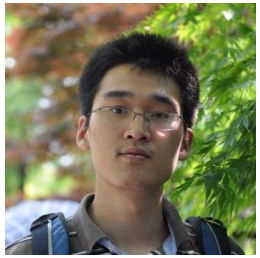
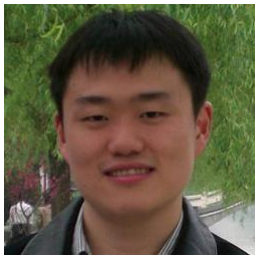
Single Image 3D Interpreter Network

Jiajun Wu*



Tianfan Xue*

Joseph Lim



Yuandong Tian

Josh Tenenbaum

Antonio Torralba

Bill Freeman

(* equal contributions)

What do we see from these images?

What do we see from these images?

What do we see from these images?

Motivation

- What do we see from these images?

Motivation

- What do we see from these images?
  - 3D object structure
  - 3D object pose/viewpoint
  - Appearance/texture

Motivation

- What do we see from these images?
  - 3D object structure
  - 3D object pose/viewpoint
  - Appearance/texture

- Humans see rich 3D information from a single image.

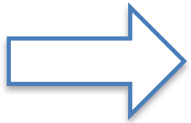
Motivation

- What do we see from these images?
  - 3D object structure
  - 3D object pose/viewpoint
  - Appearance/texture

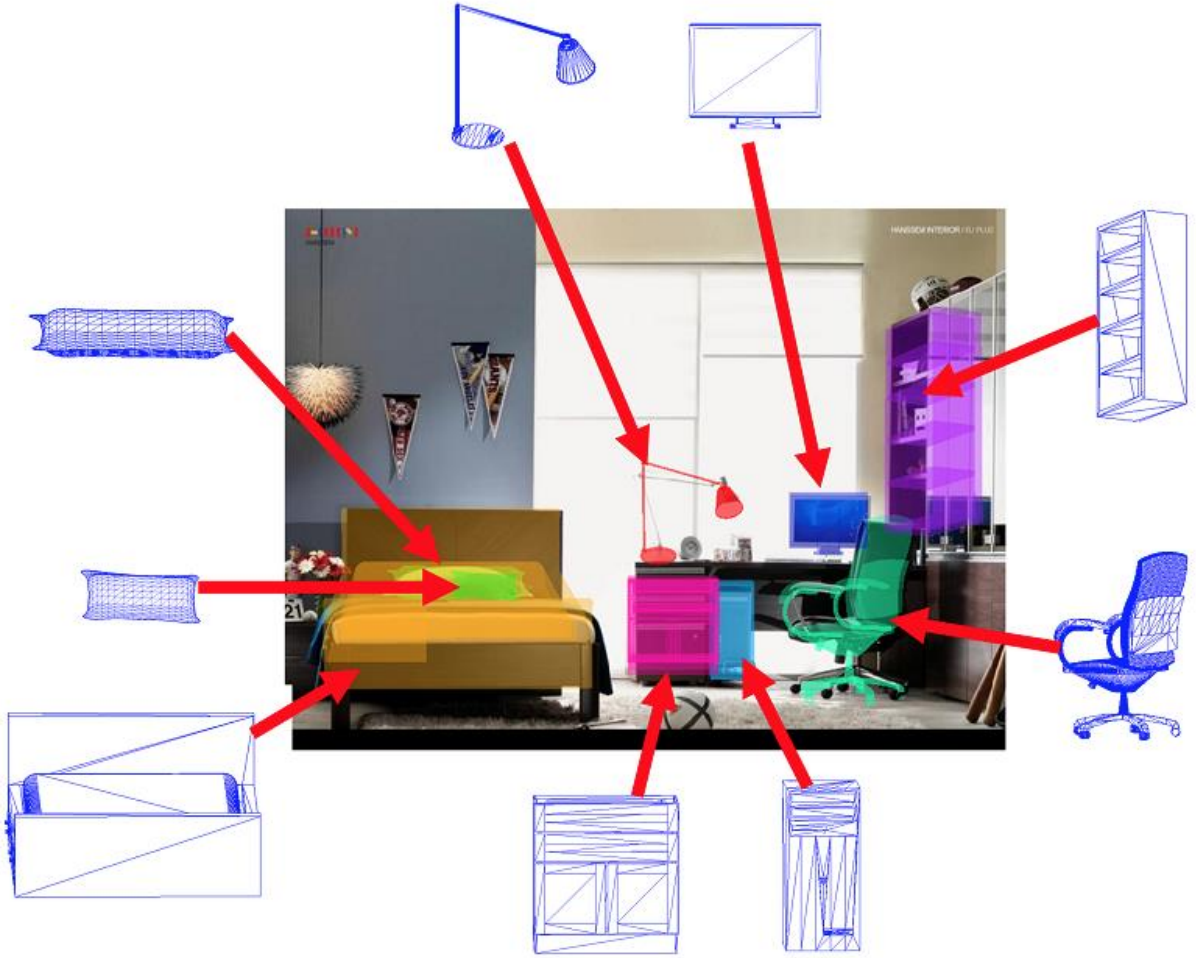
- Humans see rich 3D information from a single image.

