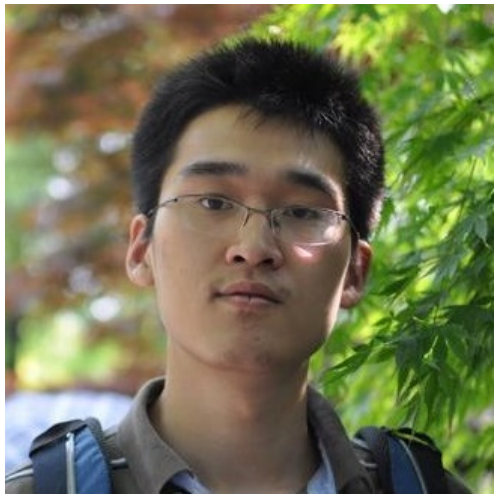
Learning a Probabilistic Latent Space of Object Shapes via 3D Generative-Adversarial Modeling

Jiajun Wu*

Chengkai Zhang*



Tianfan Xue

Bill Freeman

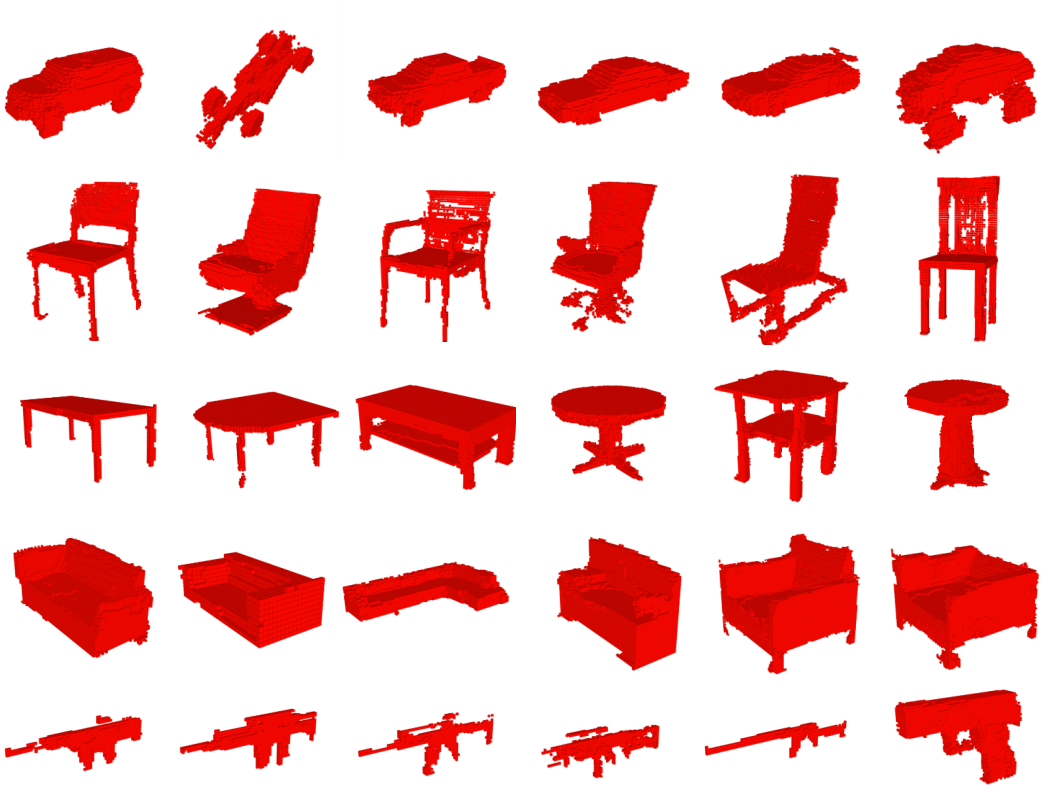
Josh Tenenbaum

NIPS 2016

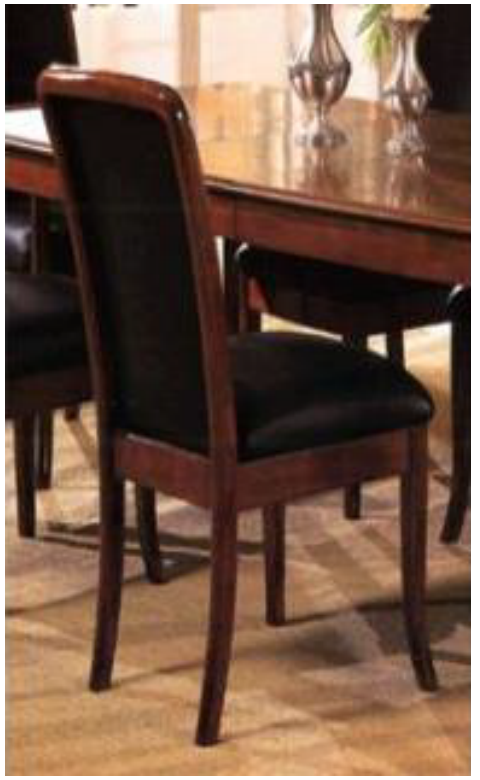
(* indicates equal contribution)

Outline

Synthesizing 3D shapes

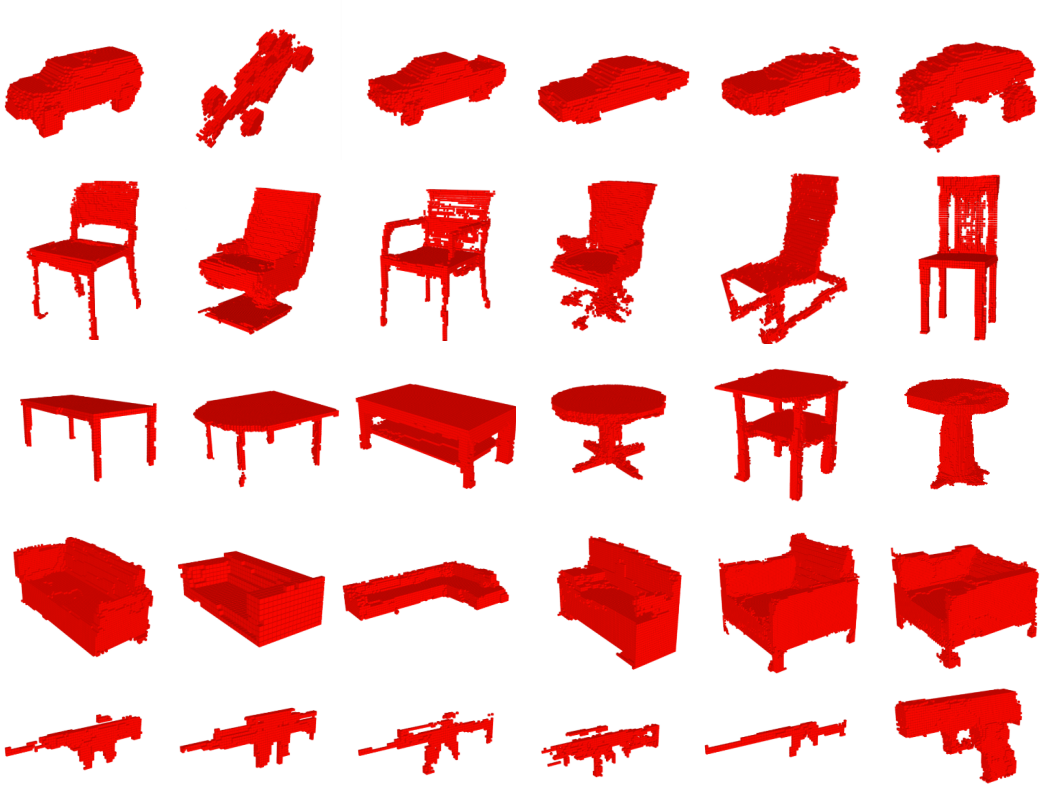


Recognizing 3D structure

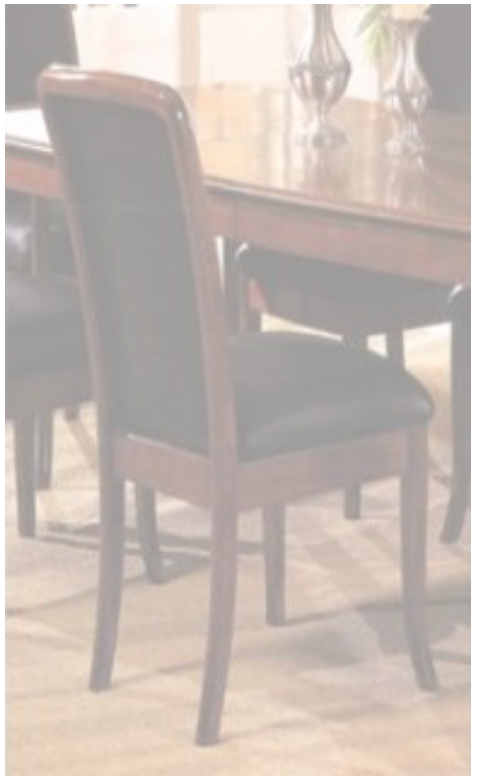


Outline

Synthesizing 3D shapes



Recognizing 3D structure



3D Shape Synthesis

Templated-based model

- Synthesizing realistic shapes
- Requiring a large shape repository
- Recombining parts and pieces

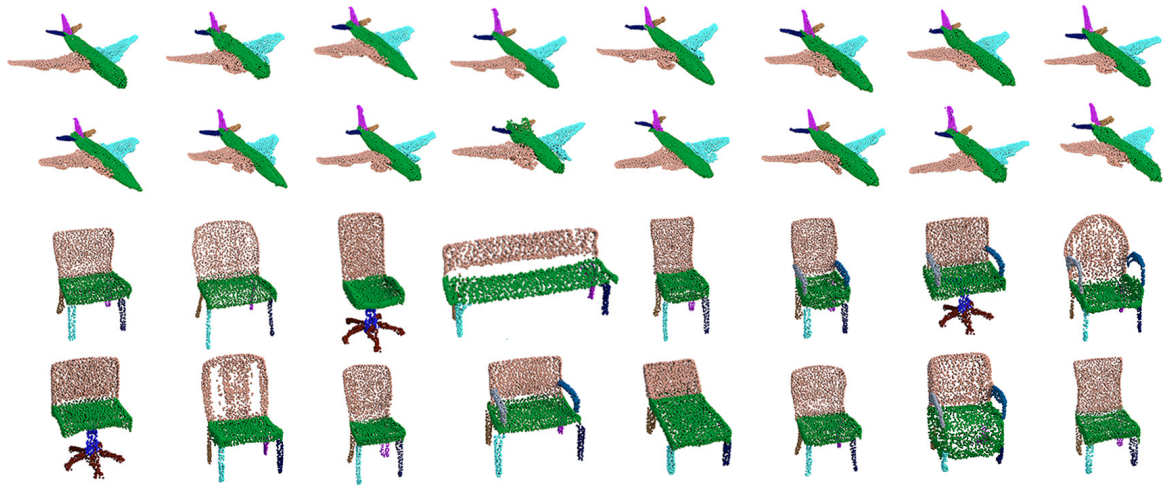


Image credit: [Huang et al., SGP 2015]

3D Shape Synthesis

Voxel-based deep generative model

- Synthesizing new shapes
- Hard to scale up to high resolution
- Resulting in not-as-realistic shapes

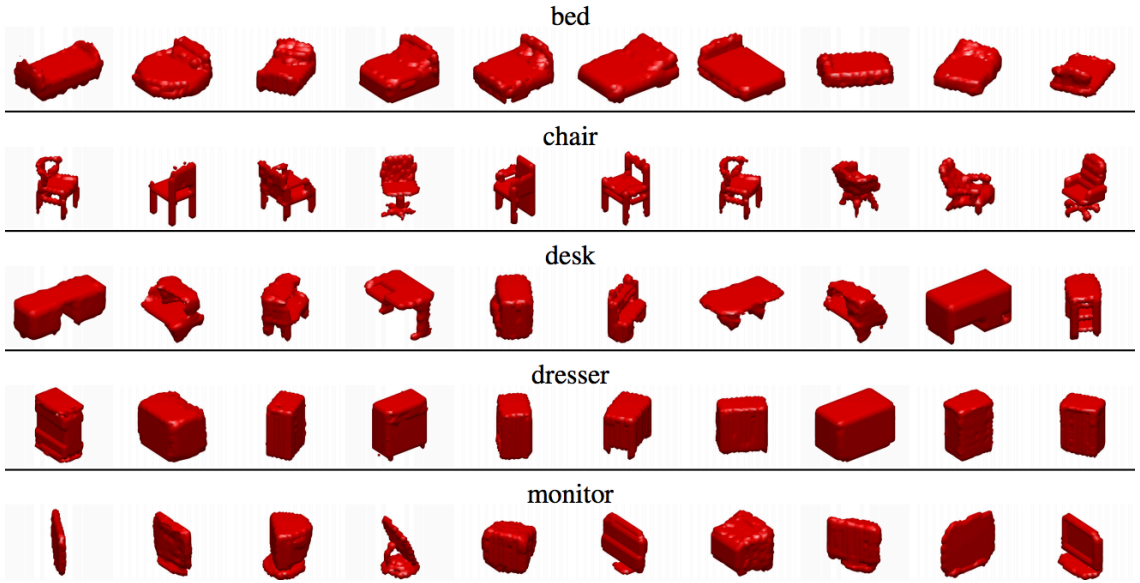
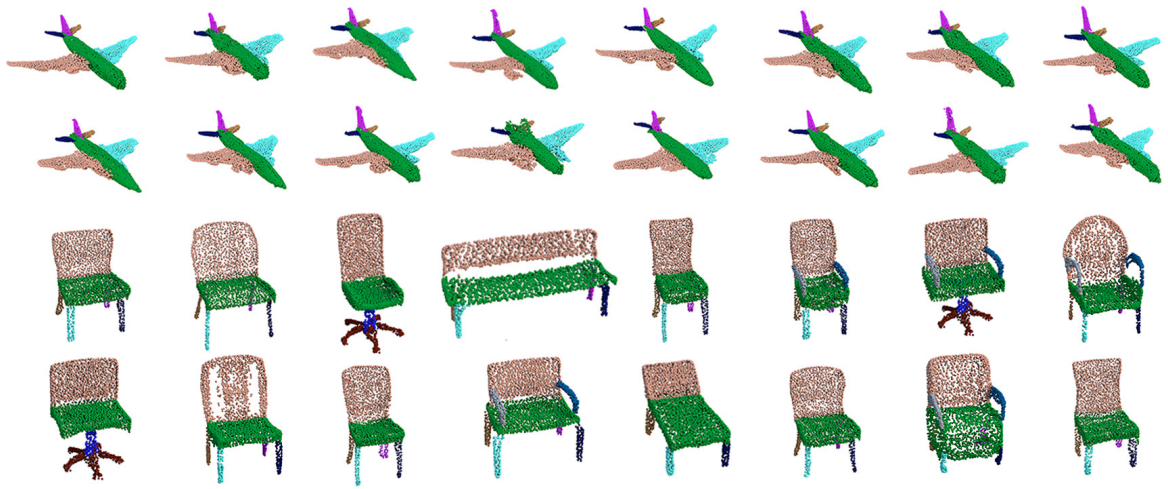


