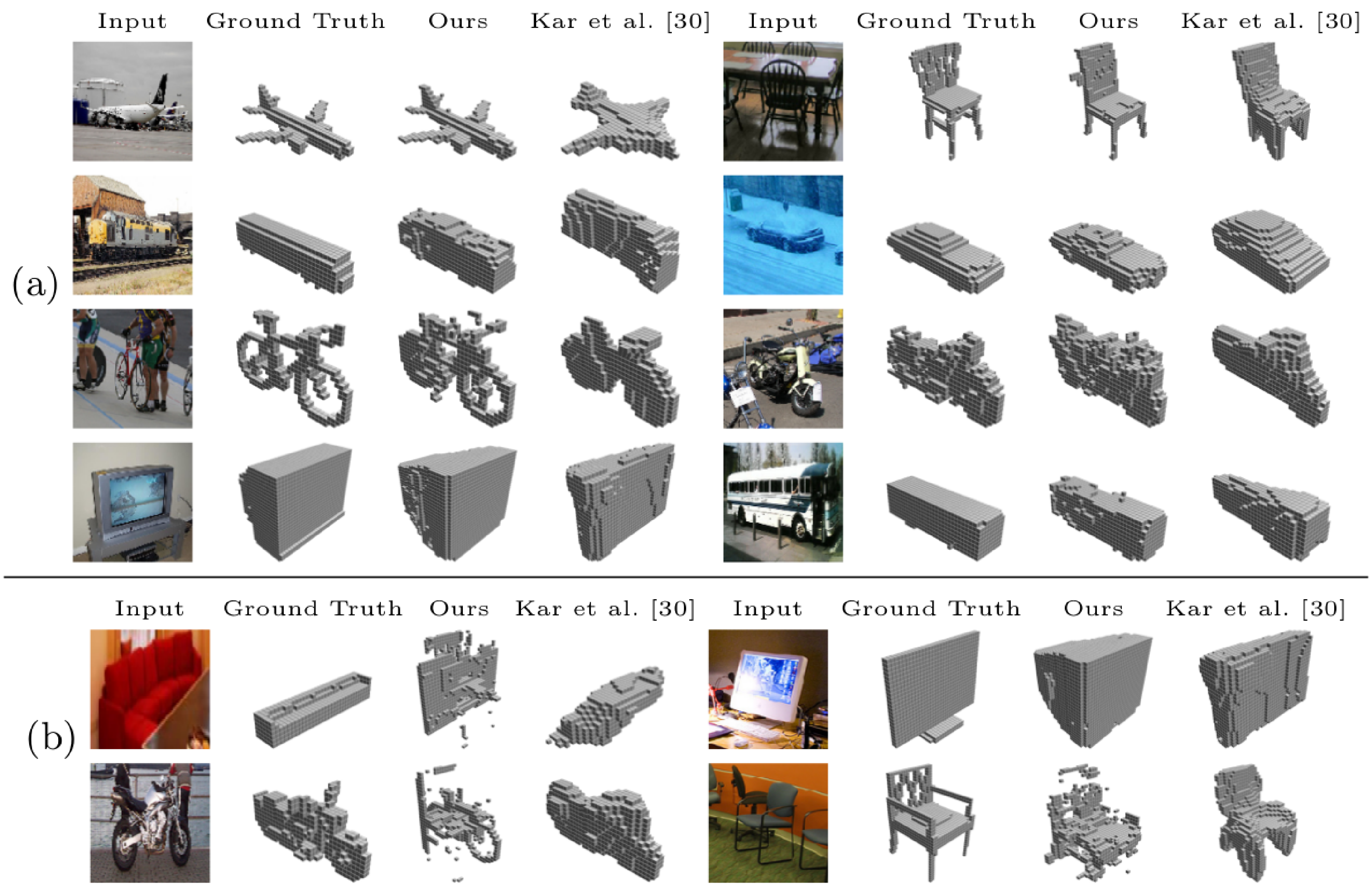



(b) Voxel IoU

| # views | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| plane | 0.513 | 0.536 | 0.549 | 0.556 | 0.561 |
| bench | 0.421 | 0.484 | 0.502 | 0.516 | 0.527 |
| cabinet | 0.716 | 0.746 | 0.763 | 0.767 | 0.772 |
| car | 0.798 | 0.821 | 0.829 | 0.833 | 0.836 |
| chair | 0.466 | 0.515 | 0.533 | 0.541 | 0.550 |
| monitor | 0.468 | 0.527 | 0.545 | 0.558 | 0.565 |