Single Image 3D Perception

Single Image 3D Perception



Approach I: Using 3D Object Labels



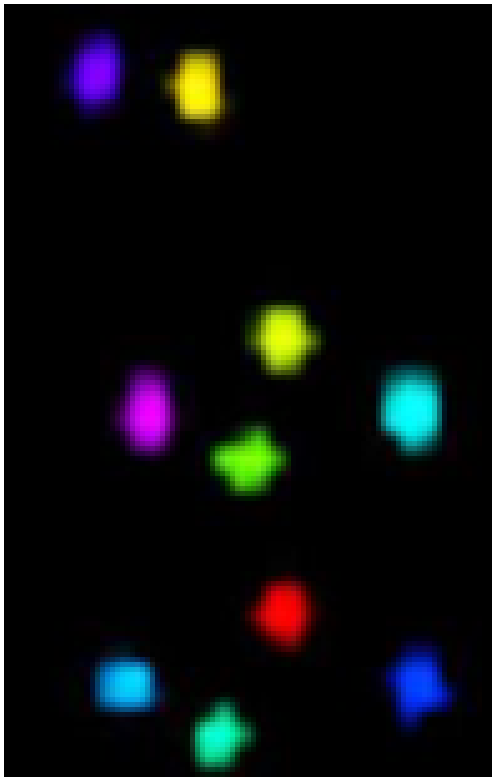
ObjectNet3D [*Xiang et al, 16*]

Approach II: Using 3D Synthetic Data

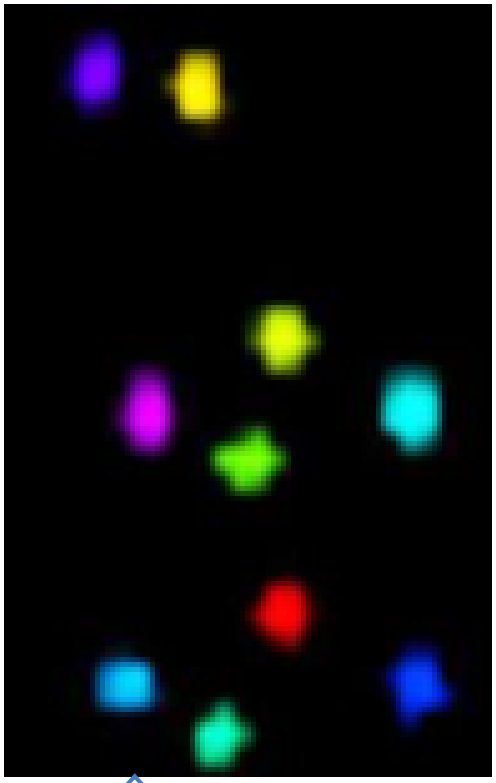
Render for CNN [*Su et al, ’15*]
Multi-view CNNs [*Dosovitskiy et al, ’16*]
TL network [*Girdhar et al, ’16*]
PhysNet [*Lerer et al, ’16*]

