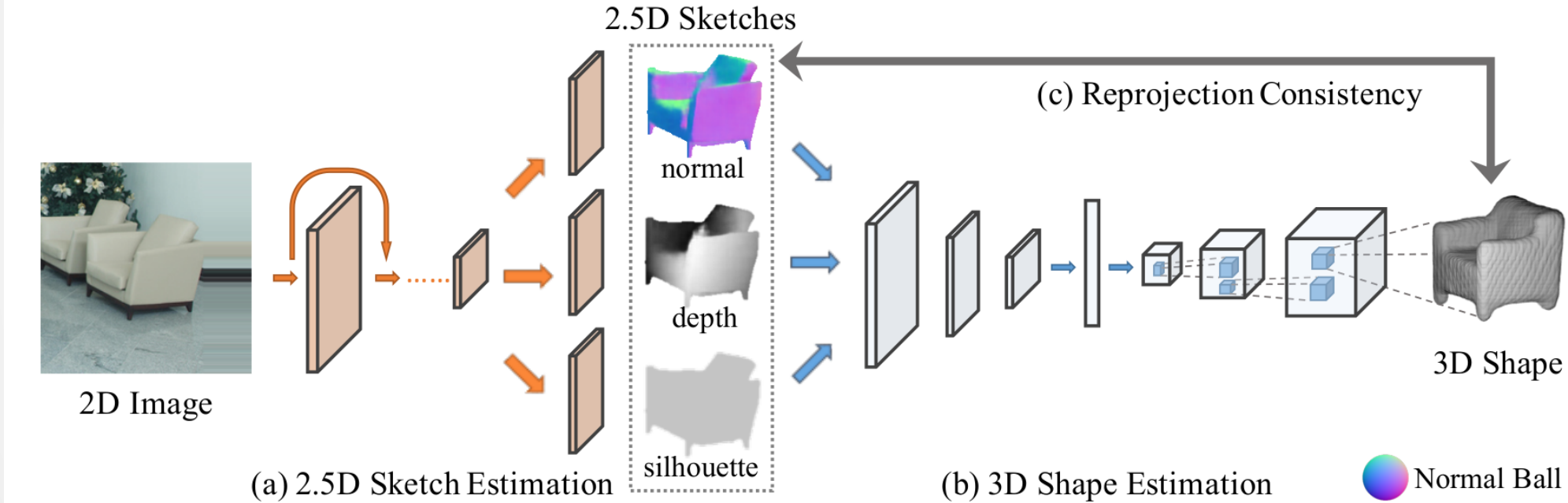
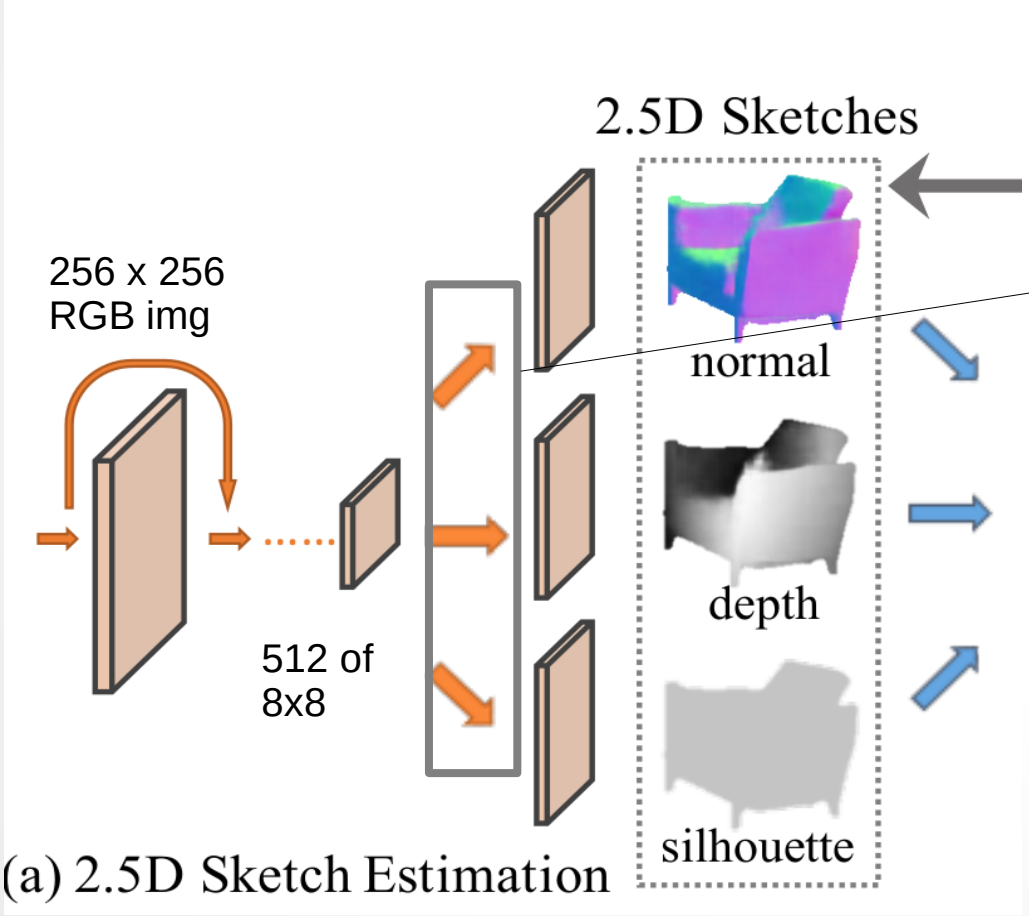
MarrNet: 3D Shape Reconstruction via 2.5D Sketches

- Authors :
  - Jiajun Wu (MIT CSAIL),
  - Yifan Wang (ShanghaiTech University),
  - Tianfan Xue (MIT CSAIL),
  - Xingyuan Sun (Shanghai Jiao Tong University),
  - William T. Freeman(MIT CSAIL, Google Research),
  - Joshua B. Tenenbaum (MIT CSAIL)
- NIPS 2017