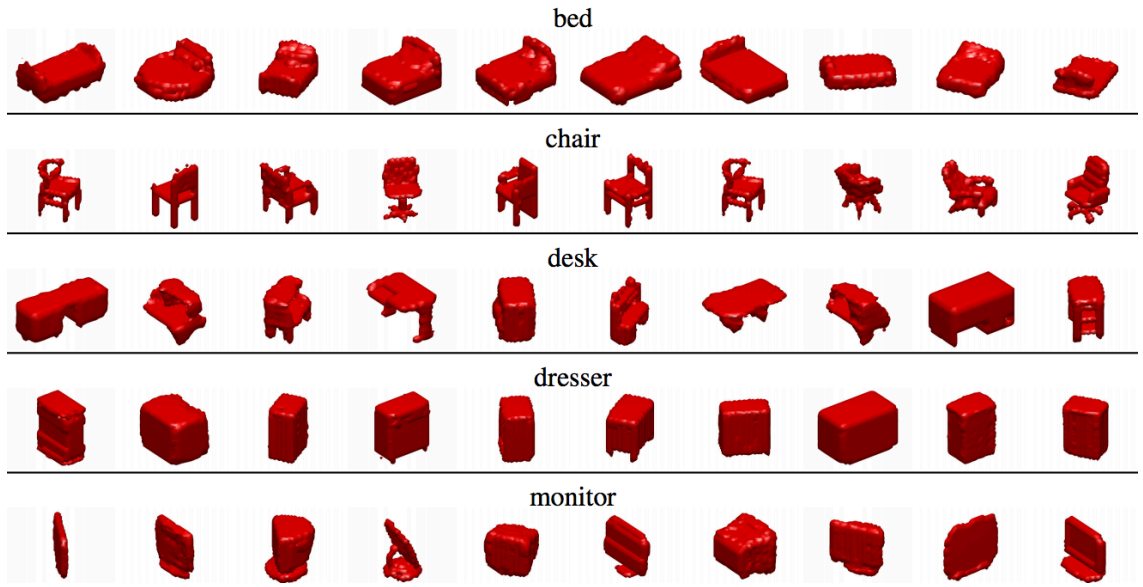
Image credit: 3D ShapeNet
[Wu et al., CVPR 2015]

3D Shape Synthesis

Realistic

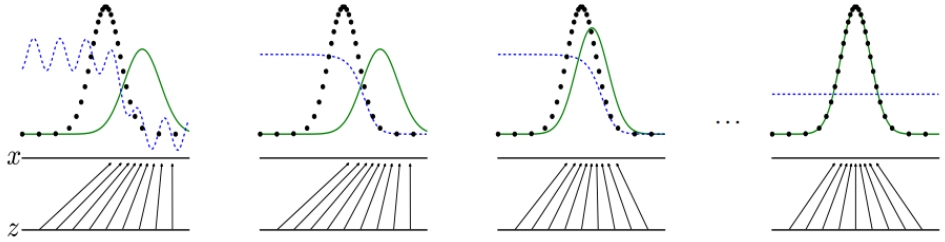


New

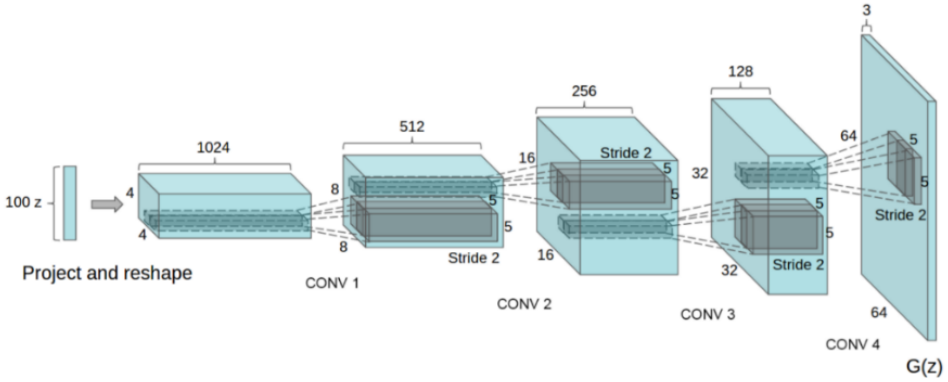


Realistic + New

Adversarial Learning



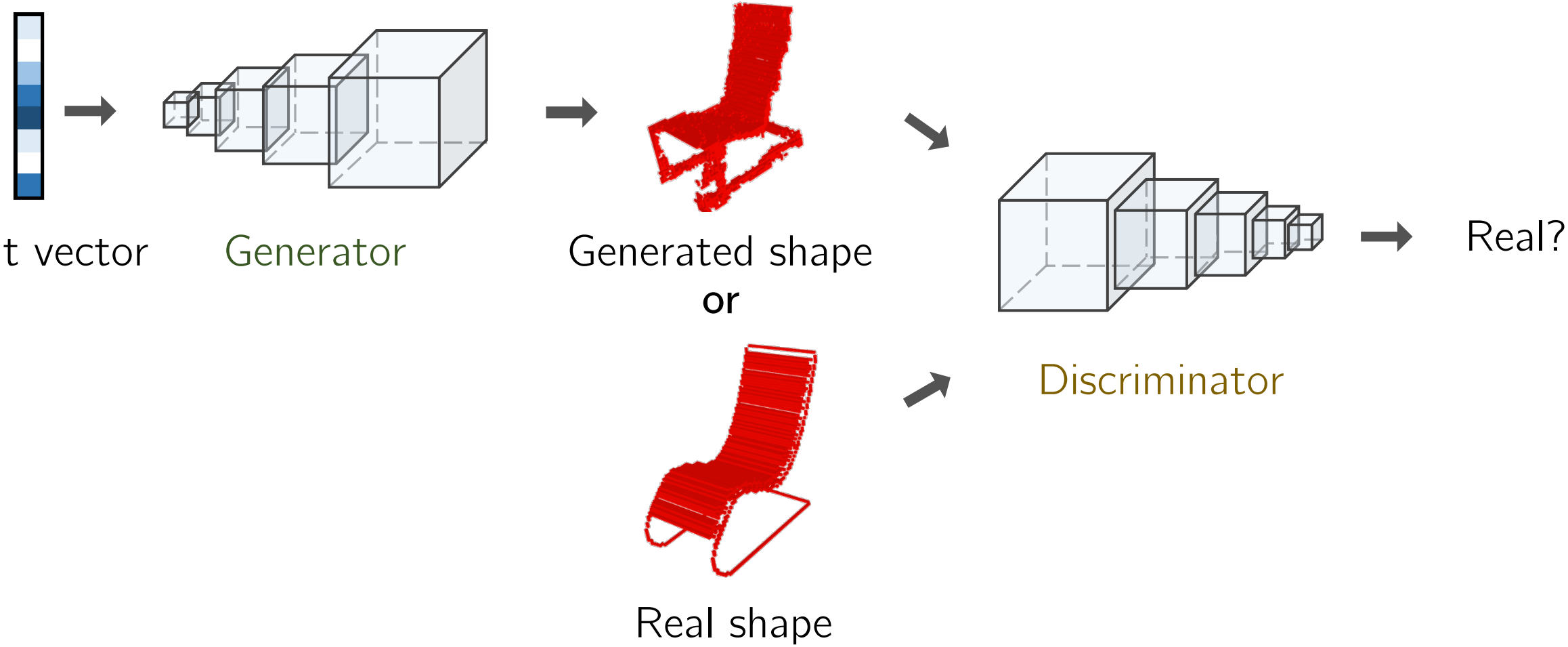
Generative adversarial networks
[Goodfellow et al., NIPS 2014]



DCGAN [Radford et al., ICLR 2016]

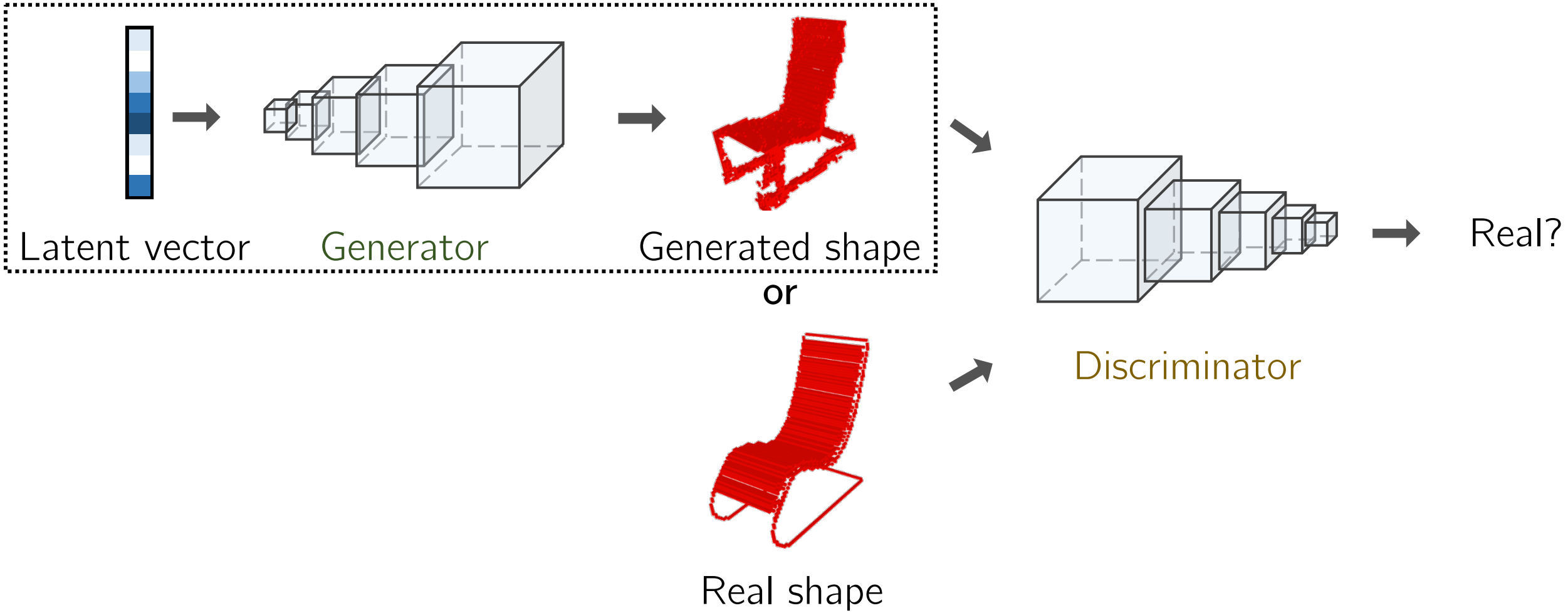
Our Synthesized 3D Shapes

3D Generative Adversarial Network



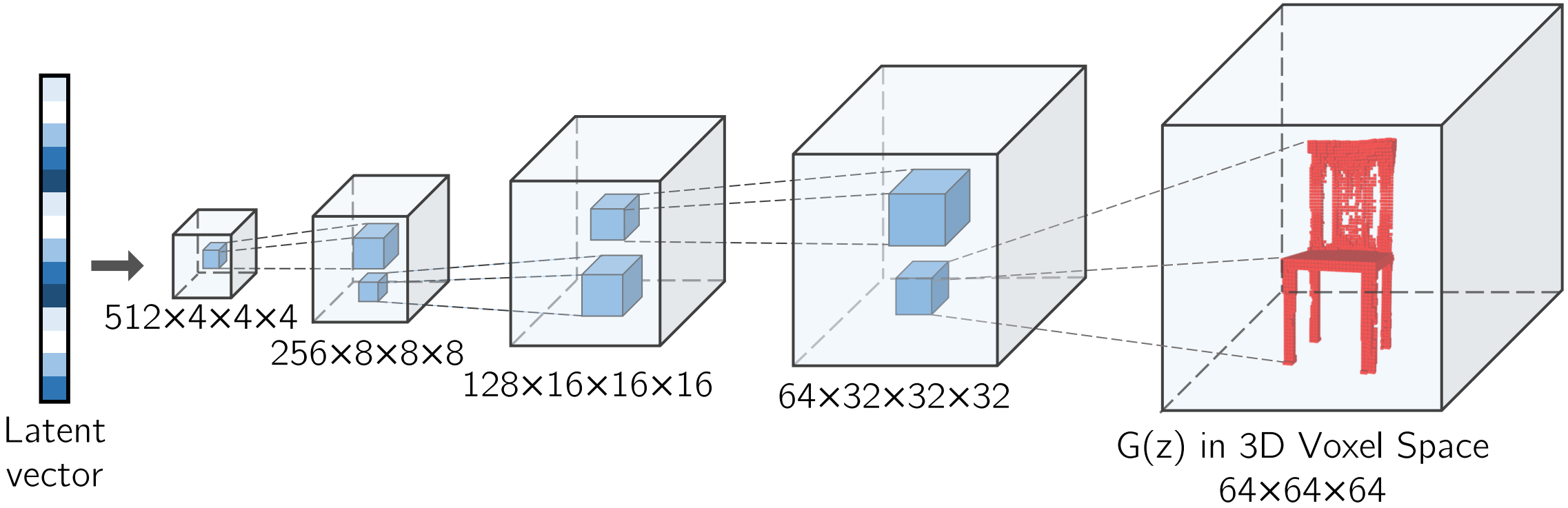
Training on ShapeNet [Chang et al., 2015]

3D Generative Adversarial Network



Training on ShapeNet [Chang et al., 2015]