Our Approach

Intermediate 2D Representation

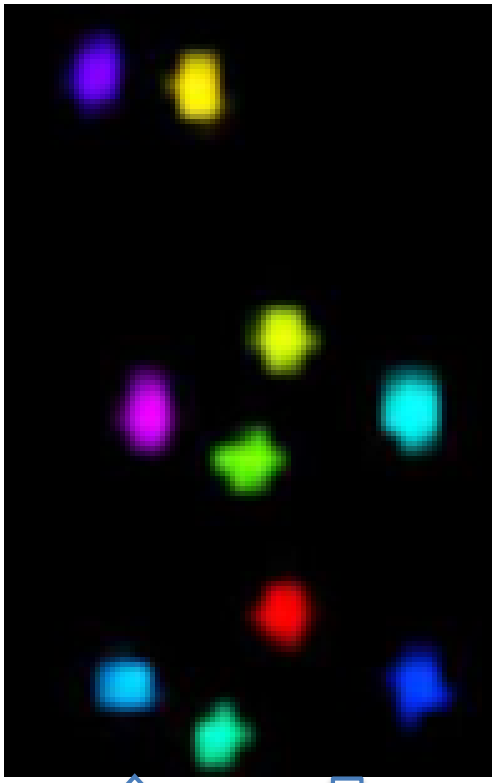


Intermediate 2D Representation



Real images with
2D keypoint labels

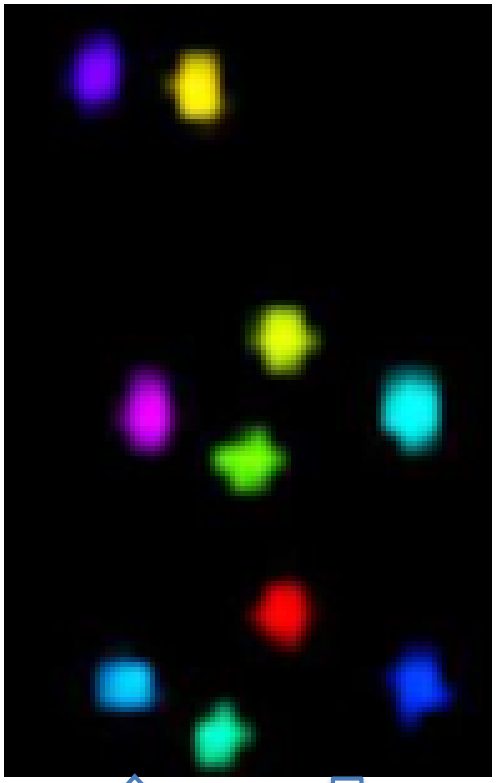
Intermediate 2D Representation



Real images with
2D keypoint labels

Synthetic
3D models

Intermediate 2D Representation



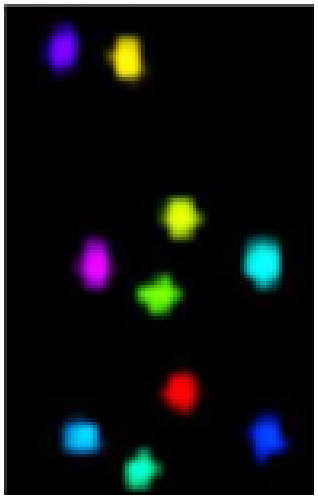
Real images with 2D keypoint labels

Synthetic 3D models

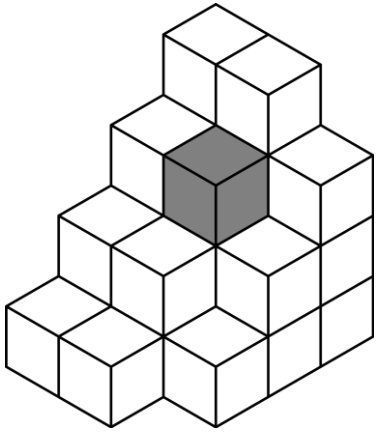
Only 2D labels!

Contribution I

- Real 2D labels + synthetic 3D models
- Keypoints as intermediate representations
- A 3D-to-2D projection layer for end-to-end training



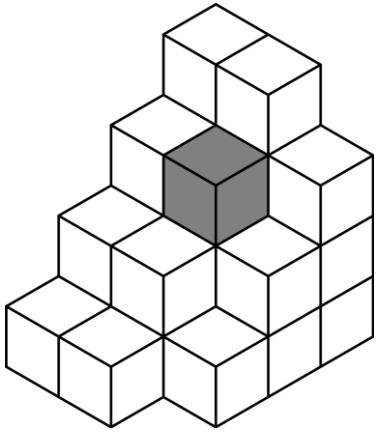
3D Object Representation



Voxel

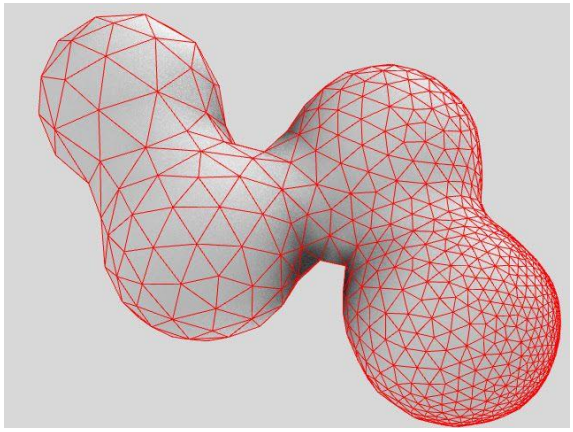
Girdhar et al. ′16
Choy et al. ′16
Xiao et al. ′12

3D Object Representation



Voxel

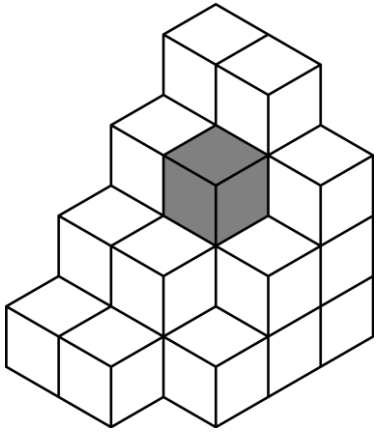
Girdhar et al. ’16
Choy et al. ’16
Xiao et al. ’12