| lamp | 0.381 | 0.406 | 0.415 | 0.416 | 0.421 |
| speaker | 0.662 | 0.696 | 0.708 | 0.714 | 0.717 |
| firearm | 0.544 | 0.582 | 0.593 | 0.595 | 0.600 |
| couch | 0.628 | 0.677 | 0.690 | 0.698 | 0.706 |
| table | 0.513 | 0.550 | 0.564 | 0.573 | 0.580 |
| cellphone | 0.661 | 0.717 | 0.732 | 0.738 | 0.754 |
| watercraft | 0.513 | 0.576 | 0.596 | 0.604 | 0.610 |

(c) Per-category IoU

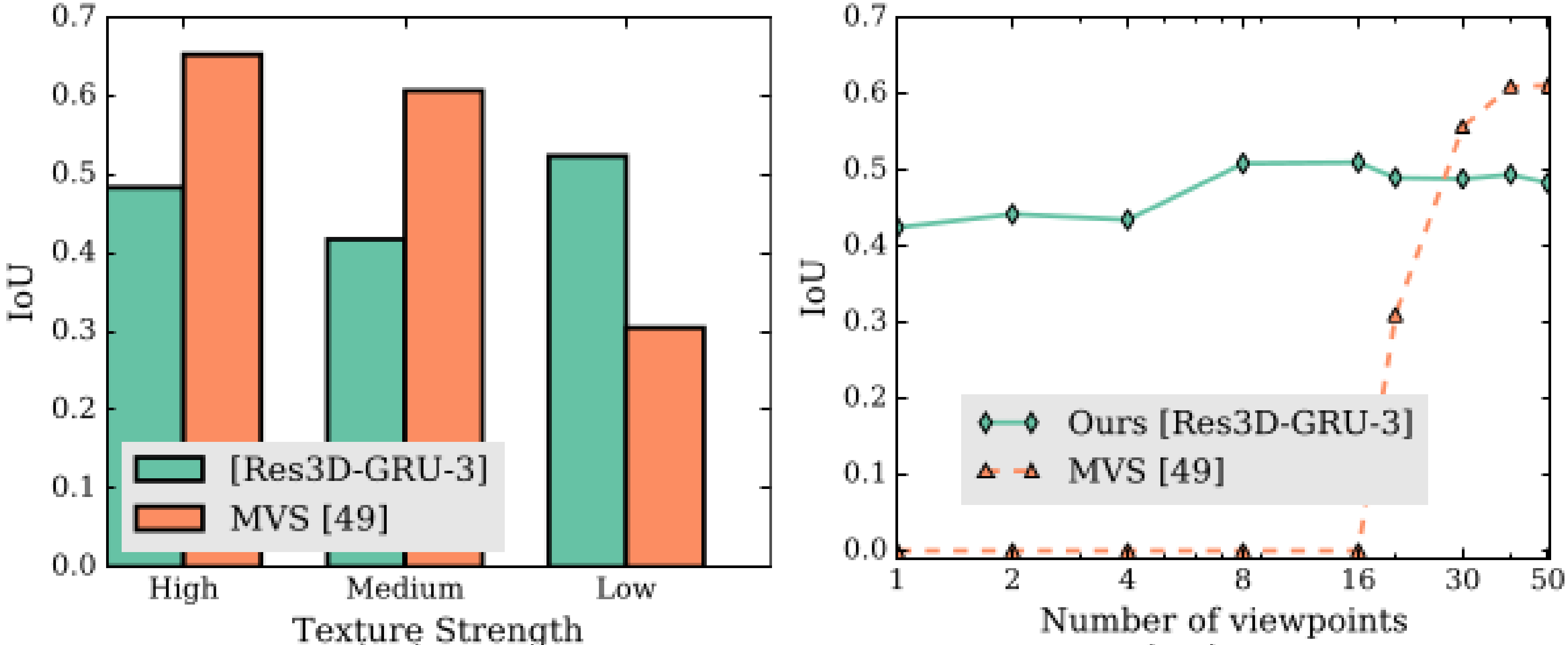
# Single Real-World Image Reconstruction