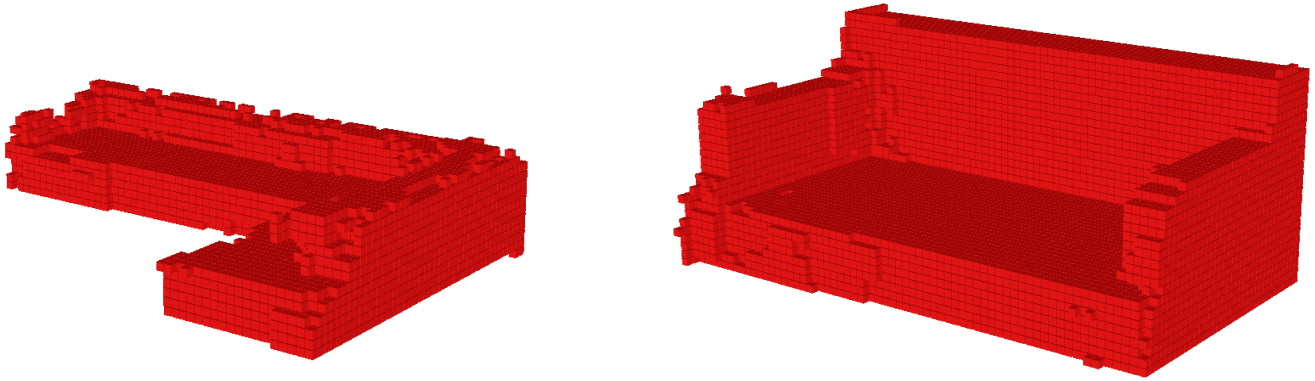
Generator Structure



Randomly Sampled Shapes

Chairs

Sofas

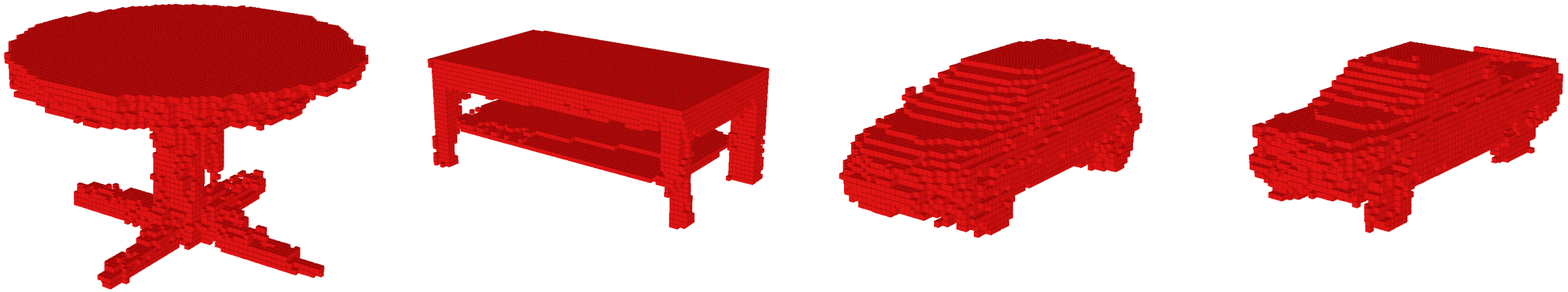


Results from 3D ShapeNet

Randomly Sampled Shapes

Tables

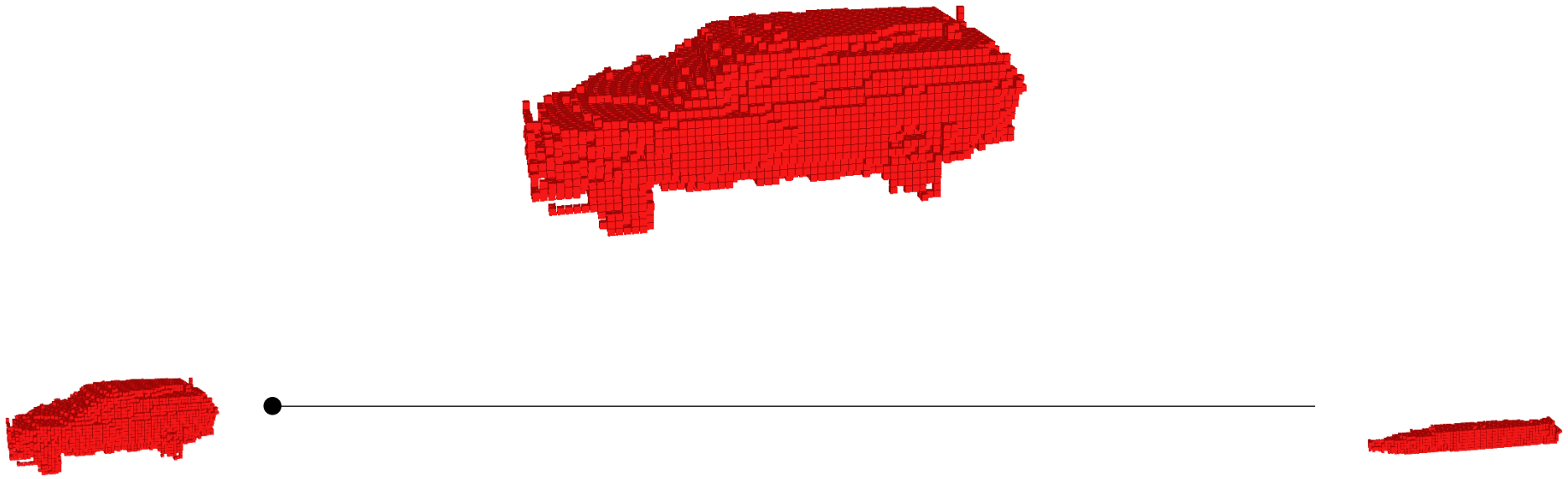
Cars



Results from 3D ShapeNet

Interpolation in Latent Space

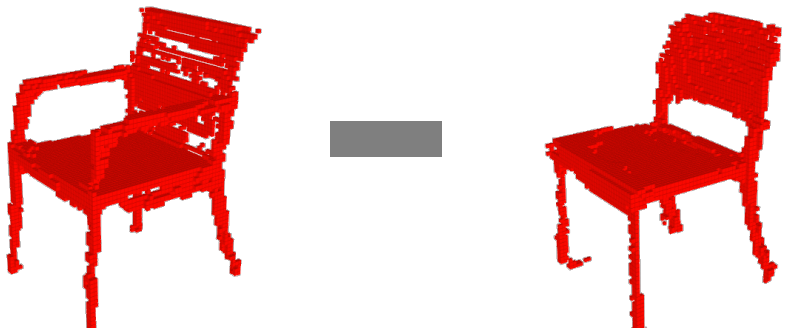
Interpolation in Latent Space



Arithmetic in Latent Space

Latent space

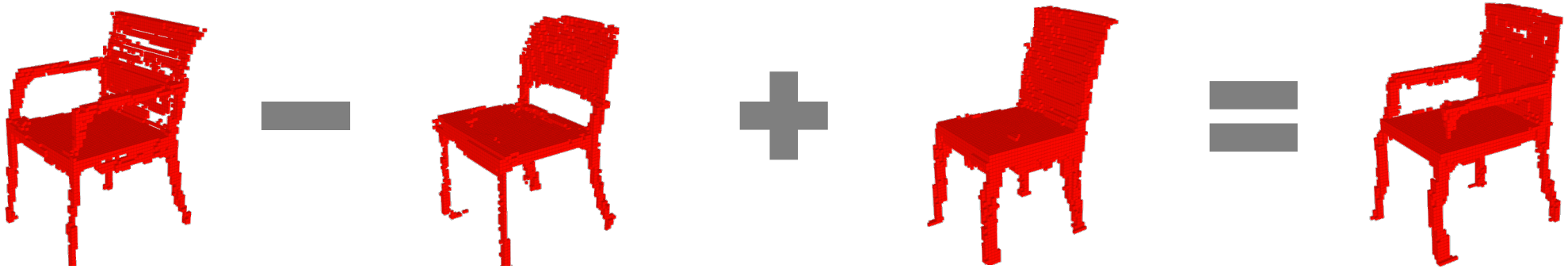
Shape space



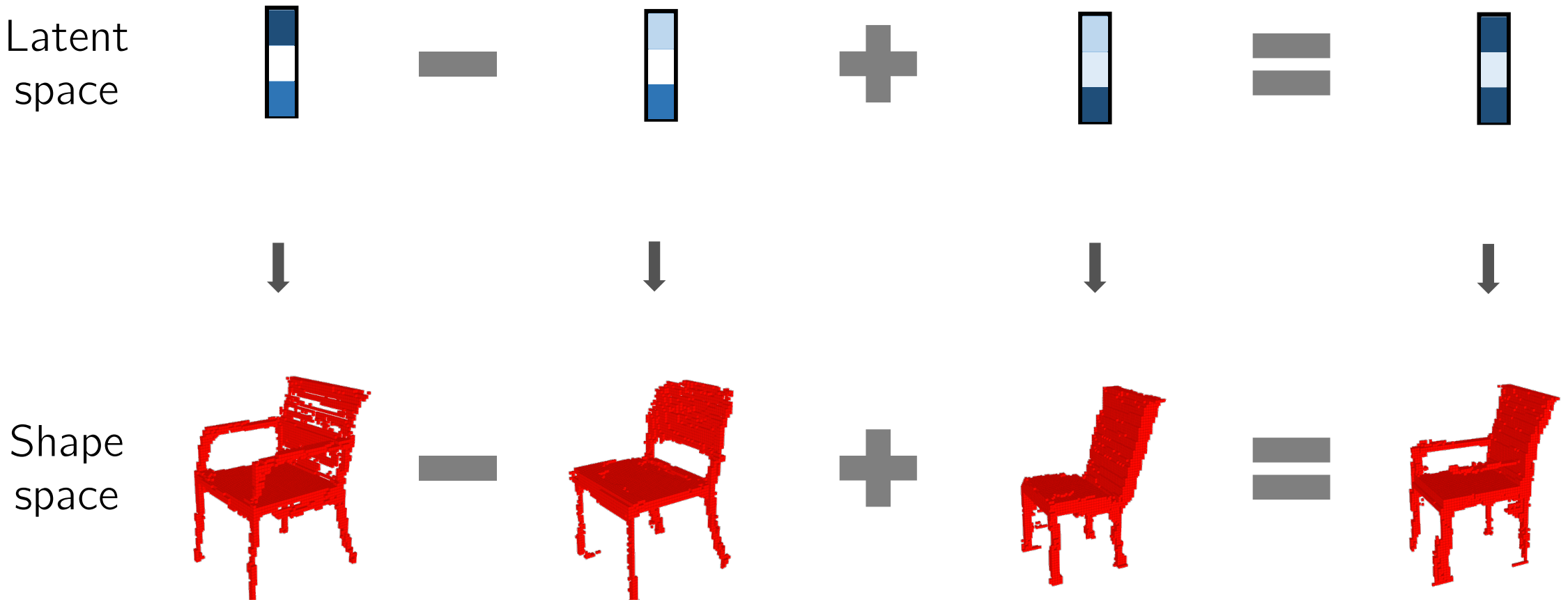
Arithmetic in Latent Space

Latent space

Shape space



Arithmetic in Latent Space



Arithmetic in Latent Space

Latent space

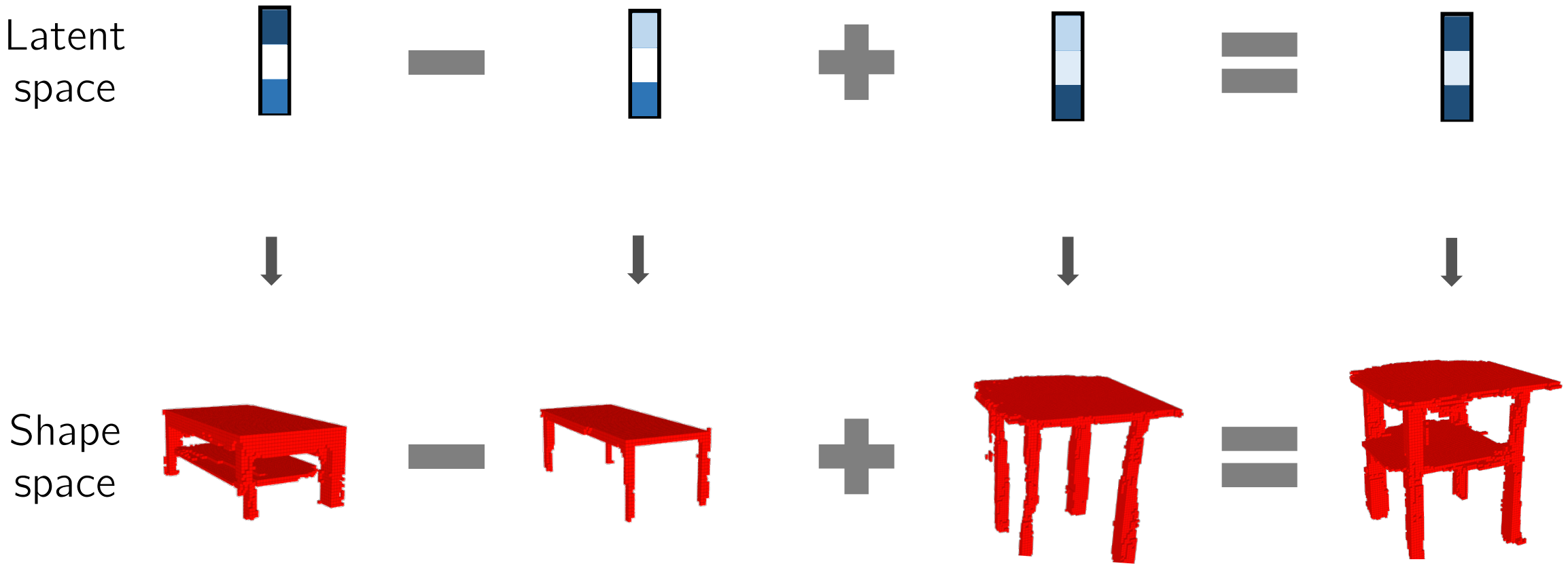
Shape space

Arithmetic in Latent Space

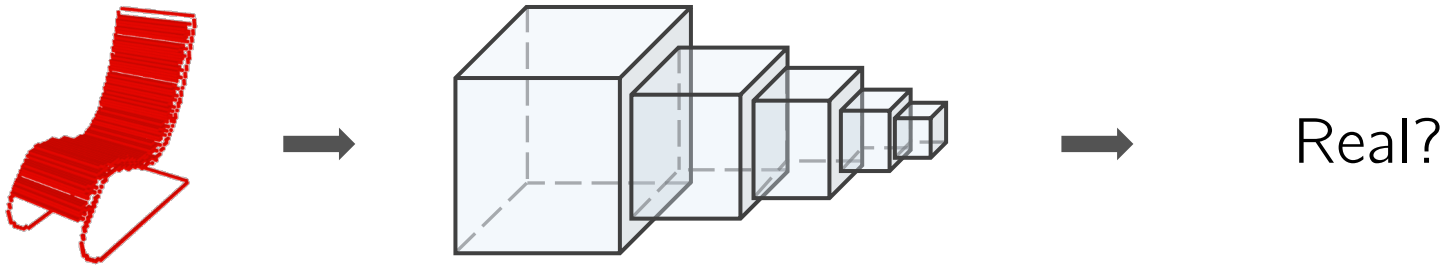
Latent space

Shape space

Arithmetic in Latent Space

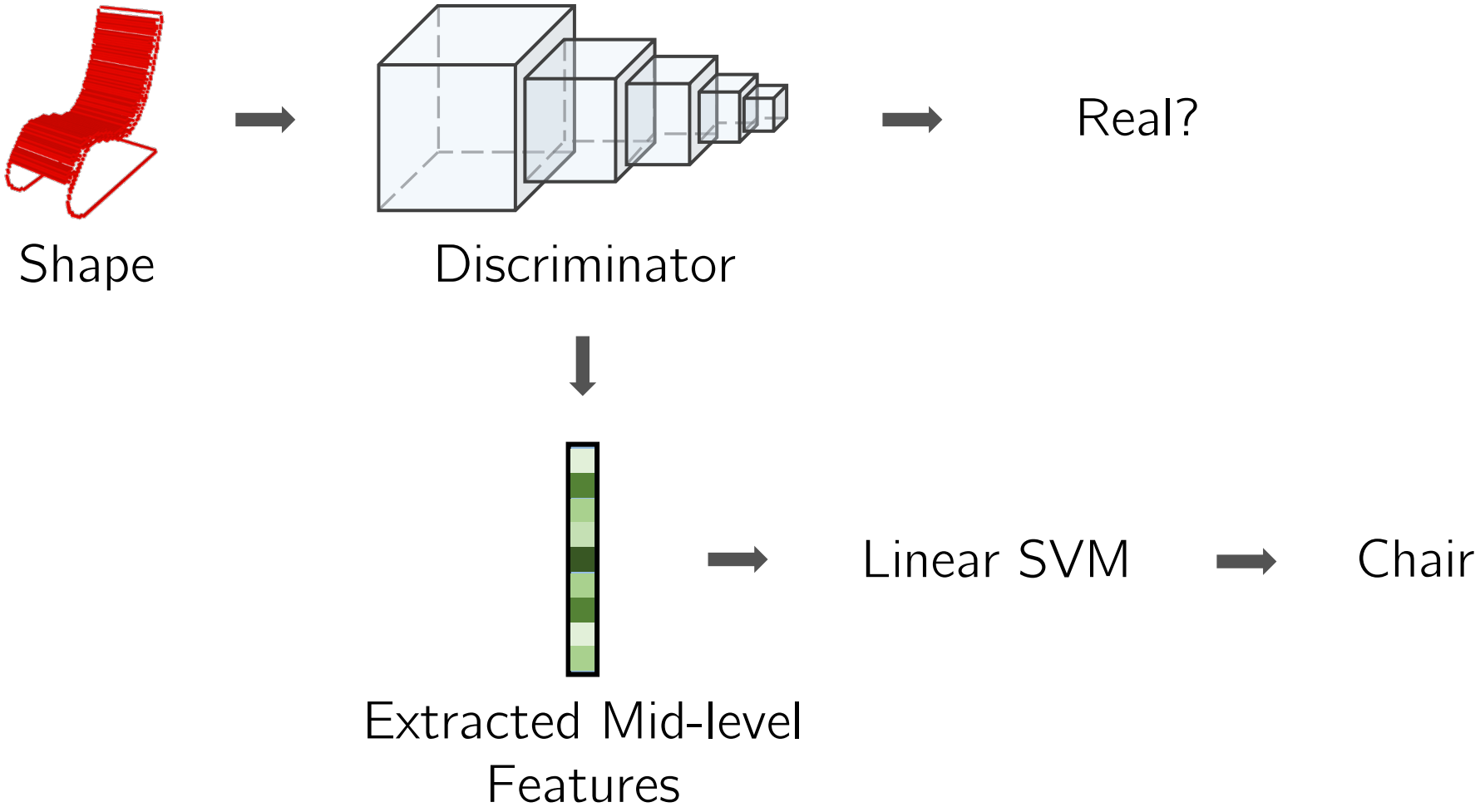


Unsupervised 3D Shape Descriptors



Extracted Mid-level Features

3D Shape Classification



3D Shape Classification Results

Supervision	Pretraining	Method	Classification (Accuracy)	
			ModelNet40	ModelNet10
Category labels	ImageNet	MVCNN [Su et al., 2015]	90.1%	-
		MVCNN-MultiRes [Qi et al., 2016]	**91.4%**	-
	None	3D ShapeNets [Wu et al., 2015]	77.3%	83.5%
		DeepPano [Shi et al., 2015]	77.6%	85.5%
		VoxNet [Maturana and Scherer, 2015]	83.0%	92.0%
		ORION [Sedaghat et al., 2016]	-	**93.8%**
Unsupervised	-	SPH [Kazhdan et al., 2003]	68.2%	79.8%
		LFD [Chen et al., 2003]	75.5%	79.9%
		T-L Network [Girdhar et al., 2016]	74.4%	-
		Vconv-DAE [Sharma et al., 2016]	75.5%	80.5%
		3D-GAN (ours)	**83.3%**	**91.0%**