Approach Overview

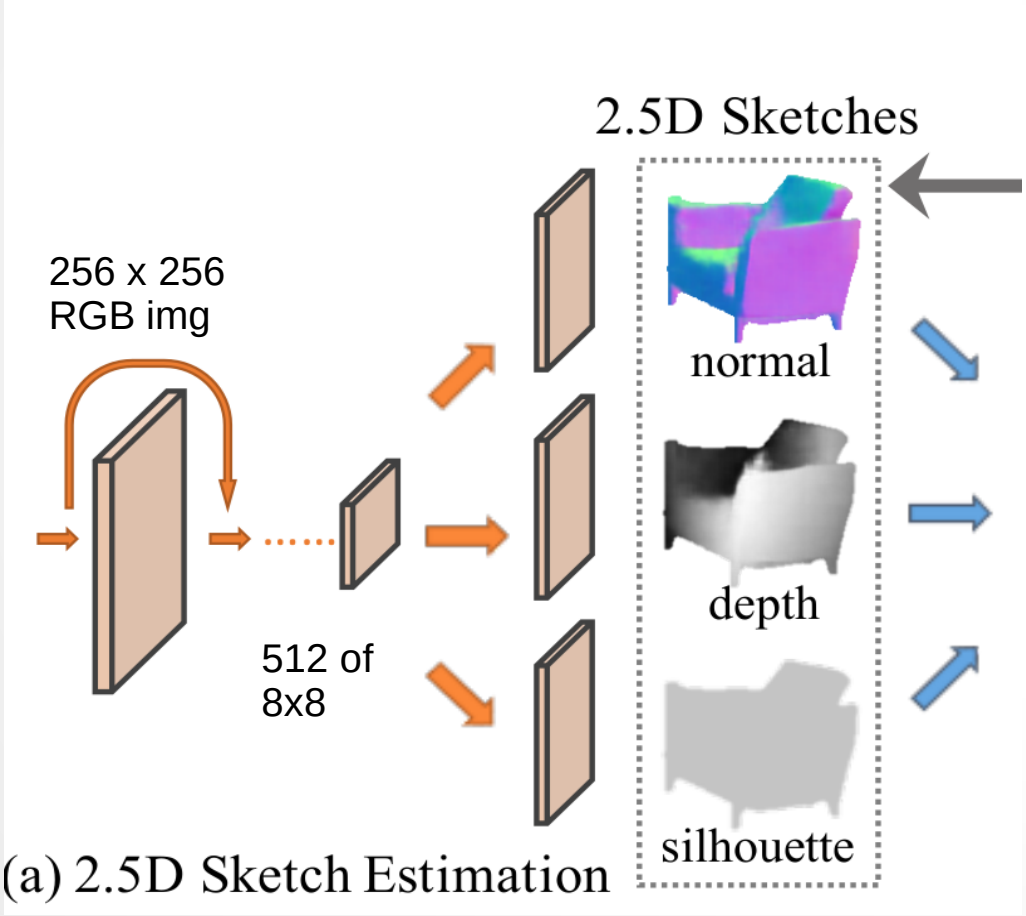


2.5D Sketch Estimation

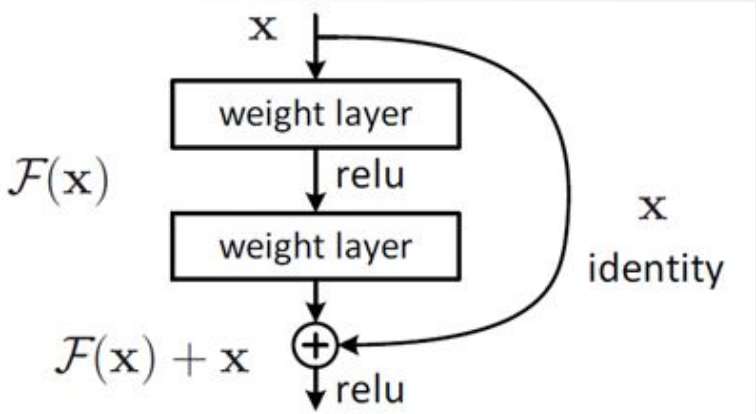


- ResNet-18 encoder-decoder network
- Decoder :
  - 5 x 5 fully convolutional and ReLU layers →
  - 1 x 1 convolutional and ReLU layers