Mesh

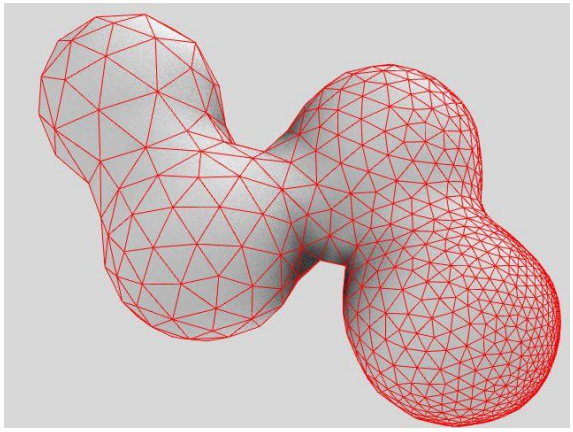
Lowe, ’91

3D Object Representation



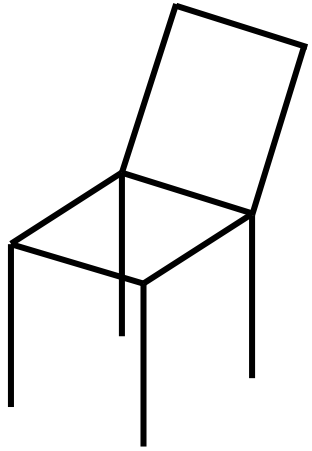
Voxel

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



Mesh

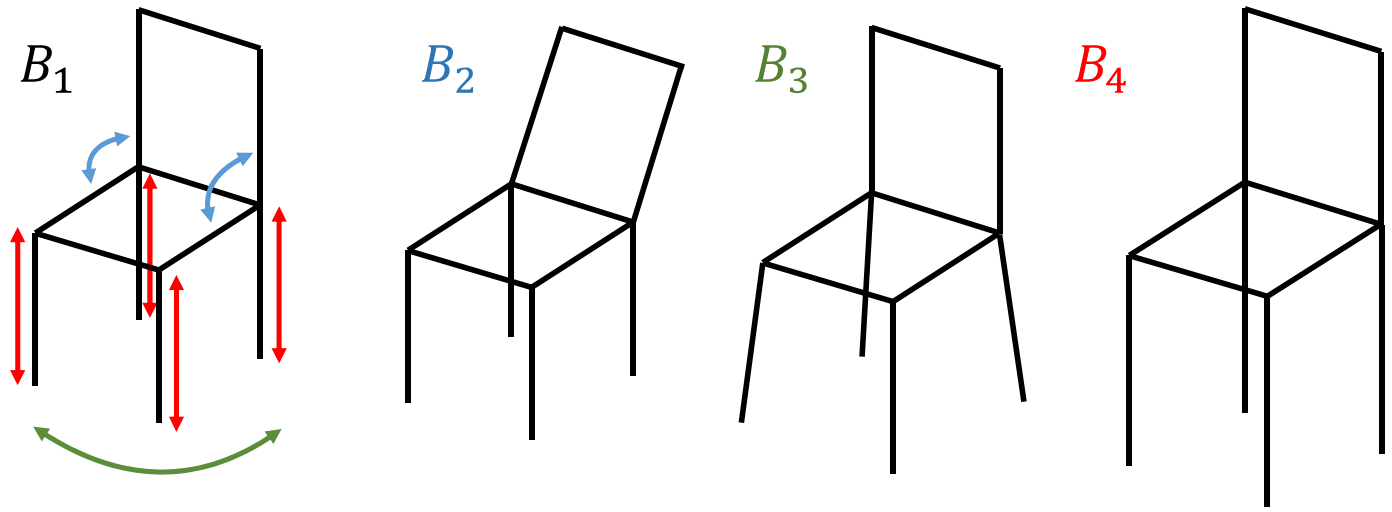
Lowe, ’91



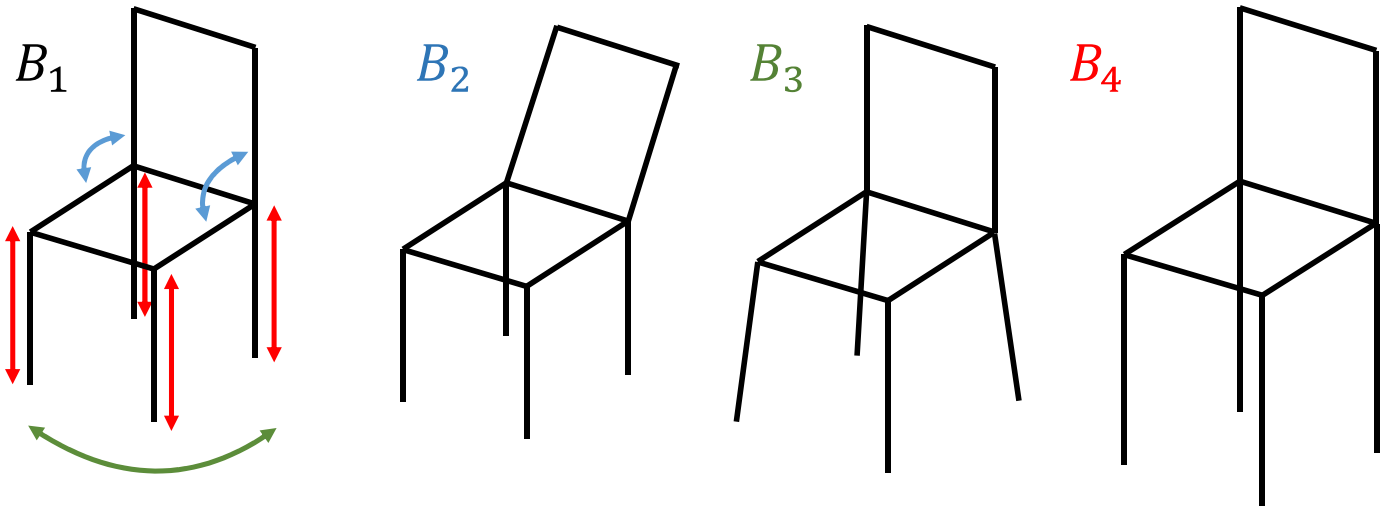
Skeleton

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