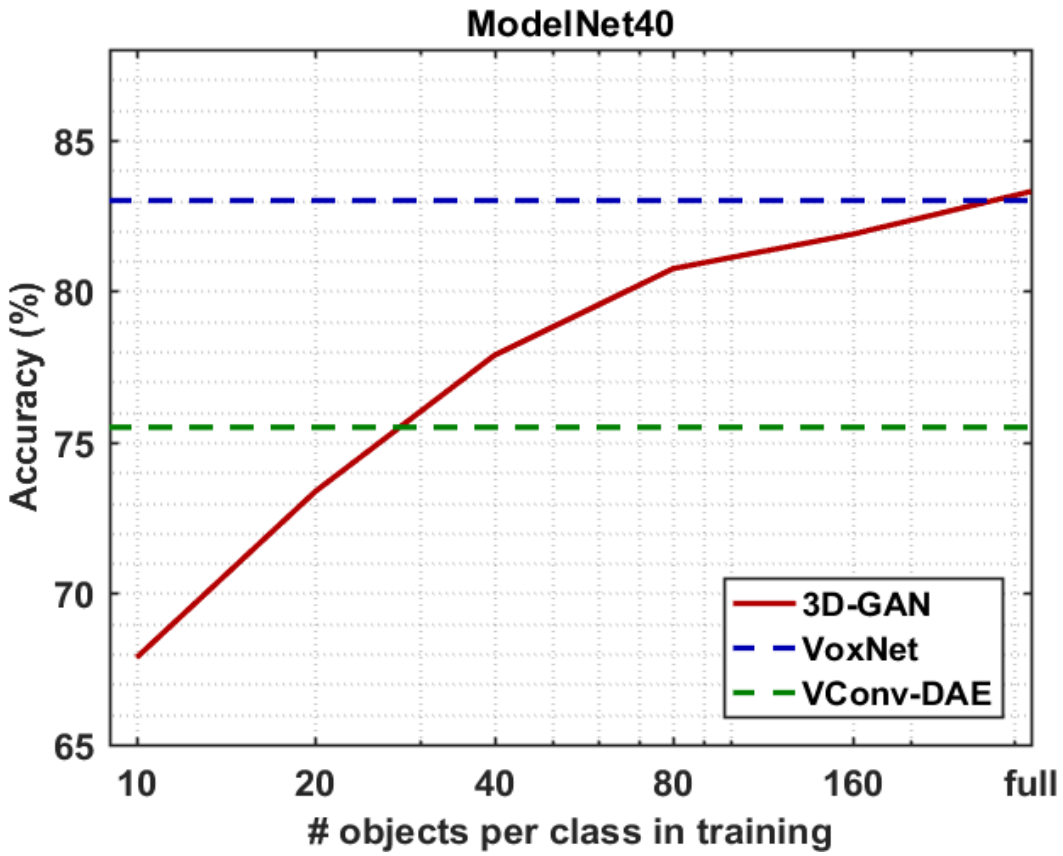
3D Shape Classification Results

Supervision	Pretraining	Method	Classification (Accuracy)	
			ModelNet40	ModelNet10
Category labels	ImageNet	MVCNN [Su et al., 2015]	90.1%	-
		MVCNN-MultiRes [Qi et al., 2016]	**91.4%**	-
	None	3D ShapeNets [Wu et al., 2015]	77.3%	83.5%
		DeepPano [Shi et al., 2015]	77.6%	85.5%
		VoxNet [Maturana and Scherer, 2015]	83.0%	92.0%
		ORION [Sedaghat et al., 2016]	-	**93.8%**
Unsupervised	-	SPH [Kazhdan et al., 2003]	68.2%	79.8%
		LFD [Chen et al., 2003]	75.5%	79.9%
		T-L Network [Girdhar et al., 2016]	74.4%	-
		Vconv-DAE [Sharma et al., 2016]	75.5%	80.5%
		3D-GAN (ours)	**83.3%**	**91.0%**

3D Shape Classification Results

Supervision	Pretraining	Method	Classification (Accuracy)	
			ModelNet40	ModelNet10
Category labels	ImageNet	MVCNN [Su et al., 2015]	90.1%	-
		MVCNN-MultiRes [Qi et al., 2016]	**91.4%**	-
	None	3D ShapeNets [Wu et al., 2015]	77.3%	83.5%
		DeepPano [Shi et al., 2015]	77.6%	85.5%
		VoxNet [Maturana and Scherer, 2015]	83.0%	92.0%
		ORION [Sedaghat et al., 2016]	-	**93.8%**
Unsupervised	-	SPH [Kazhdan et al., 2003]	68.2%	79.8%
		LFD [Chen et al., 2003]	75.5%	79.9%
		T-L Network [Girdhar et al., 2016]	74.4%	-
		Vconv-DAE [Sharma et al., 2016]	75.5%	80.5%
		3D-GAN (ours)	**83.3%**	**91.0%**

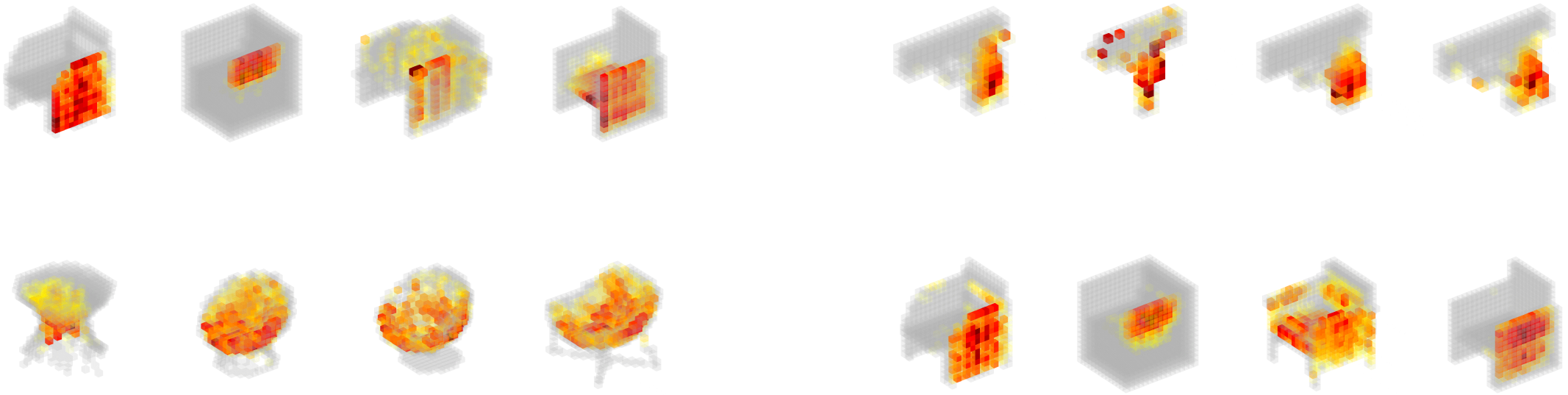
Limited Training Samples



Comparable with best unsupervisedly learned features with about 25 training samples/class

Comparable with best voxel-based supervised descriptors with the entire training set

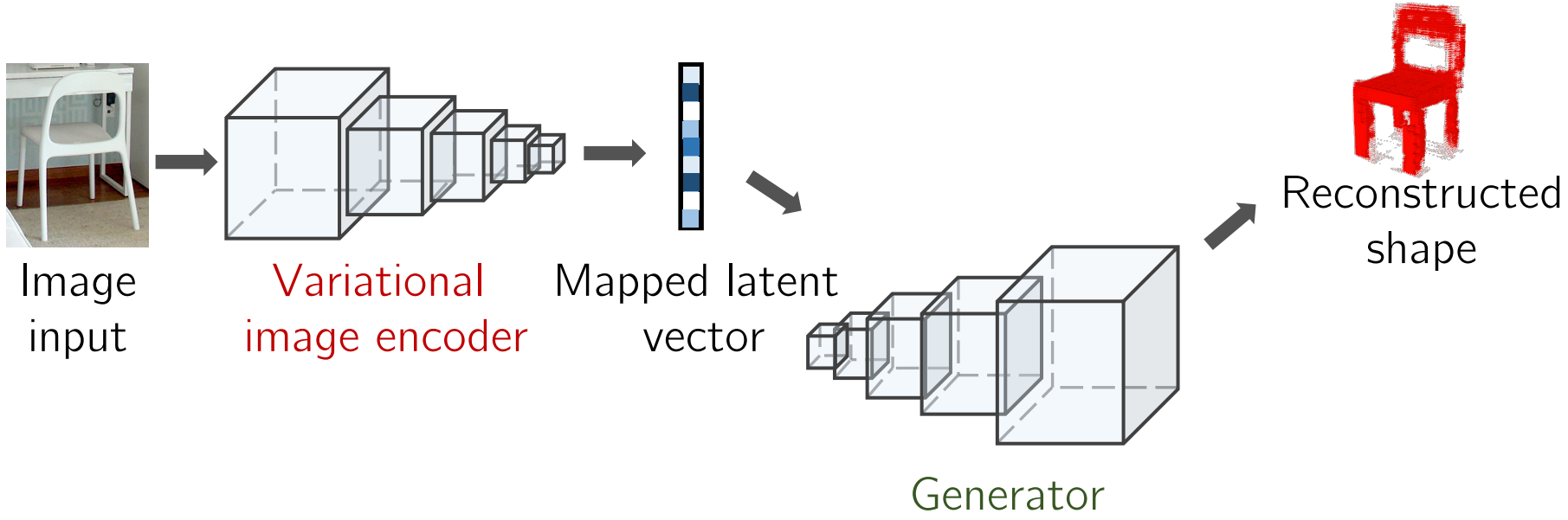
Discriminator Activations



Units respond to certain object shapes and their parts.

Extension: Single Image 3D Reconstruction

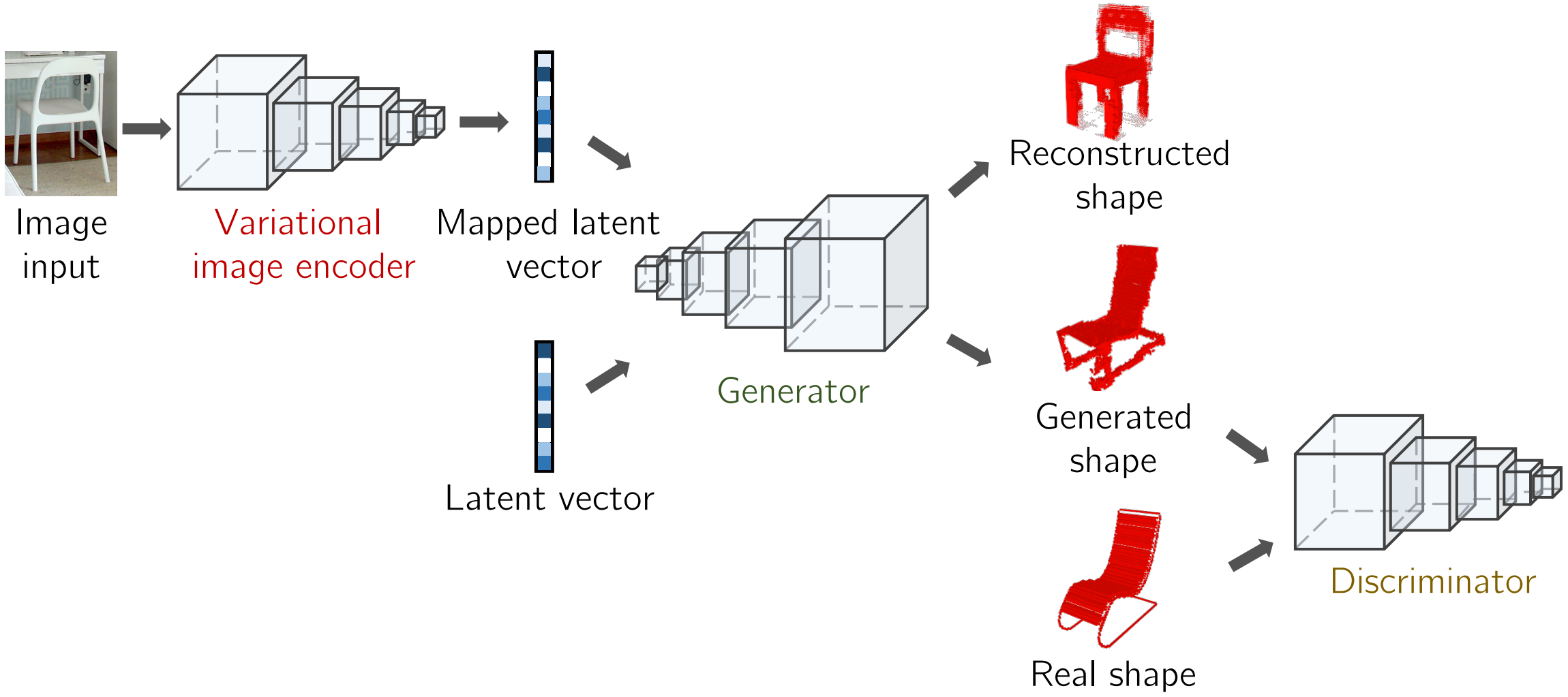
Model: 3D-VAE-GAN



A variational image encoder maps an image to a latent vector for 3D object reconstruction.

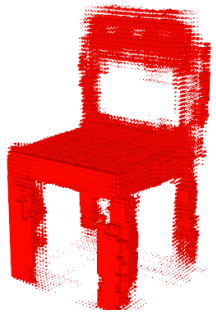
VAE-GAN [Larson et al., ICML 2016], TL-Network [Girdhar et al., ECCV 2016]

Model: 3D-VAE-GAN



We combine the encoder with 3D-GAN for reconstruction and generation.