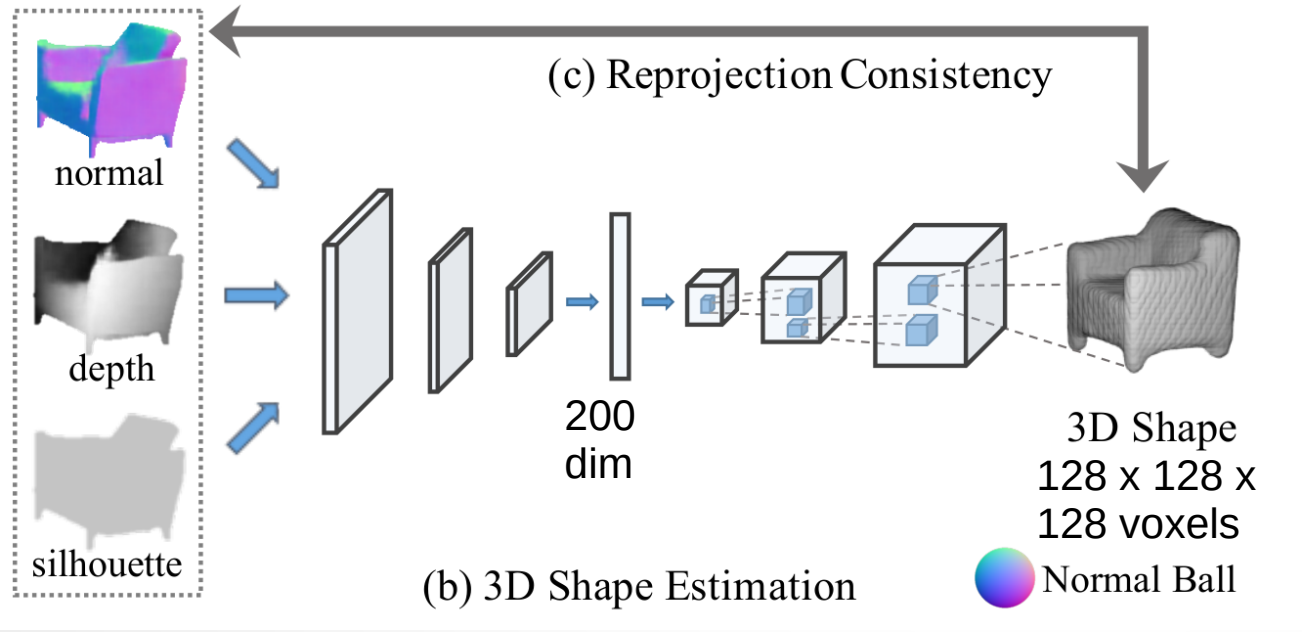
2.5D Sketch Estimation



ResNet-18 encoder-decoder network

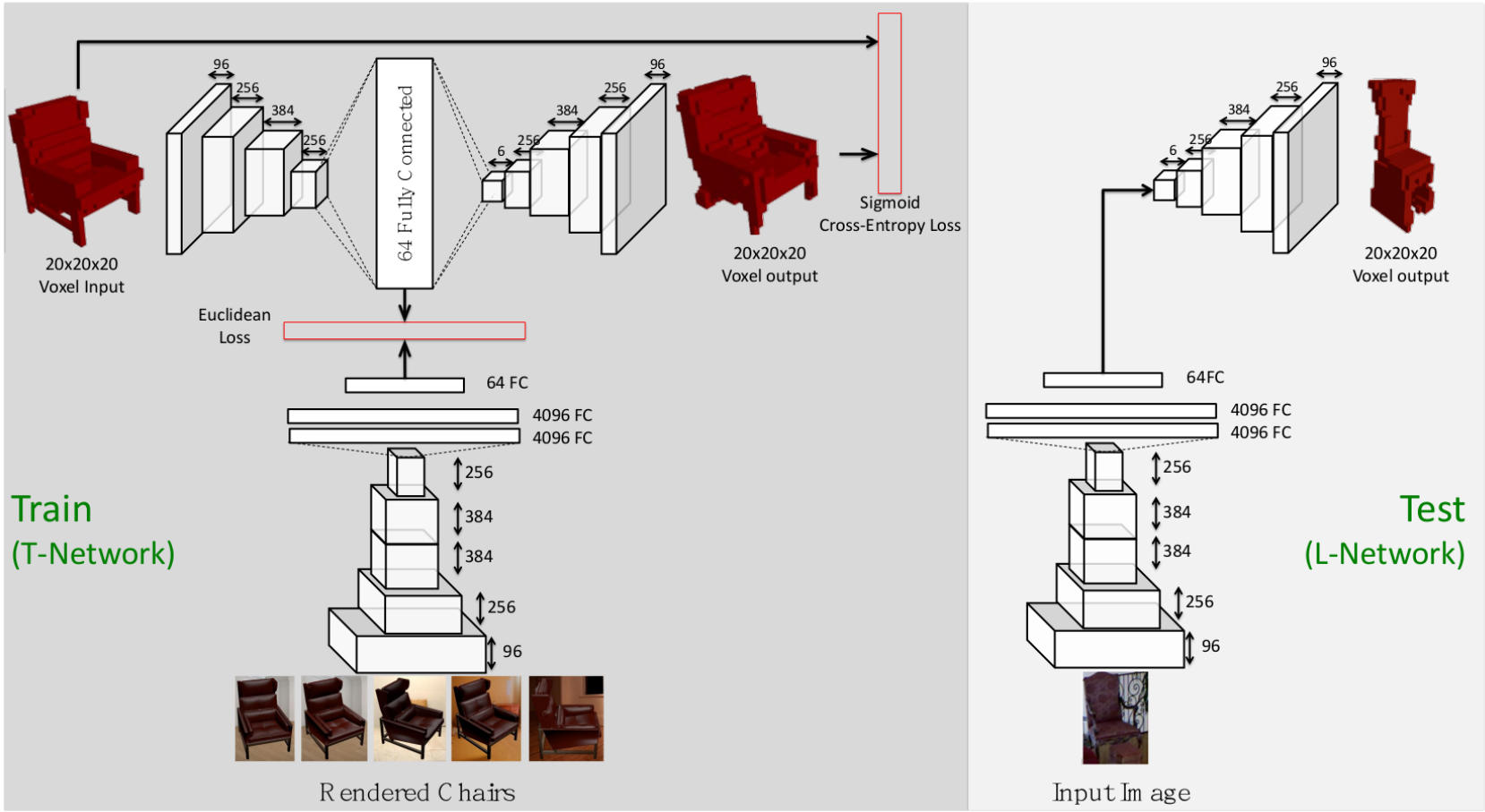


3D Shape Estimation



- Encoder-Decoder style
- References :
- TL Network
  - Learning a predictable and generative vector representation for objects. (ECCV, 2016)
- 3D GAN
  - Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling (NIPS 2016)