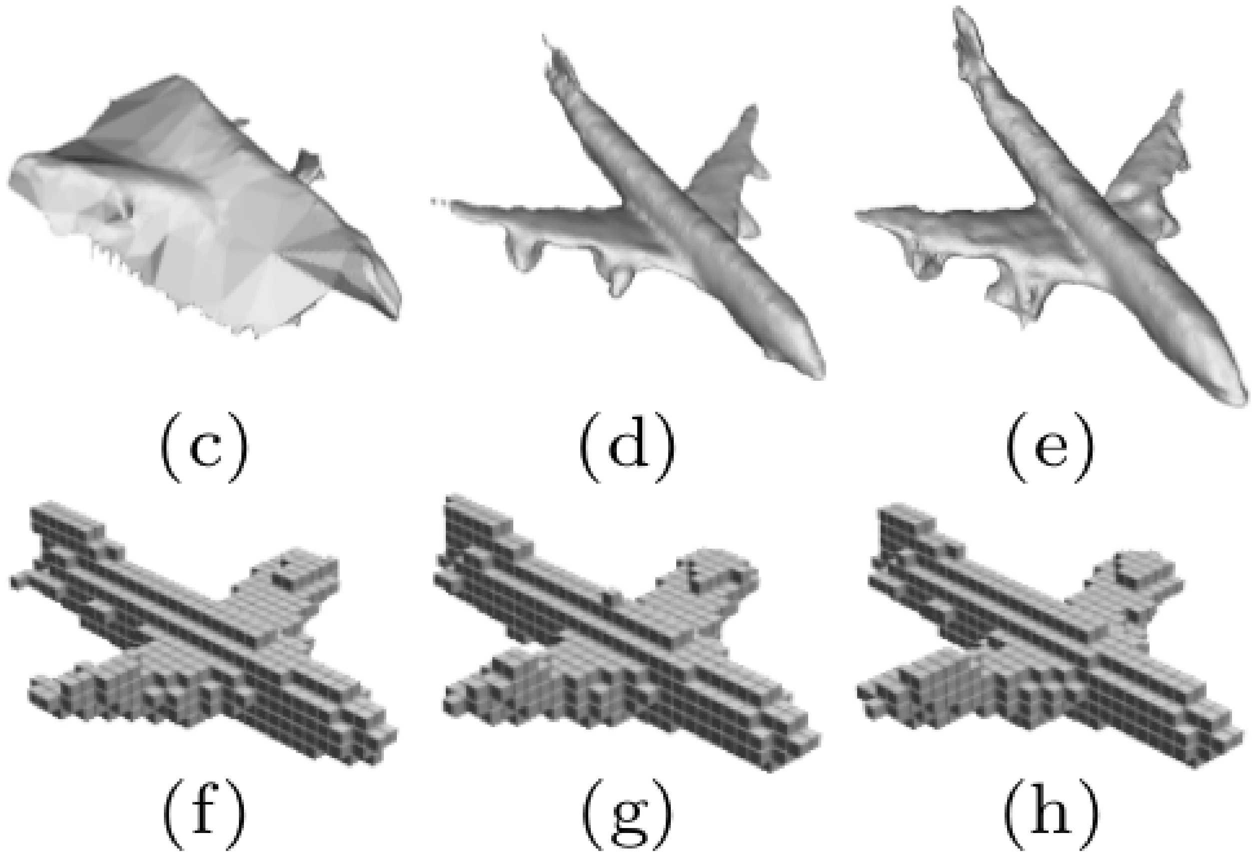
IoU

| | aero | bike | boat | bus | car | chair | mbike | sofa | train | tv | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Kar et al. [30] | 0.298 | 0.144 | 0.188 | 0.501 | 0.472 | 0.234 | 0.361 | 0.149 | 0.249 | 0.492 | 0.318 |
| ours [LSTM-1] | 0.472 | 0.330 | 0.466 | 0.677 | 0.579 | 0.203 | 0.474 | 0.251 | 0.518 | 0.438 | 0.456 |
| ours [Res3D-GRU-3] | **0.544** | **0.499** | **0.560** | **0.816** | **0.699** | **0.280** | **0.649** | **0.332** | **0.672** | **0.574** | **0.571** |

# Multi View Stereo(MVS) vs. 3D-R2N2




(a) (b)




20 views 30 views 40 views