Single Image 3D Reconstruction



Input image

Reconstructed 3D shape

Input image

Reconstructed 3D shape

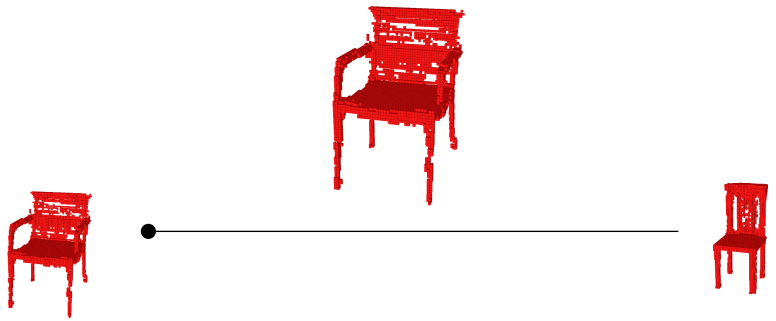
Single Image 3D Reconstruction

	Bed	Bookcase	Chair	Desk	Sofa	Table	Mean
AlexNet-fc8 [Girdhar et al., 2016]	29.5	17.3	20.4	19.7	38.8	16.0	23.6
AlexNet-conv4 [Girdhar et al., 2016]	38.2	26.6	31.4	26.6	69.3	19.1	35.2
T-L Network [Girdhar et al., 2016]	56.3	30.2	32.9	25.8	71.7	23.3	40.0
Our 3D-VAE-GAN (jointly trained)	49.1	31.9	42.6	34.8	**79.8**	33.1	45.2
Our 3D-VAE-GAN (separately trained)	**63.2**	**46.3**	**47.2**	**40.7**	78.8	**42.3**	**53.1**

Average precision on IKEA dataset [Lim et al., ICCV 2013]

Contributions of 3D-GAN

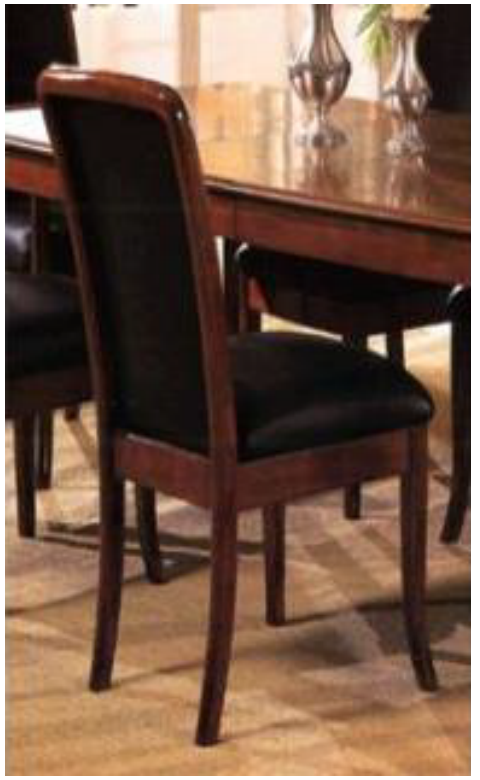
- Synthesizing new and realistic 3D shapes via adversarial learning
  - Exploring the latent shape space
- Extracting powerful shape descriptors for classification
- Extending 3D-GAN for single image 3D reconstruction



Outline

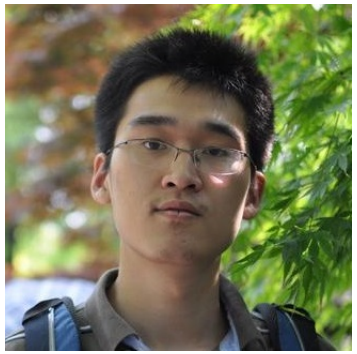
3D-GAN: Synthesizing 3D shapes

Recognizing 3D structure

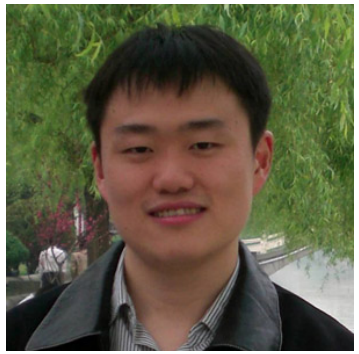


Single Image 3D Interpreter Network

Jiajun Wu*


Tianfan Xue*


Joseph Lim


Yuandong Tian


Josh Tenenbaum

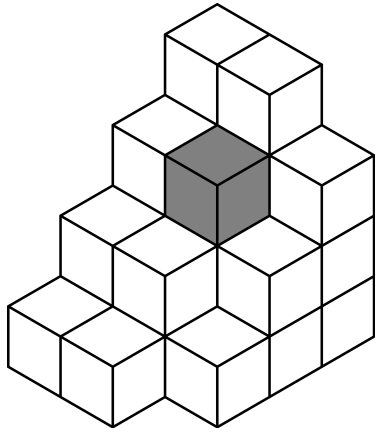

Antonio Torralba


Bill Freeman

ECCV 2016

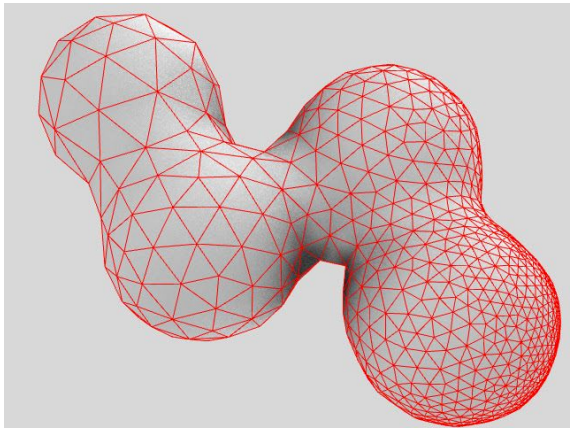
(* indicates equal contribution)

3D Object Representation



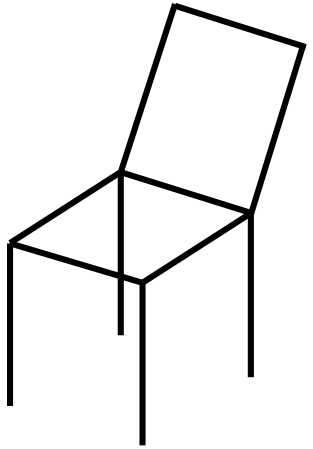
Voxel

Girdhar et al. '16
Choy et al. '16
Xiao et al. '12



Mesh

Goesele et al. '10
Furukawa and Ponce, '07
Lensch et al. '03

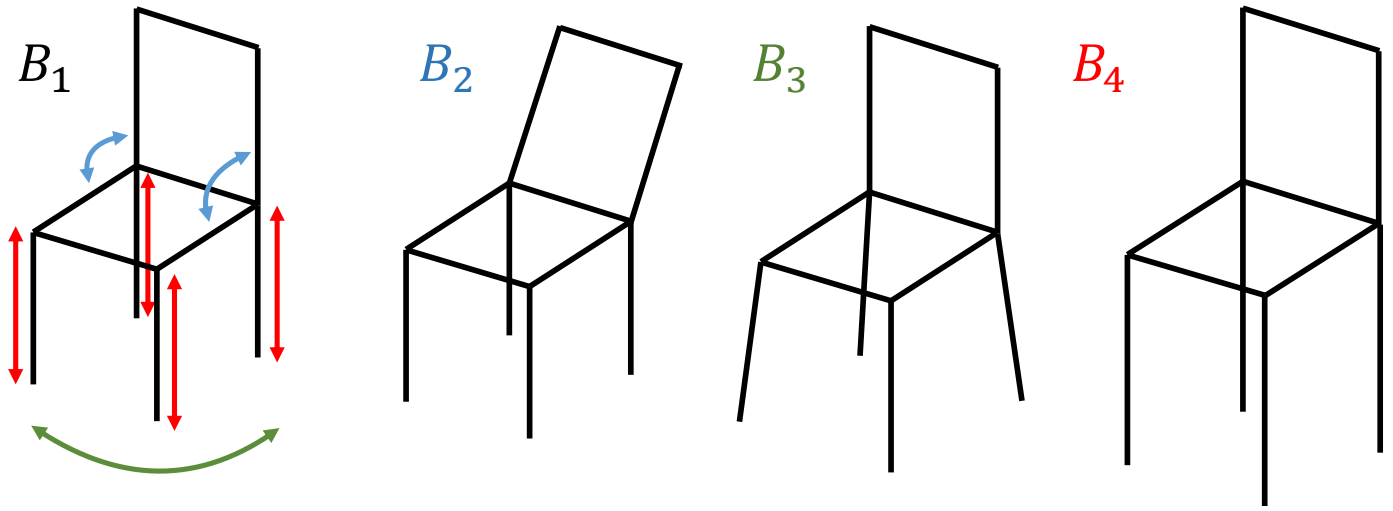


Skeleton

Zhou et al. '16
Biederman et al. '93
Fan et al. '89

Goal

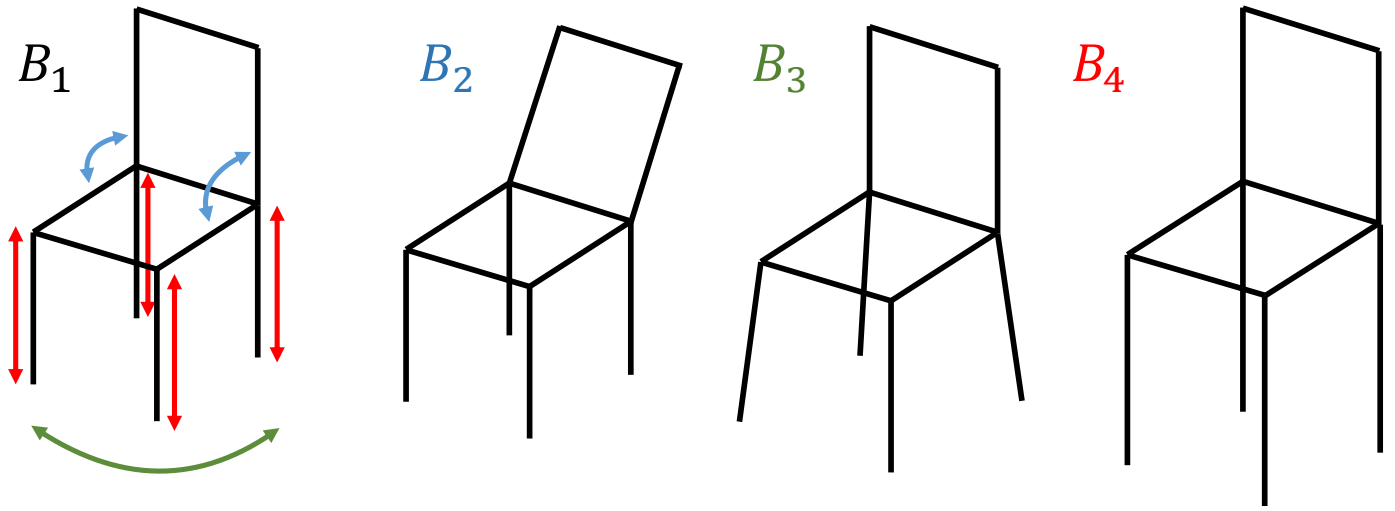
Skeleton Representation

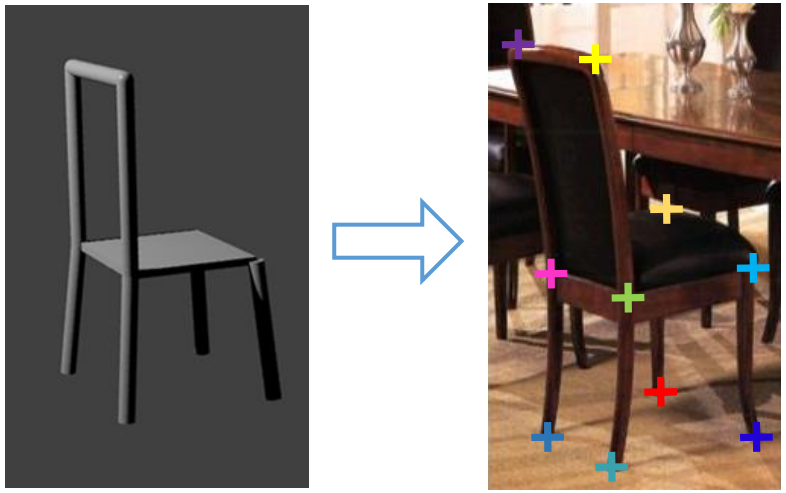


$$\sum_{k=1}^{K} \alpha_k B_k$$

structure parameter

3D Skeleton to 2D Image





$$P(R\sum_{k=1}^{K}\alpha_k B_k + T)$$

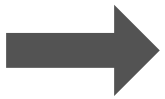
projection

rotation

structure parameter

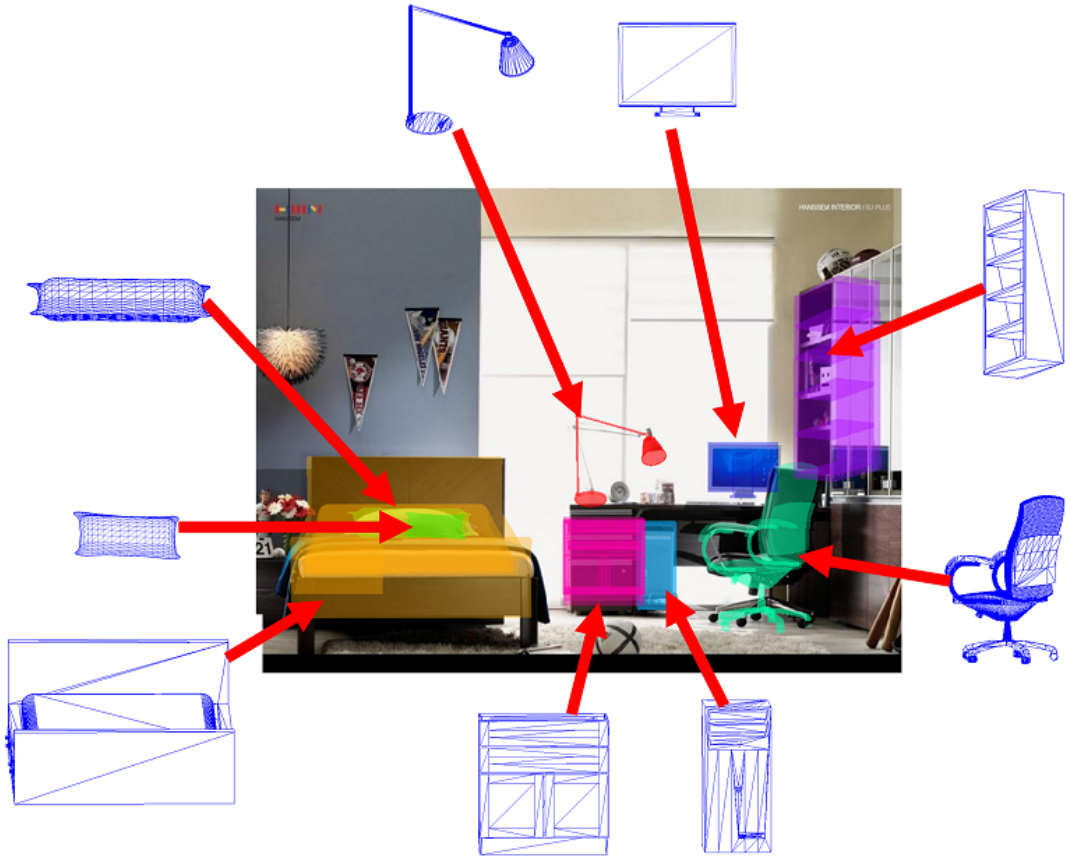
translation

Goal



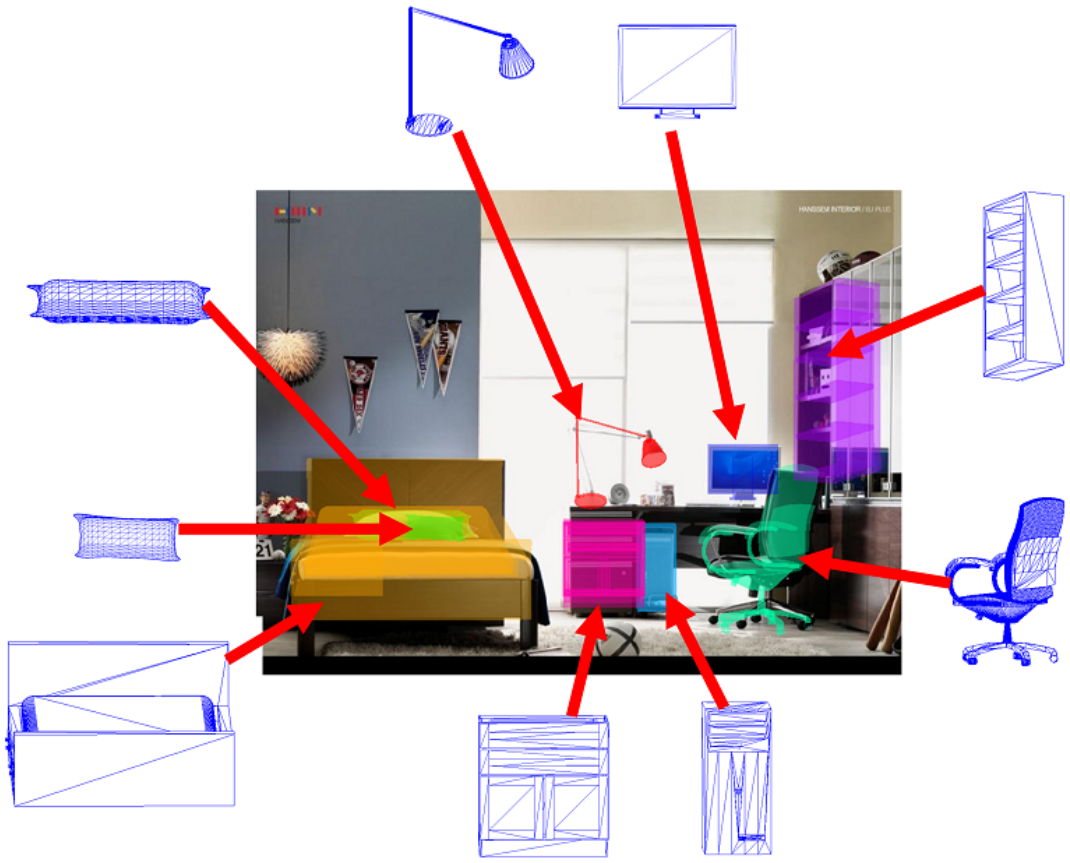
$$P, R, \vec{\alpha}, T$$

Approach I: Using 3D Object Labels



ObjectNet3D [Xiang et al, 16]