Architecture of TL-Network



Architecture of 3D GAN

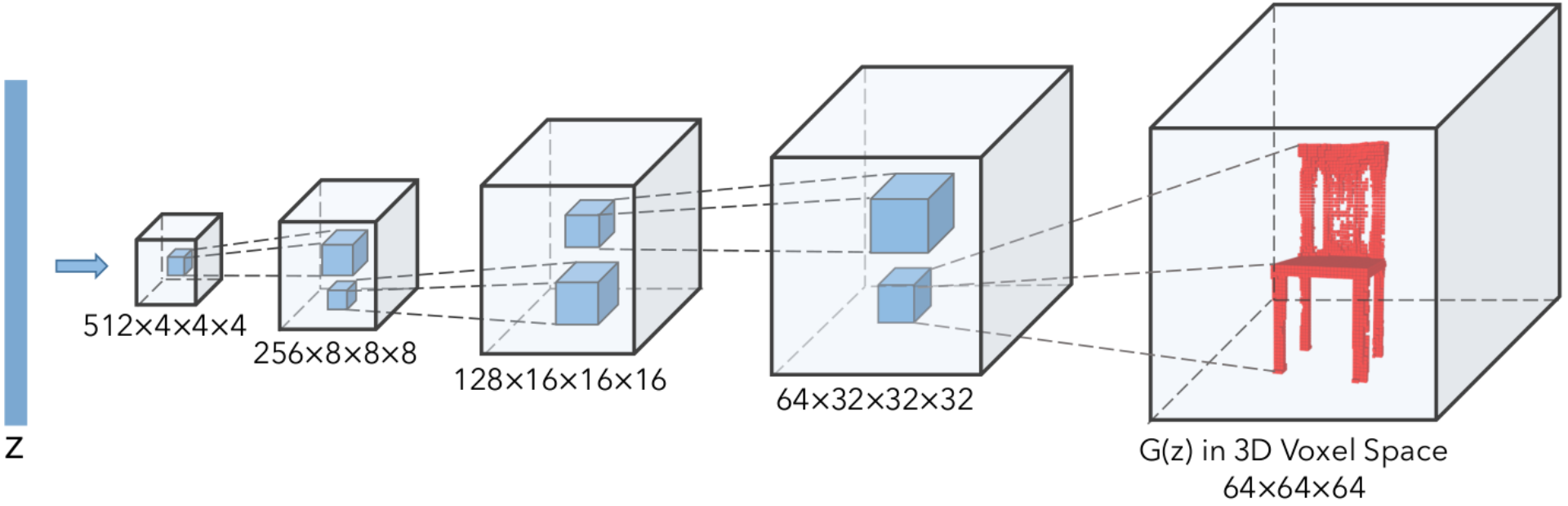
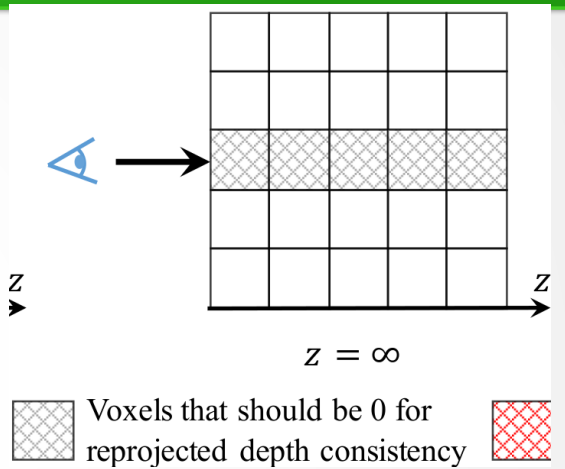


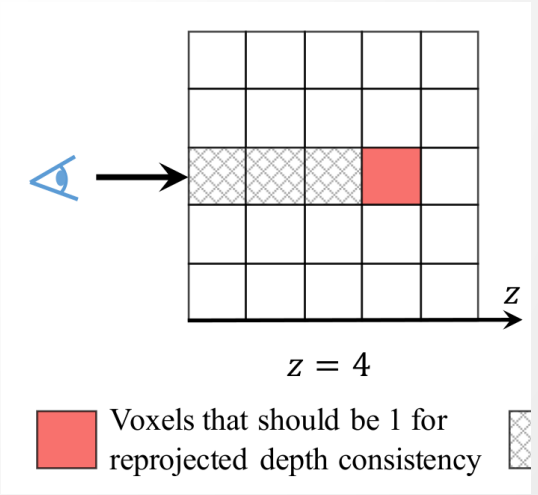
Figure 1: The generator in 3D-GAN. The discriminator mostly mirrors the generator.

Reprojection Consistency : Depth Reprojection loss

$v_{x,y,z} \in [0, 1], \forall x, y, z$

$$L_{\text{depth}}(x, y, z) = \begin{cases} v_{x,y,z}^2, & z < d_{x,y} \\ (1 - v_{x,y,z})^2, & z = d_{x,y} \\ 0, & z > d_{x,y} \end{cases}$$





Gradients :

$$\frac{\partial L_{\text{depth}}(x, y, z)}{\partial v_{x,y,z}} = \begin{cases} 2v_{x,y,z}, & z < d_{x,y} \\ 2(v_{x,y,z} - 1), & z = d_{x,y} \\ 0, & z > d_{x,y} \end{cases}$$

Reprojection Consistency : Normal Reprojection loss

$$n_x = (0, -n_c, n_b)$$

$$n_y = (-n_c, 0, n_a)$$

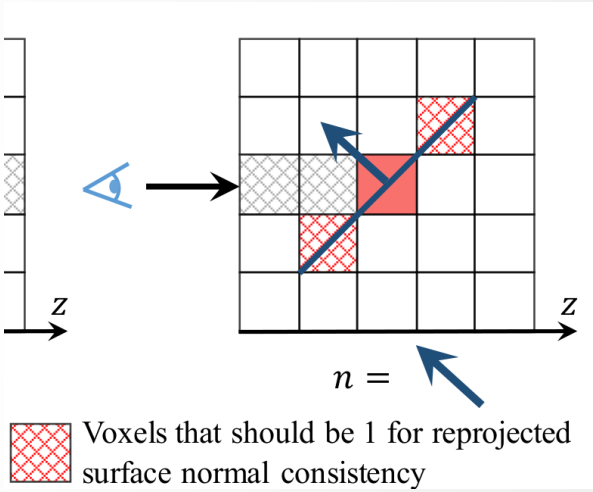
Orthogonal

$$n_{x,y} = (n_a, n_b, n_c)$$

$$n'_x = (0, -1, n_b/n_c)$$

$$n'_y = (-1, 0, n_a/n_c)$$

$$L_{\text{normal}}(x,y,z) = \left(1 - v_{x,y-1,z+\frac{n_b}{n_c}}\right)^2 + \left(1 - v_{x,y+1,z-\frac{n_b}{n_c}}\right)^2 + \left(1 - v_{x-1,y,z+\frac{n_a}{n_c}}\right)^2 + \left(1 - v_{x+1,y,z-\frac{n_a}{n_c}}\right)^2.$$



Gradients :