Skeleton Representation



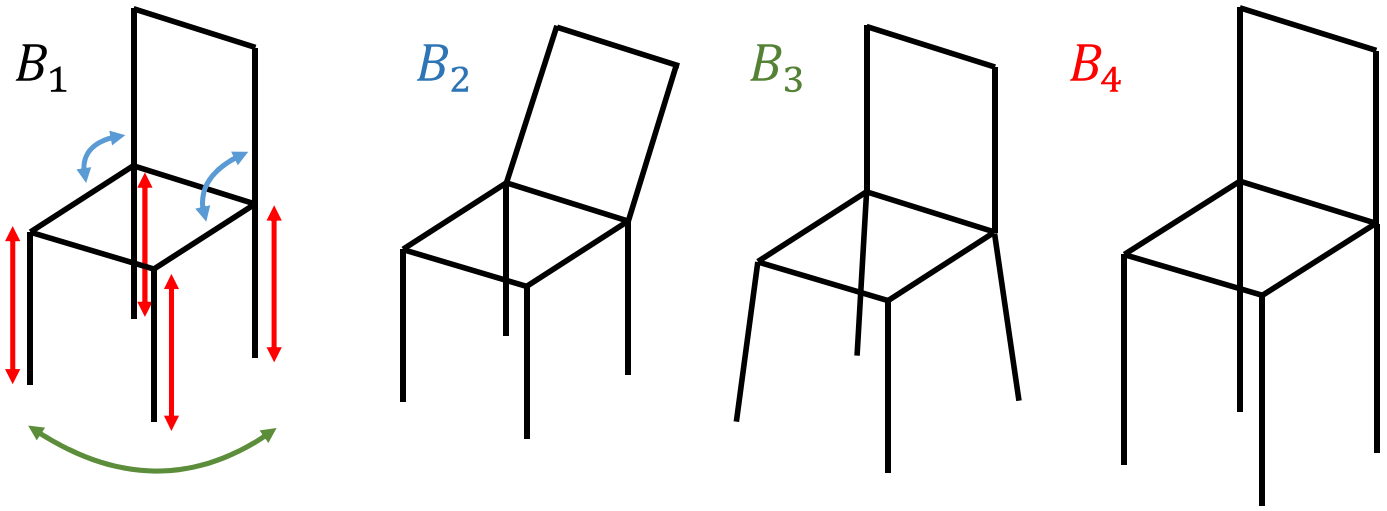
Skeleton Representation



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

structure parameter

3D Skeleton to 2D Keypoints

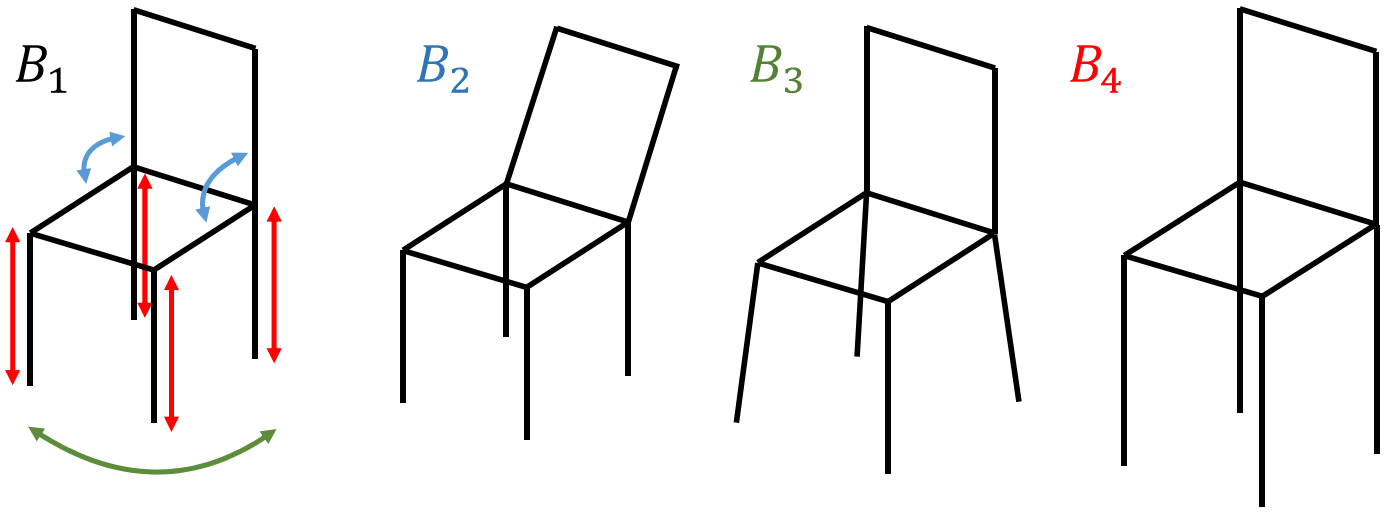


$$R \sum_{k=1}^{K} \alpha_k B_k$$

rotation

structure parameter

3D Skeleton to 2D Keypoints



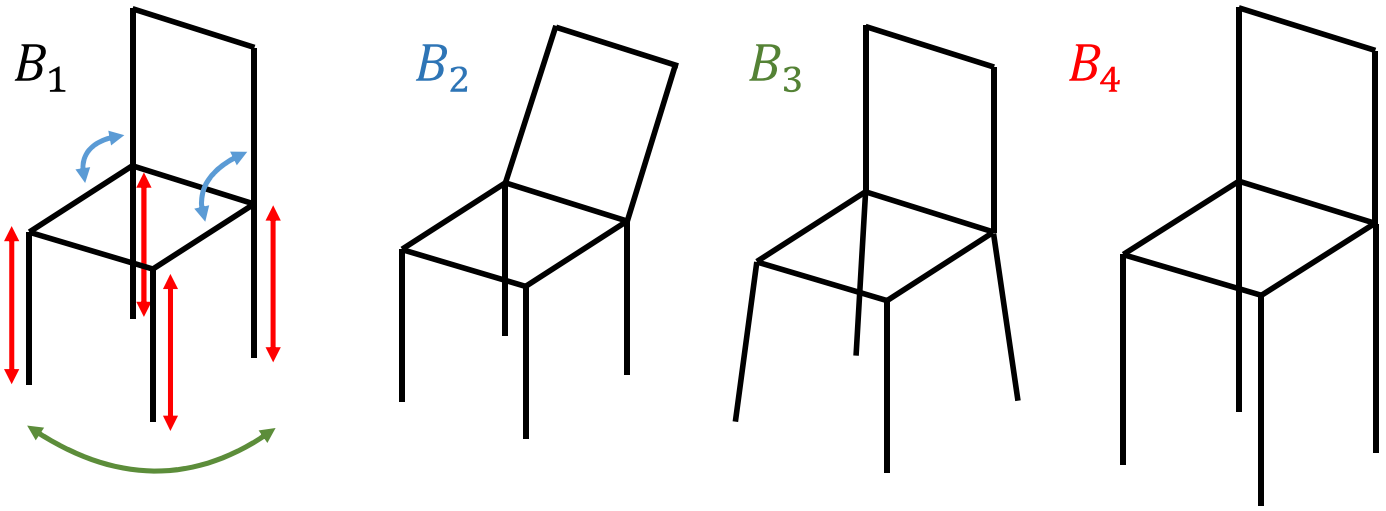