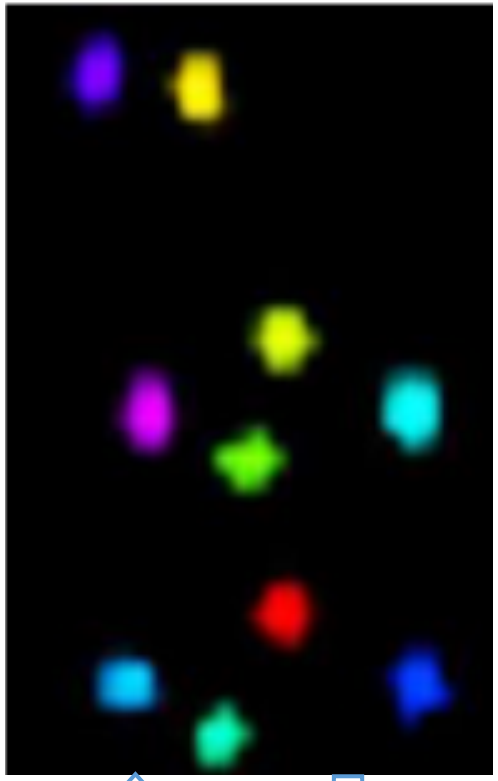
Approach II: Using 3D Synthetic Data



ObjectNet3D [Xiang et al, 16]

Render for CNN [Su et al, '15]
Multi-view CNNs [Dosovitskiy et al, '16]
TL network [Girdhar et al, '16]
PhysNet [Lerer et al, '16]

Intermediate 2D Representation

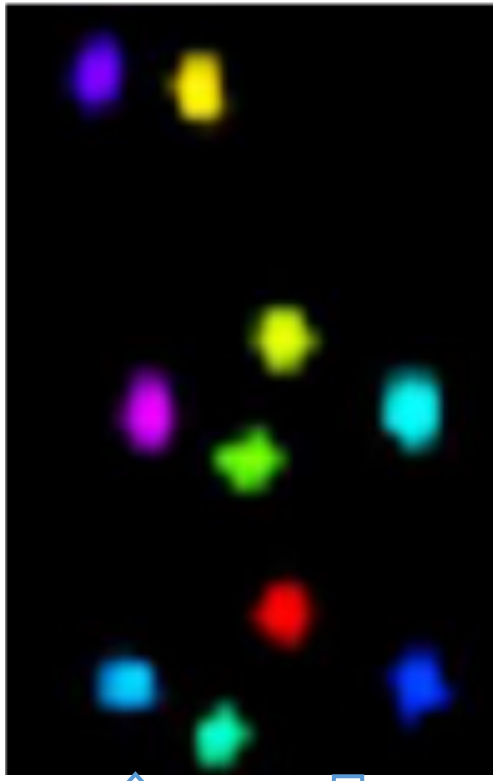


Real images with
2D keypoint labels

Synthetic
3D models

Only 2D labels!

3D INterpreter Network (3D-INN)



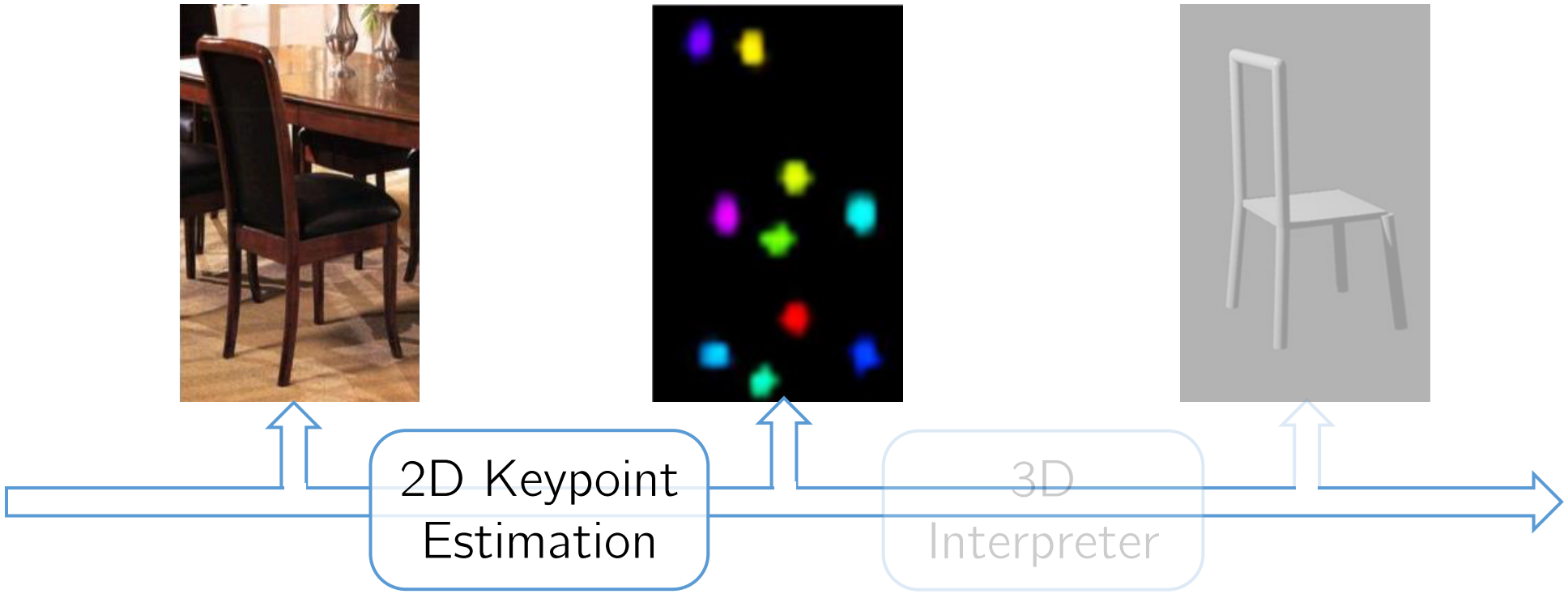
Real images with
2D keypoint labels

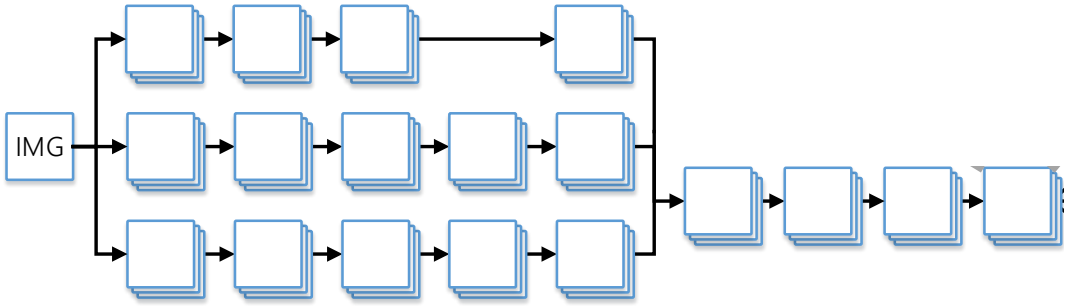
Synthetic
3D models

Only 2D labels!

Ramakrishna et al. '12
Grinciunaite et al. '13

3D-INN: Image to 2D Keypoints

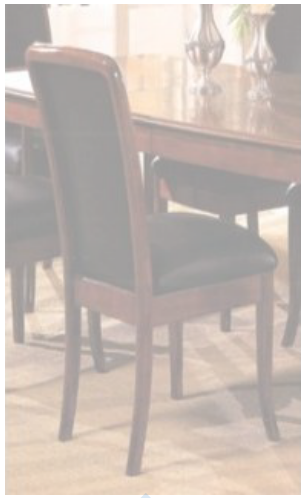


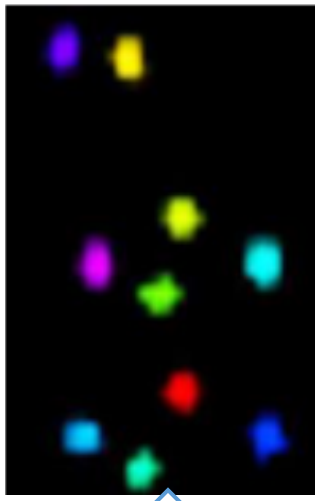


Inspired by Tompson et al. '15

Using 2D-annotated real data
Input: an RGB image
Output: keypoint heatmaps

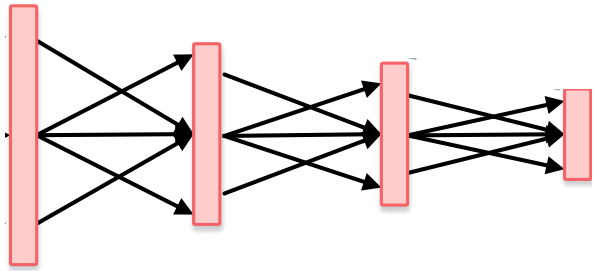
3D-INN: 2D Keypoints to 3D Skeleton





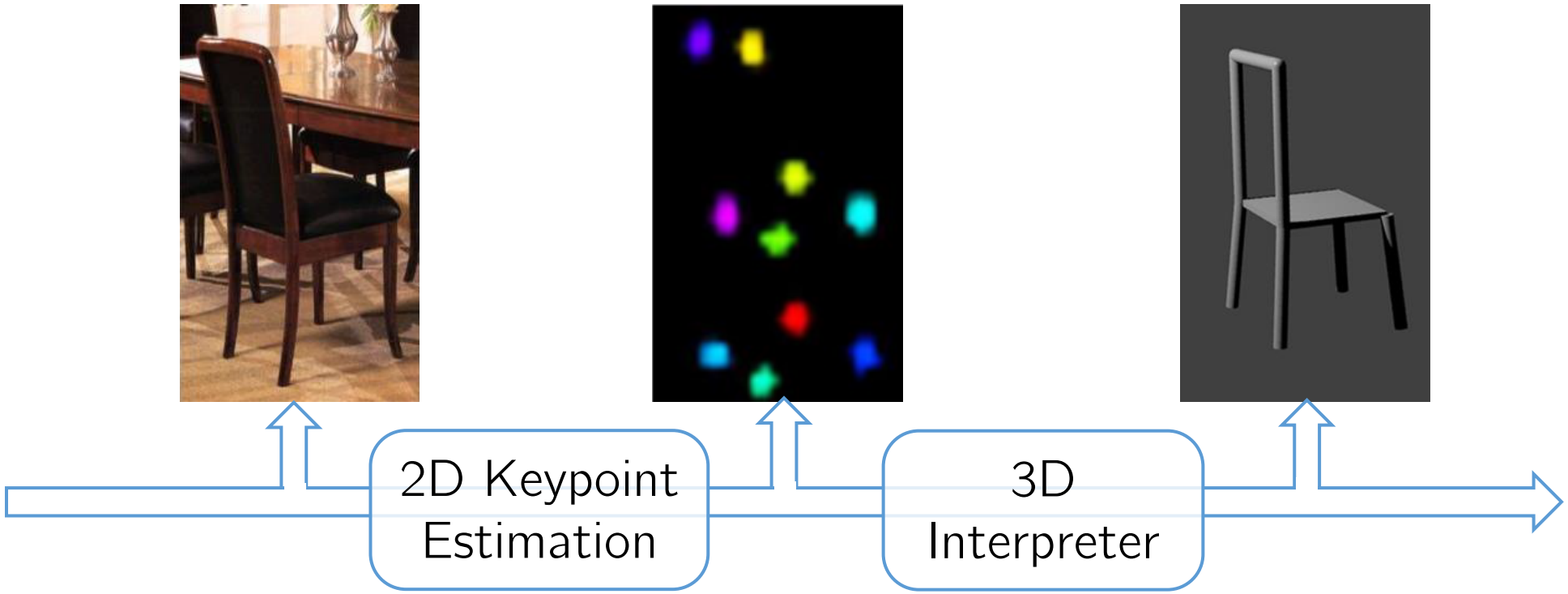
2D Keypoint Estimation

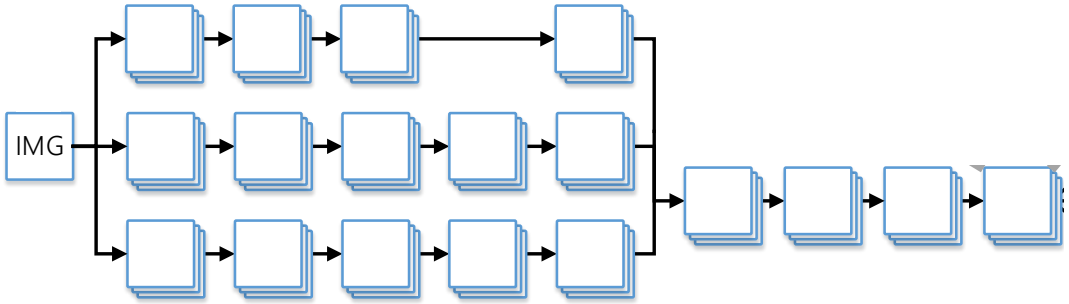
3D Interpreter

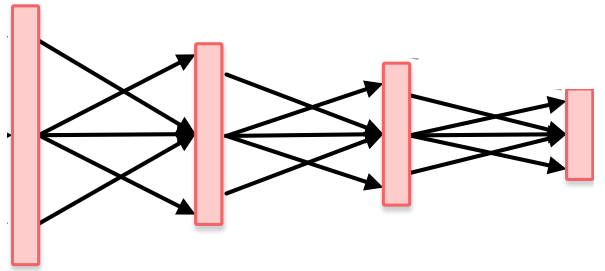


Using 3D synthetic data
Input: rendered keypoint heatmaps
Output: 3D parameters $\{P, R, \vec{\alpha}, T\}$