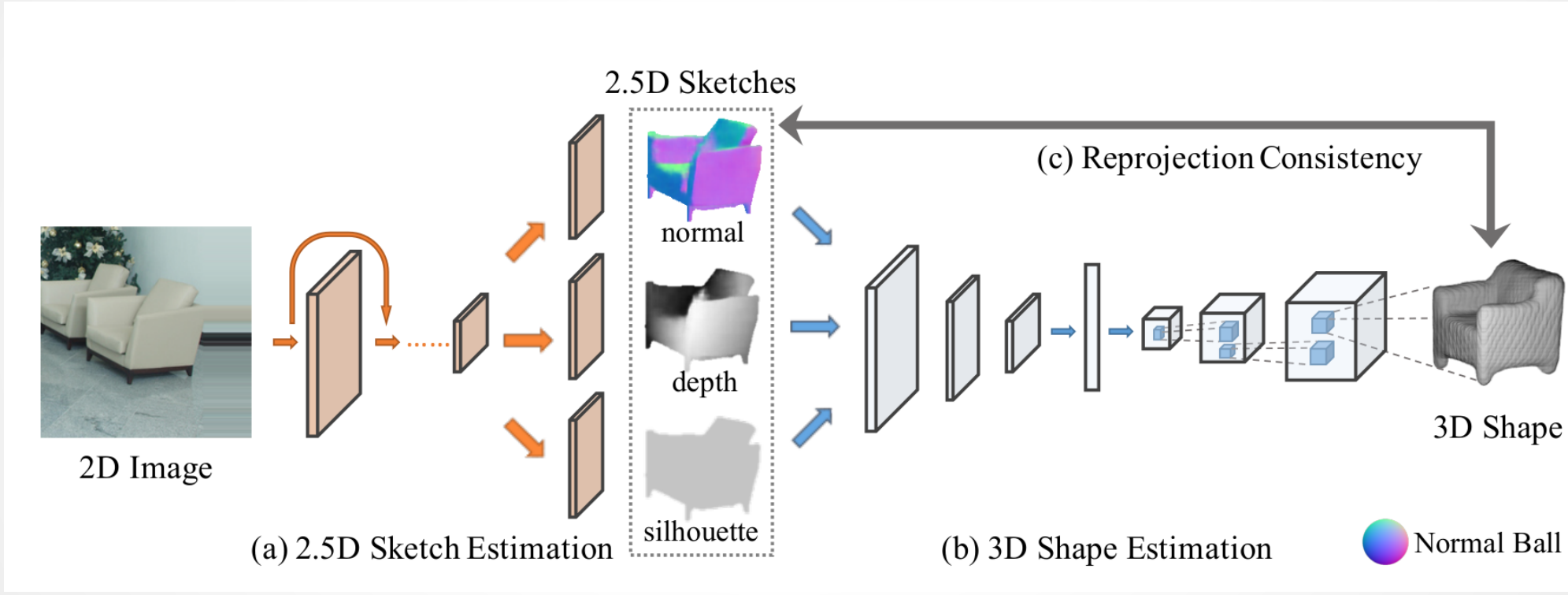
$$\frac{\partial L_{\text{normal}}(x,y,z)}{\partial v_{x-1,y,z+\frac{n_a}{n_c}}} = 2\left(v_{x-1,y,z+\frac{n_a}{n_c}} - 1\right)$$

$$\frac{\partial L_{\text{normal}}(x,y,z)}{\partial v_{x+1,y,z-\frac{n_a}{n_c}}} = 2\left(v_{x+1,y,z-\frac{n_a}{n_c}} - 1\right)$$

Training Paradigm

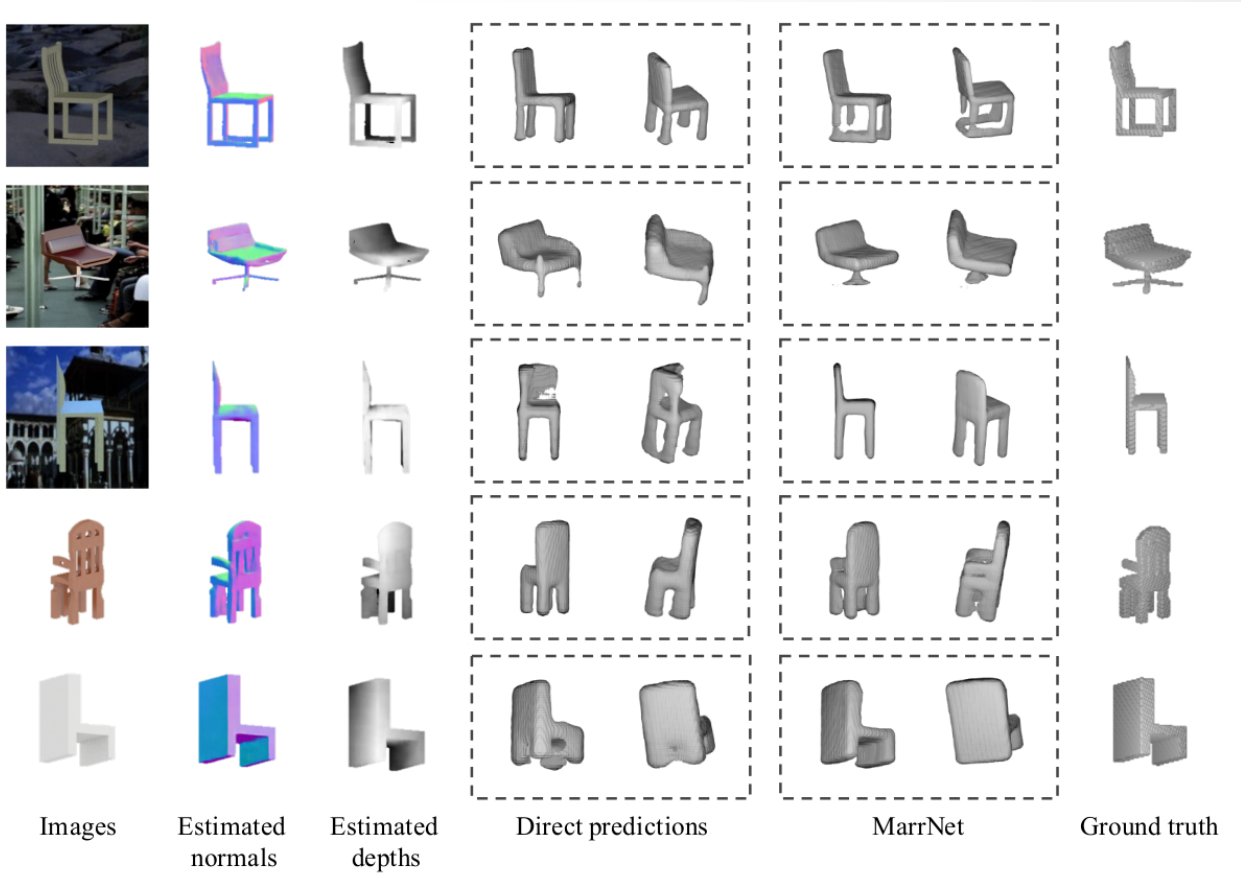
- Train with synthetic data of random angle, then fine-tune with real image.
- Pretrain :
  - ShapeNet objects
  - 2.5 Image : Normal, depth, silhouette, with L2 loss.
  - 3D : Ground truth voxels with cross entropy.
- Fine-tuning:
  - Reprojection Consistency
  - Each image for 40 iterations
  - Image overfits
  - No annotation required.
  - 10 seconds for single image. 100 milliseconds without fine-tuning.