$$R\sum_{k=1}^{K}\alpha_k B_k + T$$

rotation

structure parameter

translation

3D Skeleton to 2D Keypoints



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

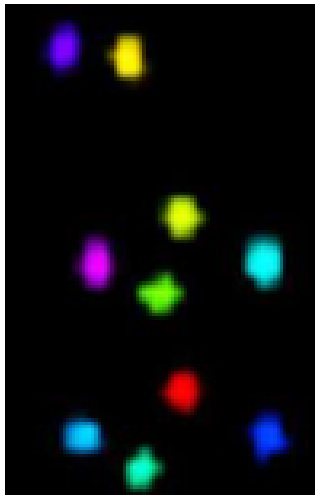
projection

rotation

structure parameter

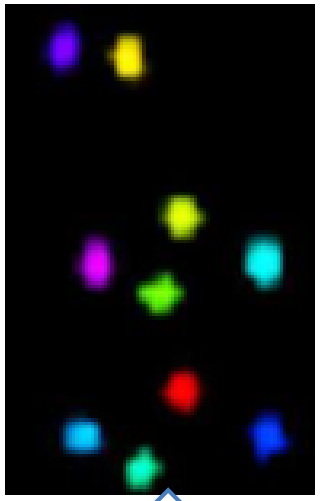
translation

3D INterpreter Network (3D-INN)