3D-INN: Initial Design



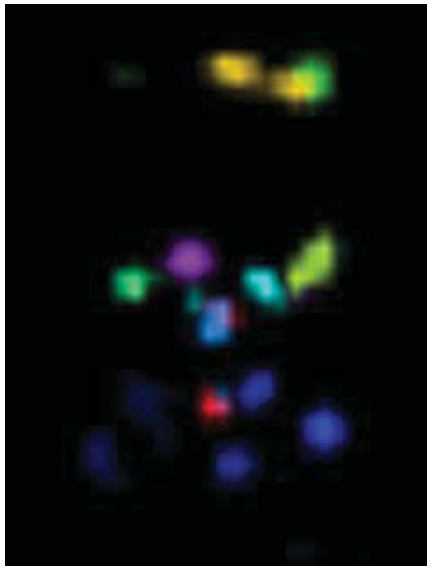




Initial Results

Image

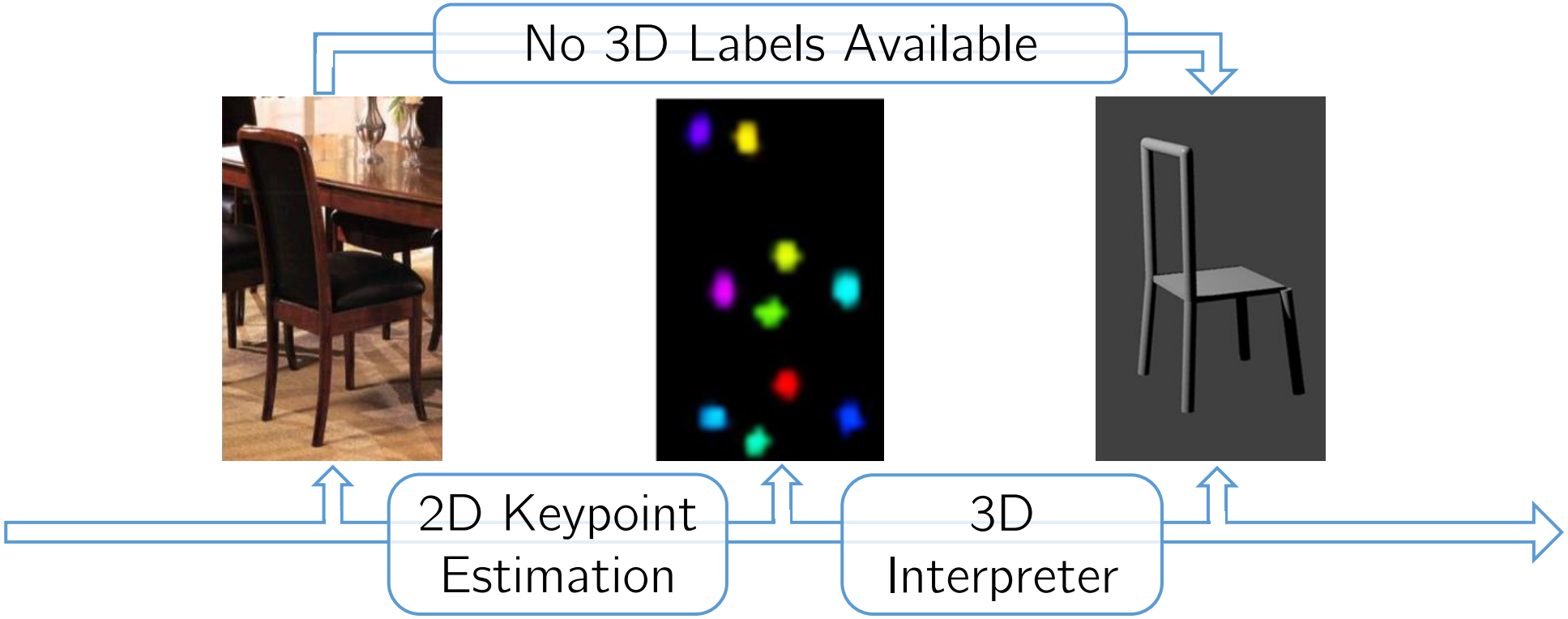
Inferred Keypoint Heatmap



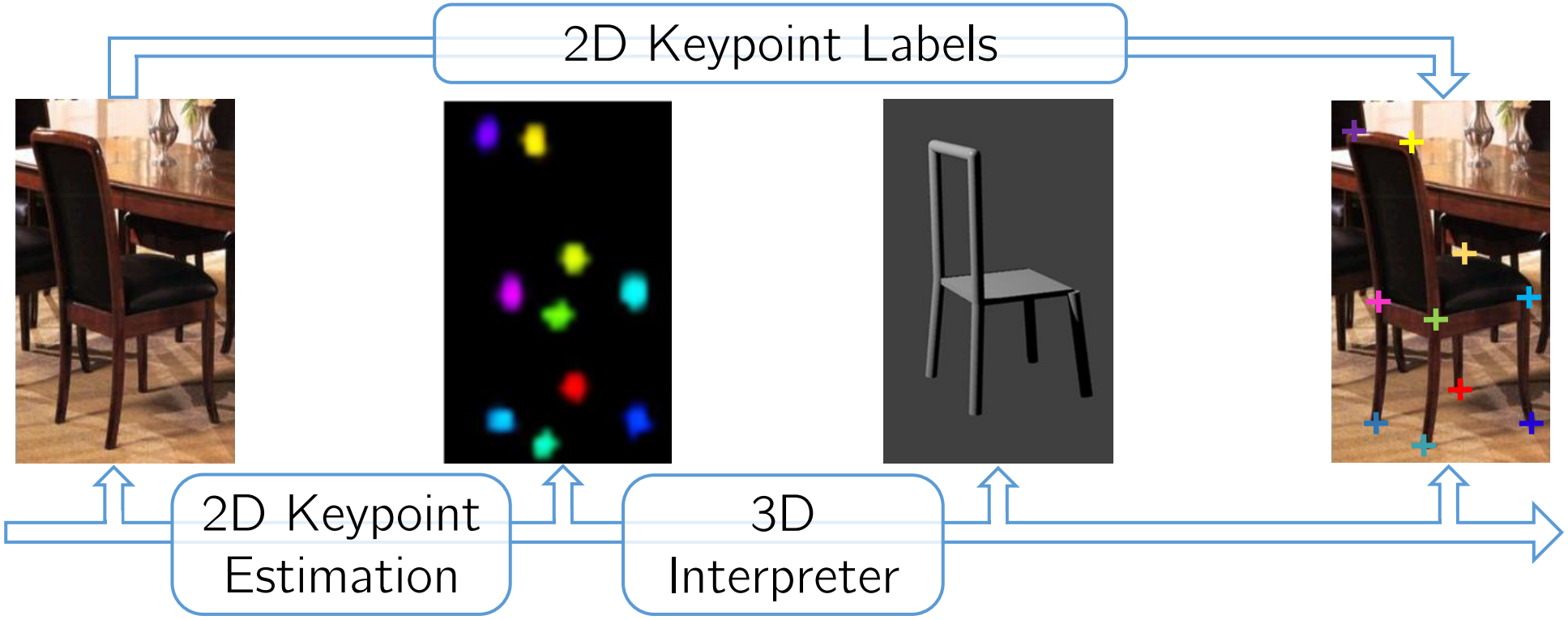
Inferred 3D Skeleton

Errors in the first stage propagate to the second

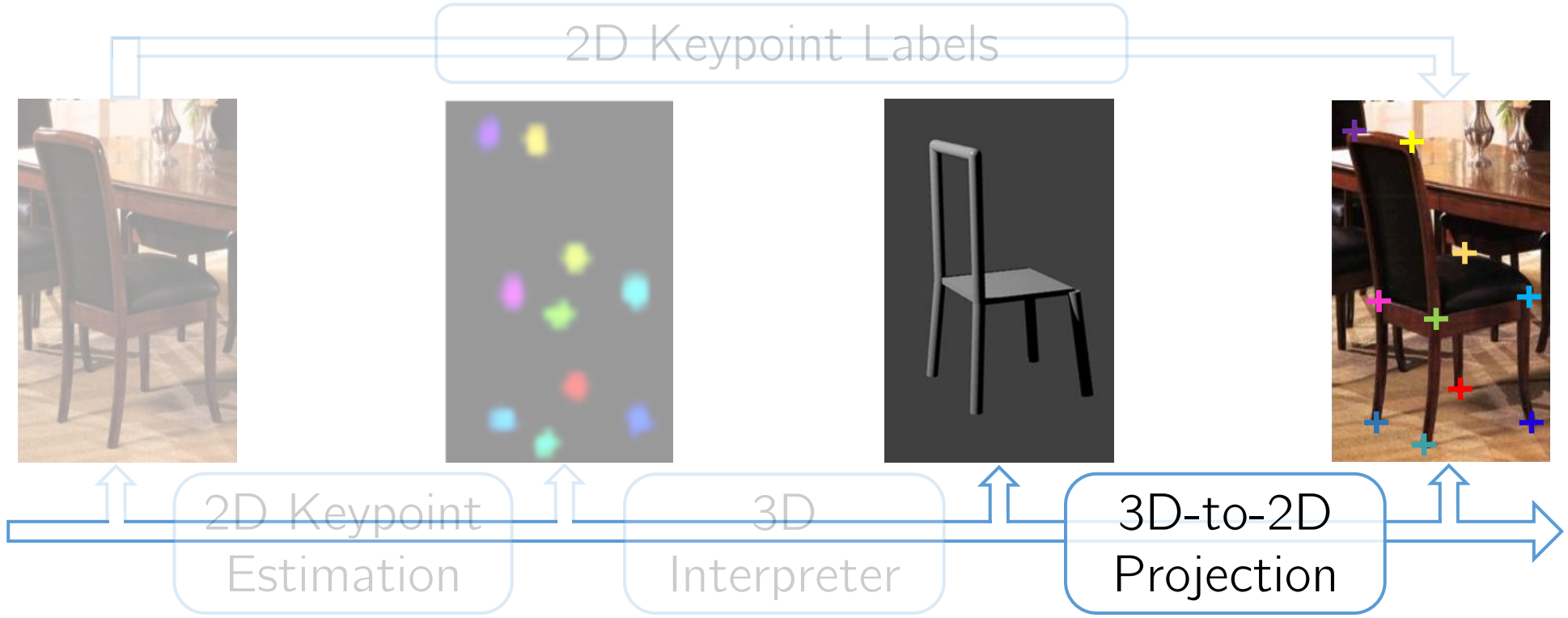
3D-INN: End-to-End Training?



3D-INN: End-to-End Training?



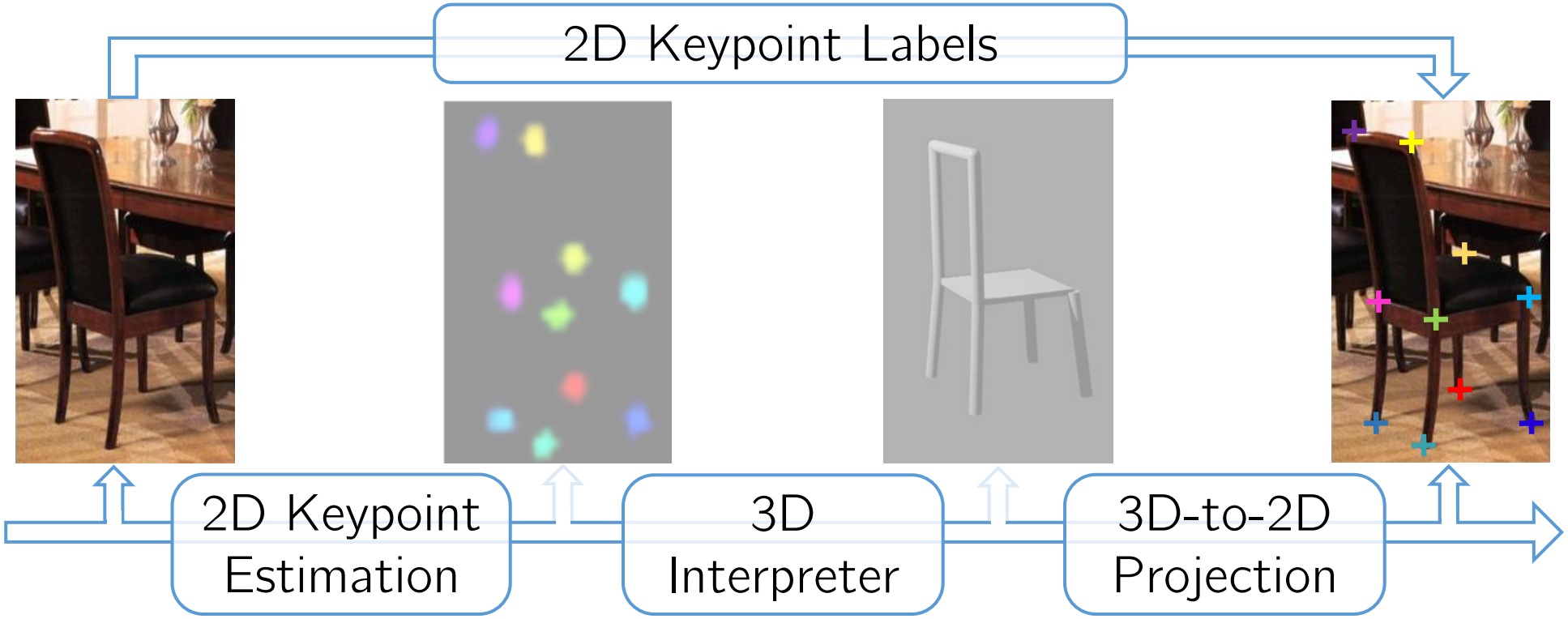
3D-INN: 3D-to-2D Projection Layer



$$P(R\sum_{k=1}^{K}\alpha_k B_k + T)$$

3D-to-2D projection is fully differentiable.

3D-INN: 3D-to-2D Projection Layer

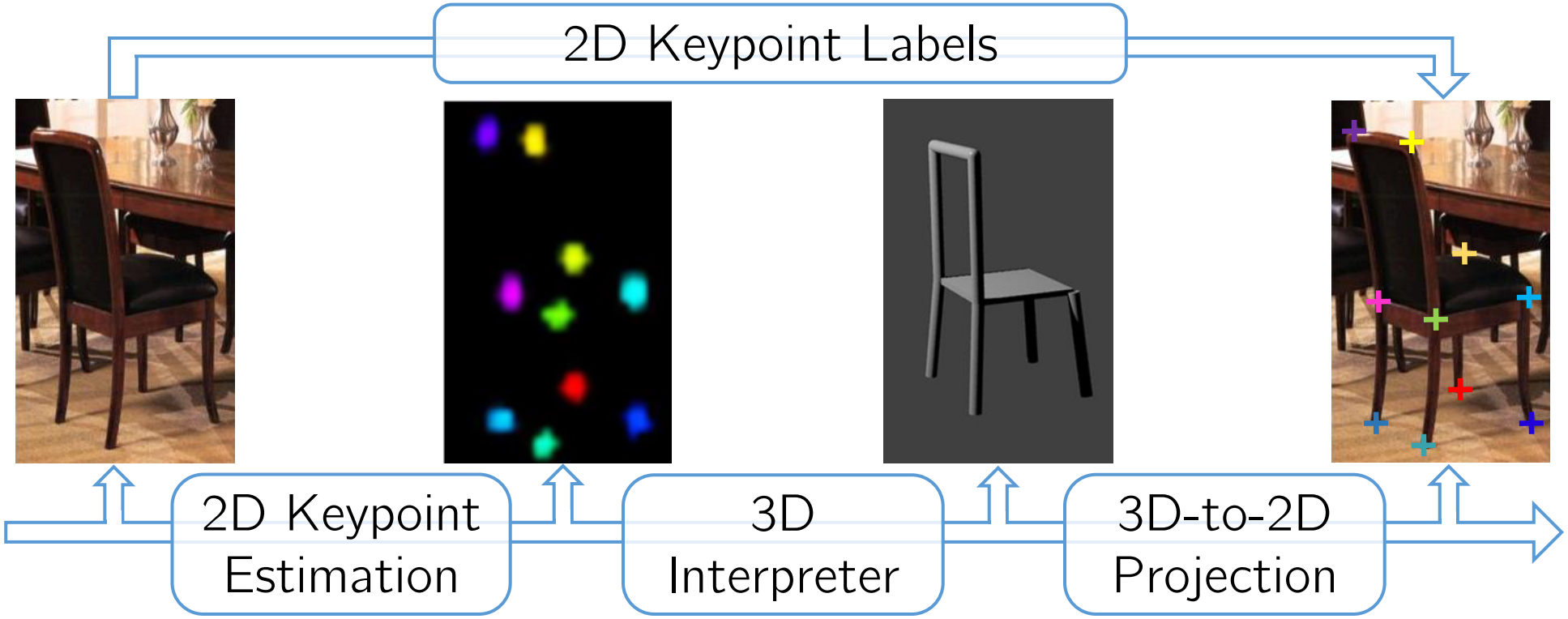


Using 2D-annotated real data
Input: an RGB image
Output: keypoint coordinates

Objective function:

$$\min \left\| P(R \sum_{k=1}^{K} \alpha_k B_k + T) - X_{2D} \right\|_2$$

3D-INN: Training Paradigm



Three-step training paradigm

II: 3D Interpreter

I: 2D Keypoint Estimation

III: End-to-end Finetuning

Refined Results

Image

Initial Estimation

After End-to-End Fine-tuning

3D Estimation: Qualitative Results

Training: our Keypoint-5 dataset, 2K images per category

Keypoint-5 dataset

3D Estimation: Qualitative Results

Training: our Keypoint-5 dataset, 2K images per category

IKEA Dataset [Lim et al, '13]

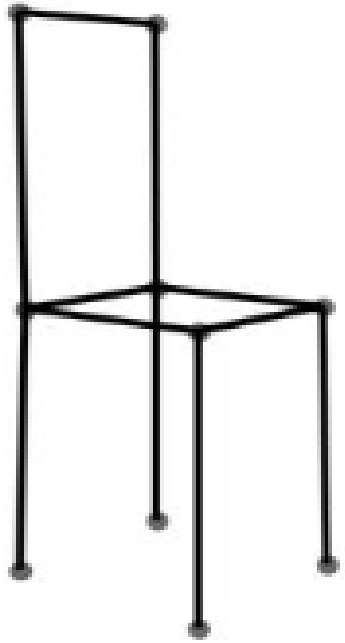
3D Estimation: Qualitative Results

Training: our Keypoint-5 dataset, 2K images per category

SUN Database [Xiao et al, '11]