Modules

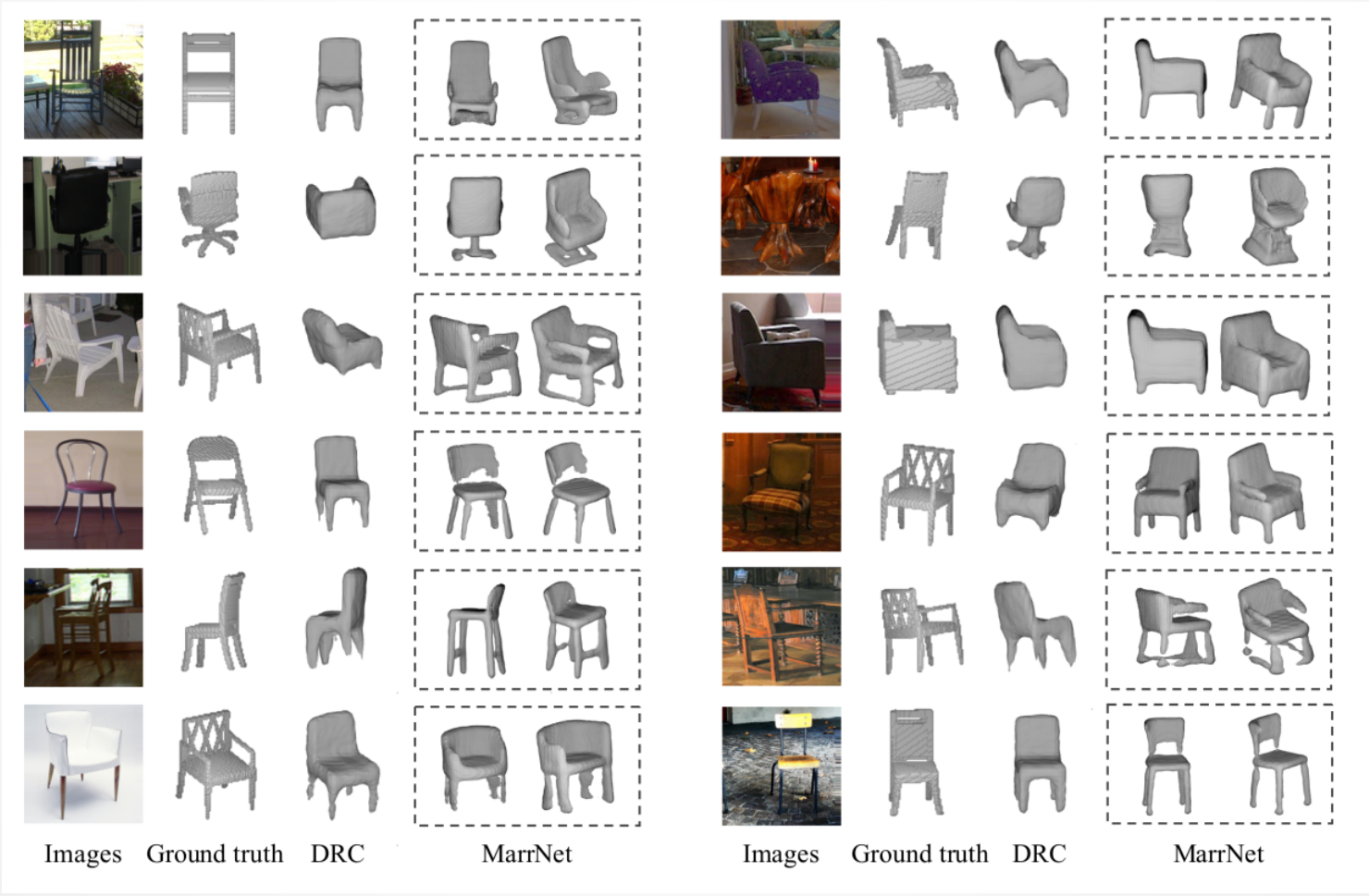

2.5D Sketches
(c) Reprojection Consistency
normal
depth
silhouette
2D Image
3D Shape
(a) 2.5D Sketch Estimation
(b) 3D Shape Estimation
Normal Ball

Results : ShapeNet

- Random background + (physics-based) rendered object inside
  - 6778 shapes
  - 20 random angles each
  - No fine-tuning
  - Ground truth available in this stage