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

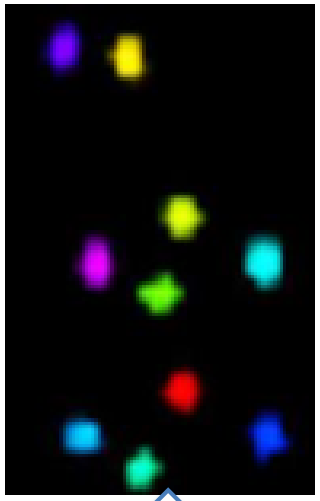
3D-INN: Image to Keypoint



2D Keypoint Estimation

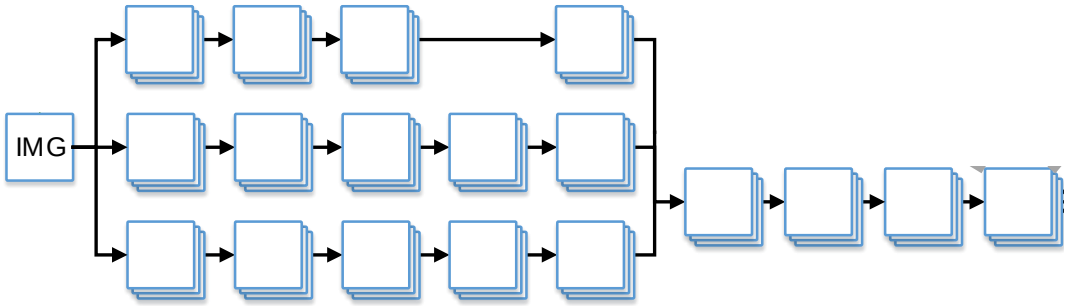
3D Interpreter

3D-INN: Image to Keypoint



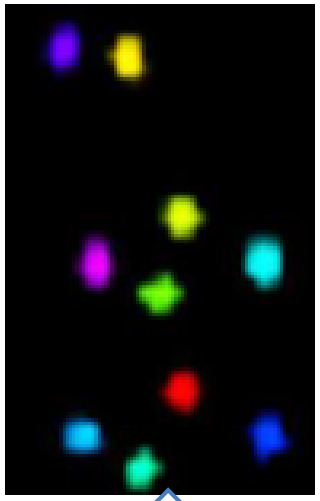
2D Keypoint Estimation

3D Interpreter



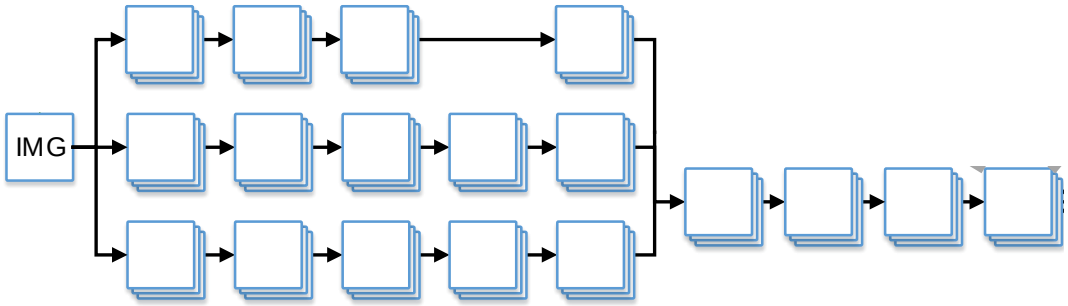
Inspired by [*Tompson et al. 15*]

3D-INN: Image to Keypoint



2D Keypoint Estimation

3D Interpreter



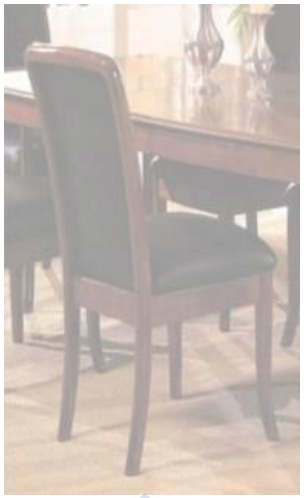
Inspired by [*Tompson et al. '15*]