3D Estimation: Qualitative Results

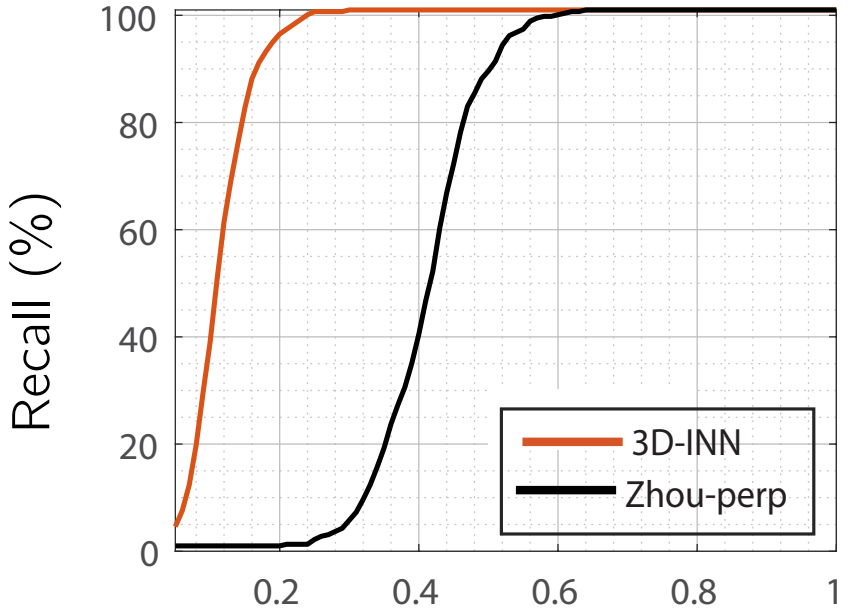
Training: our Keypoint-5 dataset, 2K images per category



SUN Database [Xiao et al, '11]

3D Structure Estimation

Images Results



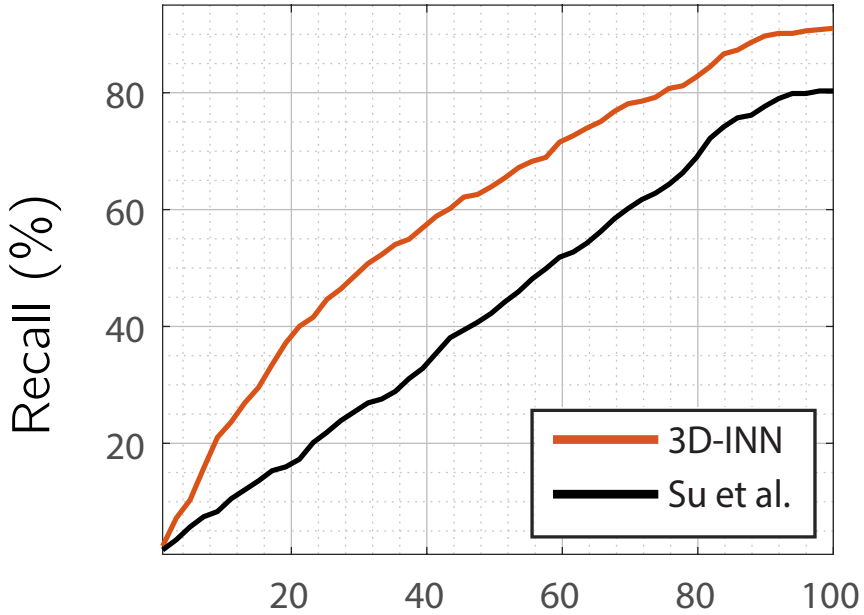
RMSE of estimated 3D keypoints on IKEA [Lim et al, '13]

Method	Bed	Sofa	Chair	Avg.
3D-INN	88.6	88.0	87.8	88.0
Zhou, '16	52.3	58.0	60.8	58.5

Average recall (%)

Viewpoint Estimation

Images Results

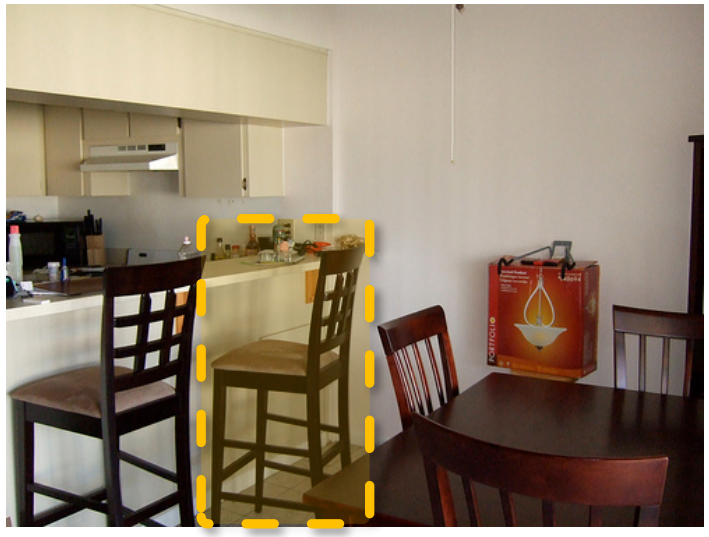


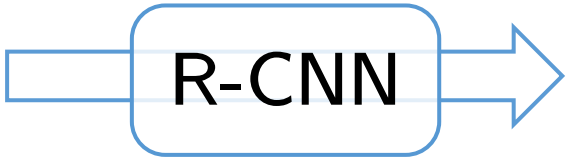
Azimuth angular error on IKEA [Lim et al, '13]

Method	Table	Sofa	Chair	Avg.
3D-INN	55.0	64.7	63.5	60.3
Su, '15	52.7	35.7	37.7	43.3

Average recall (%)

Localization and Viewpoint Estimation



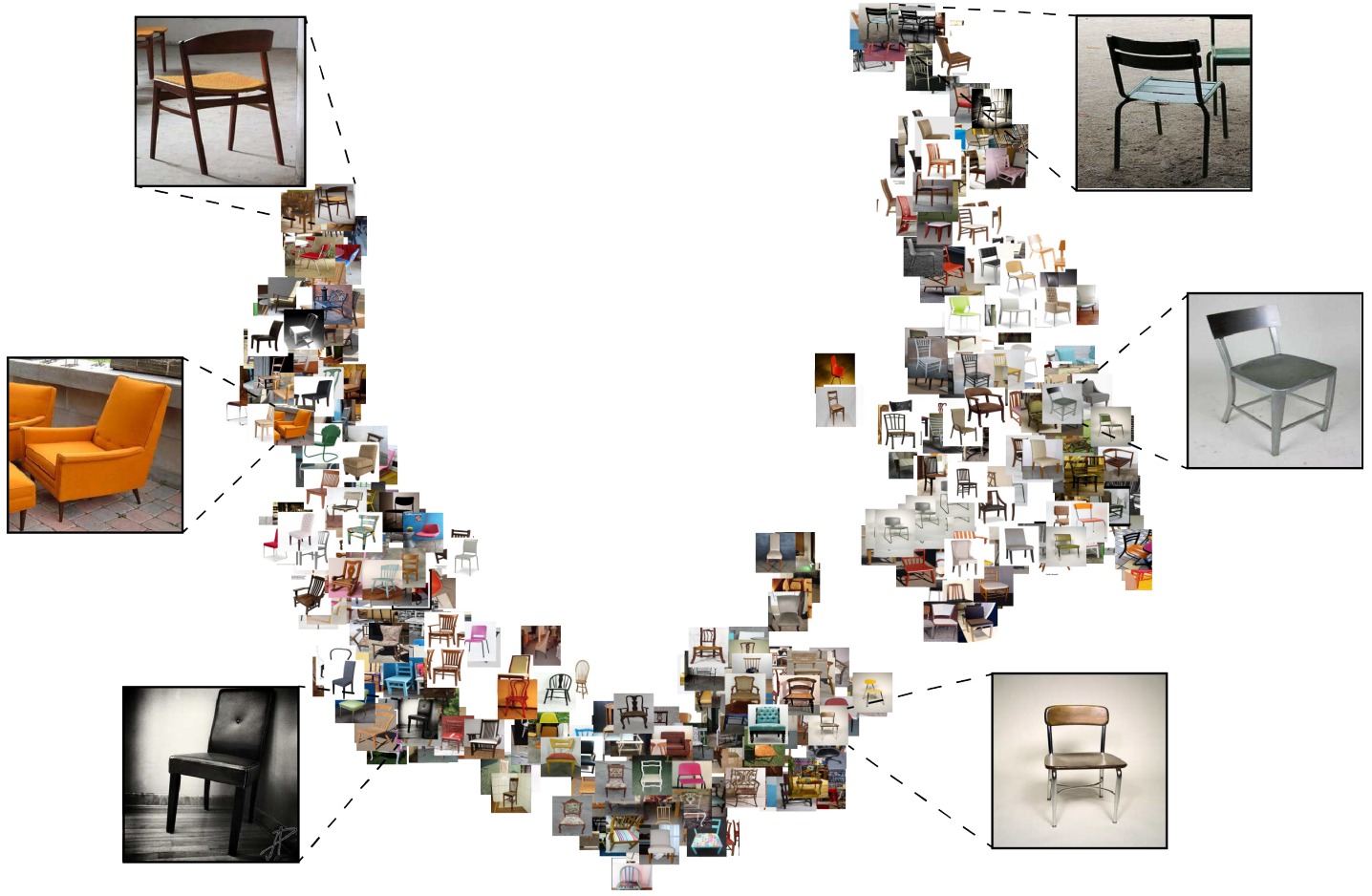


Girshick et al, '14

Category	VDPM	DPM+VP	Su et al.	V & K	3D-INN
Chair	6.8	6.1	15.7	**25.1**	23.1
Sofa	5.1	11.8	18.6	43.8	**45.8**

Viewpoint estimation on the PASCAL 3D+ dataset [Xiang et al, '14]

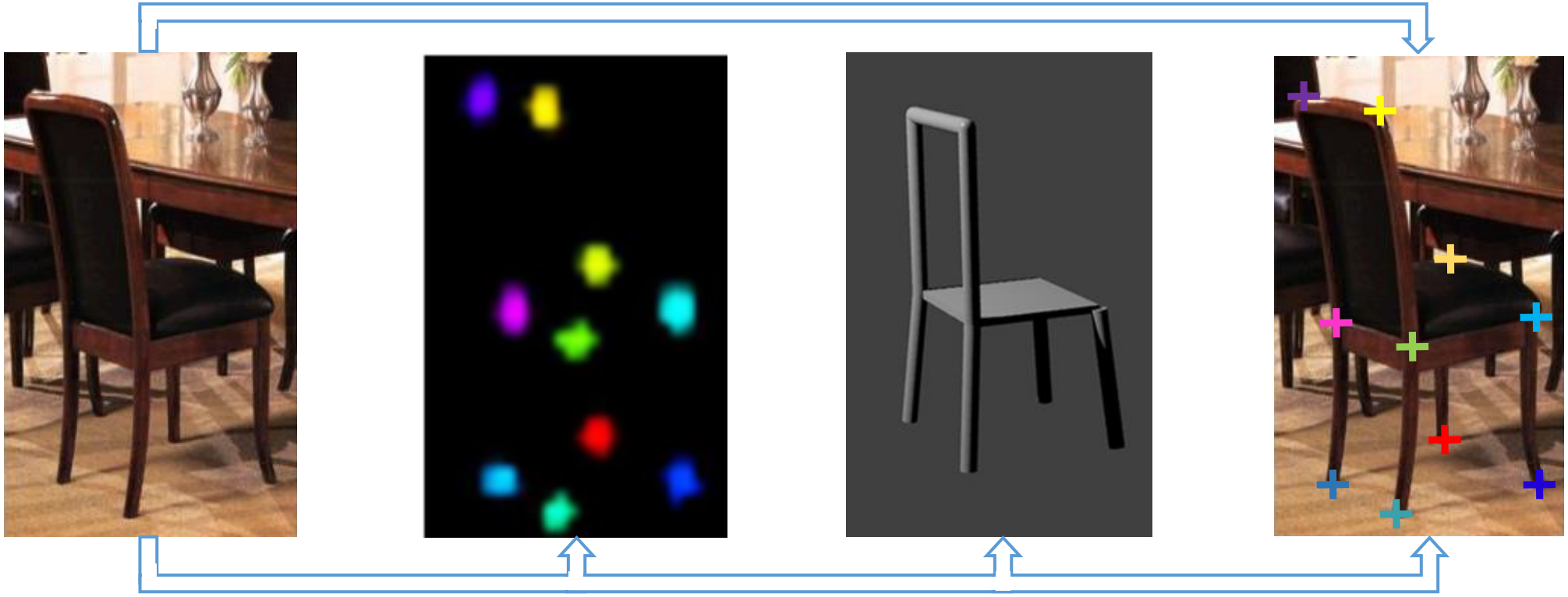
Chair Embedding



Manifold of chairs based on their inferred viewpoint

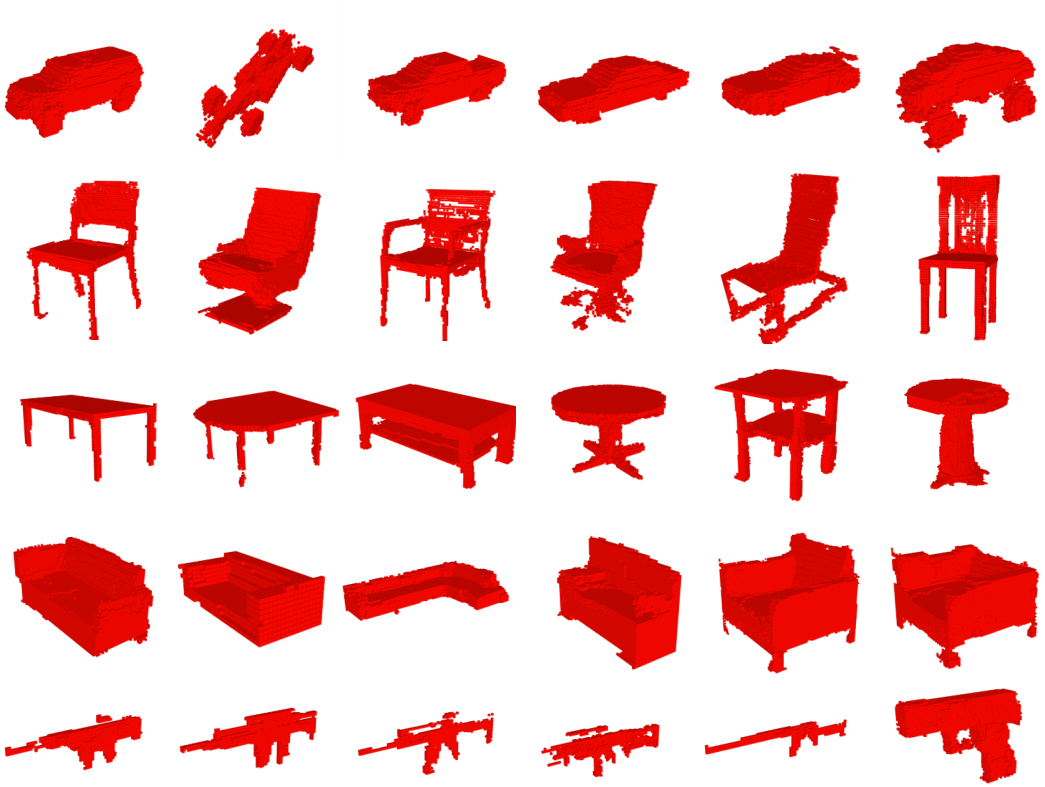
Contributions of 3D-INN

- Single image 3D perception
  - Real 2D labels + synthetic 3D models, connected via keypoints
  - A 3D-to-2D projection layer for end-to-end training



Summary

3D-GAN: Synthesizing 3D shapes



3D-INN: Recognizing 3D structure