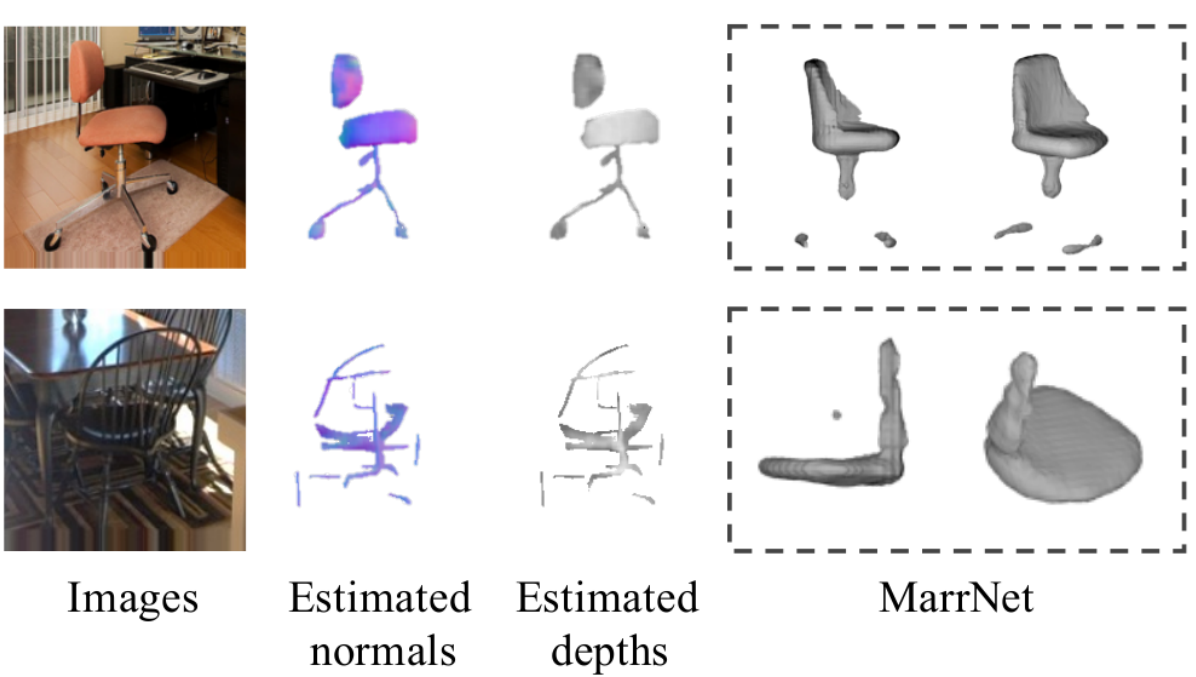


Results : Pascal 3D+

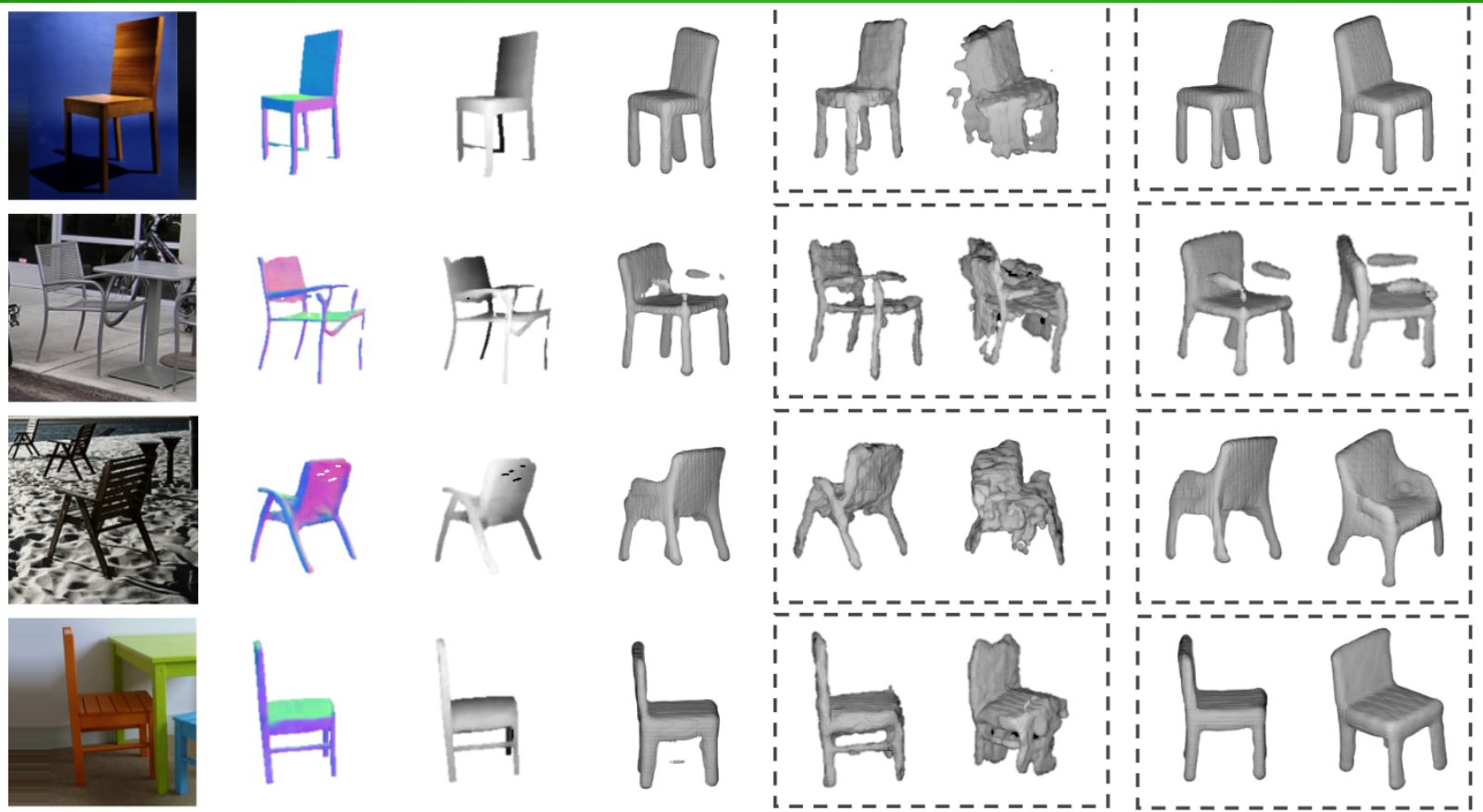


Results : Pascal 3D+ : Cont.d

	DRC	MarrNet	GT
DRC	50	26	17
MarrNet	74	50	42
Ground truth	83	58	50



Result : Ablation Study



Results : IKEA Dataset

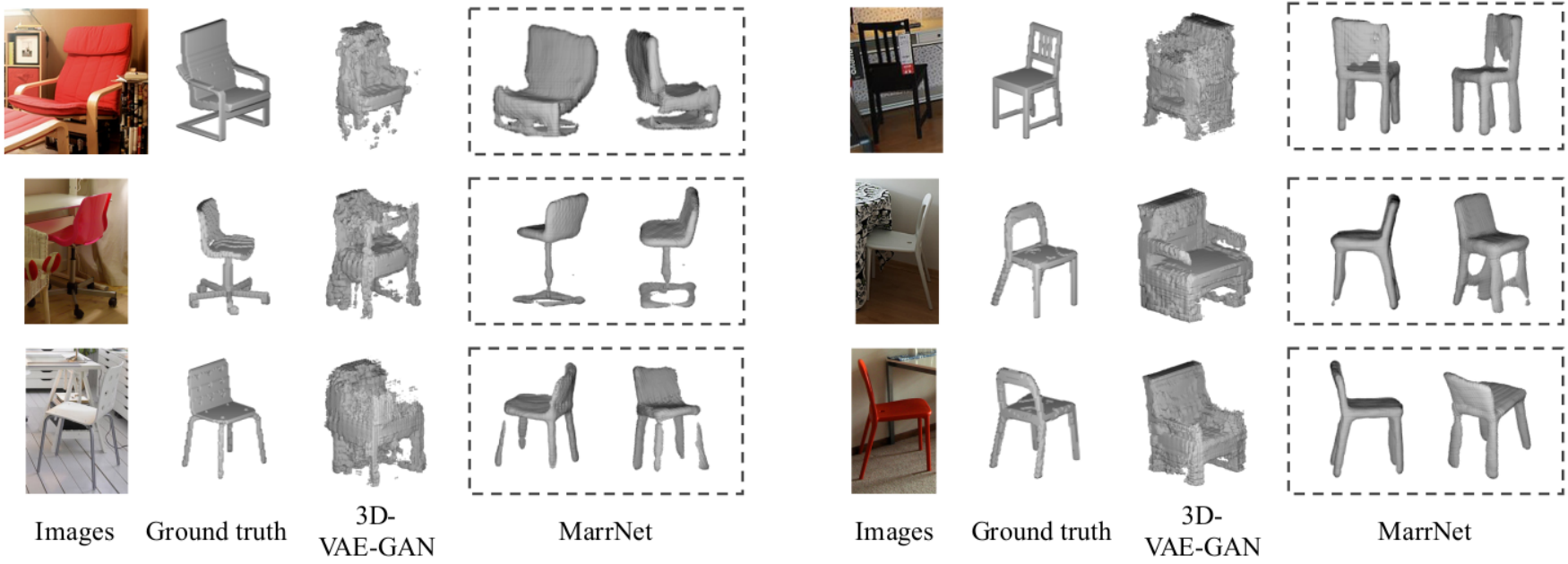


Figure 8: 3D reconstruction of chairs on the IKEA [Lim et al., 2013] dataset. From left to right: input, ground truth, 3D estimation by 3D-VAE-GAN [Wu et al., 2016b], and two views of MarrNet predictions. Our model recovers more details compared to 3D-VAE-GAN.

Other Results

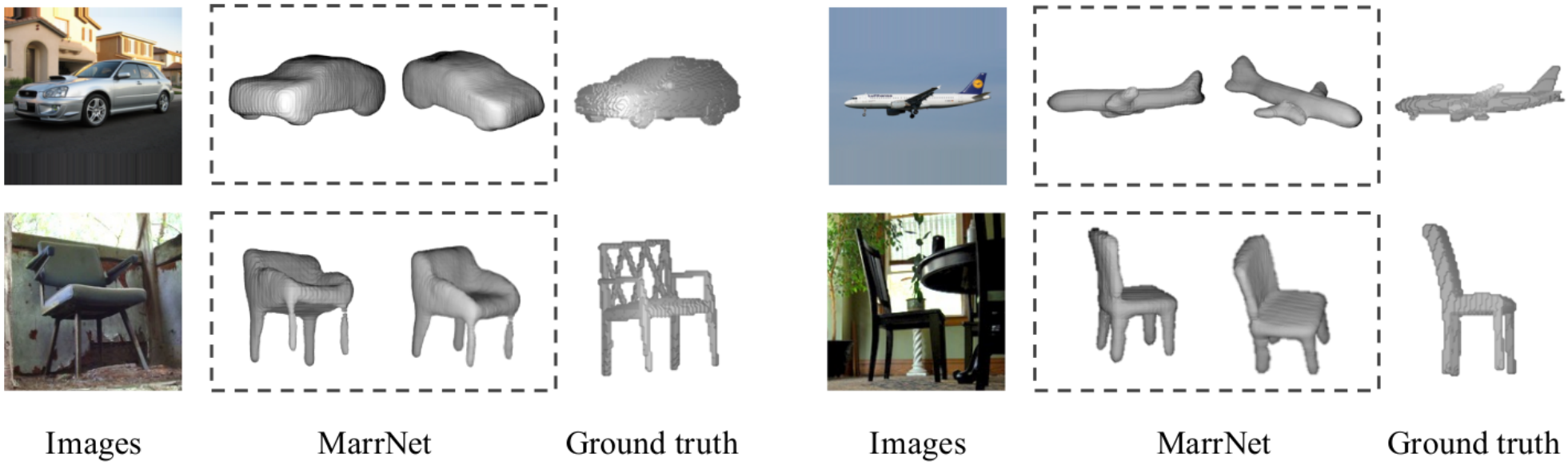


Figure 10: 3D reconstruction of objects from multiple categories on the PASCAL 3D+ [Xiang et al., 2014] dataset. MarrNet also recovers 3D shape well when it is trained on multiple categories.

Conclusion

- MarrNet : A novel model that explicitly models 2.5D sketches for single-image 3D shape reconstruction.
- Differentiable loss functions for the consistency between 3D shape and 2.5D sketches, enabling end-to-end fine-tuning on real images without annotations.