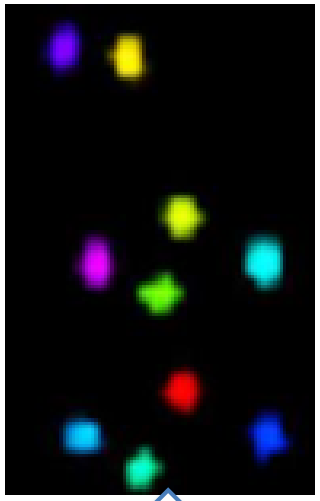
Using 2D-annotated real data

Input: an RGB image

Output: Keypoint heatmaps

3D-INN: Keypoint to 3D Skeleton

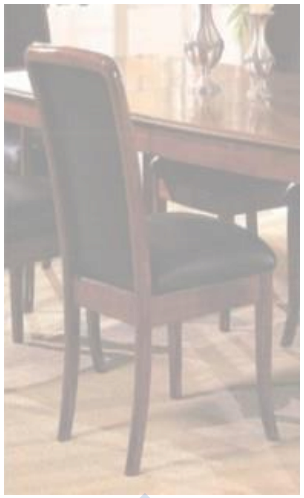


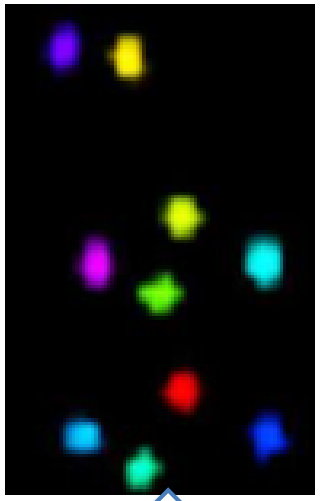


2D Keypoint Estimation

3D Interpreter

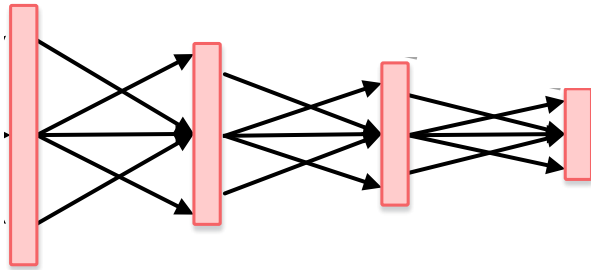
3D-INN: Keypoint to 3D Skeleton



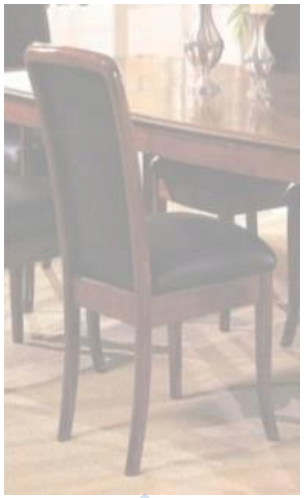


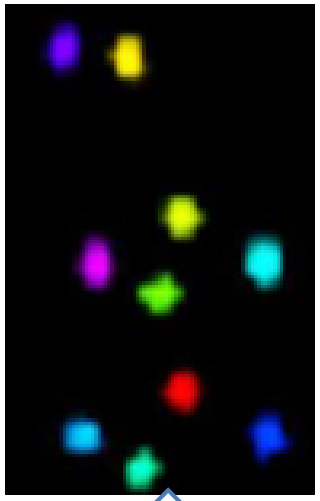
2D Keypoint Estimation

3D Interpreter



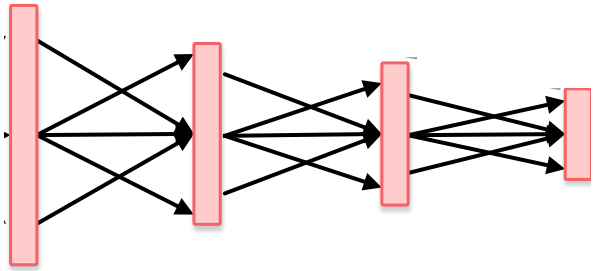
3D-INN: Keypoint to 3D Skeleton





2D Keypoint Estimation

3D Interpreter

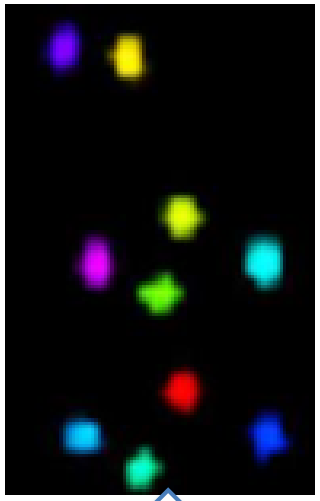


Using 3D synthetic data

Input: rendered keypoint heatmaps

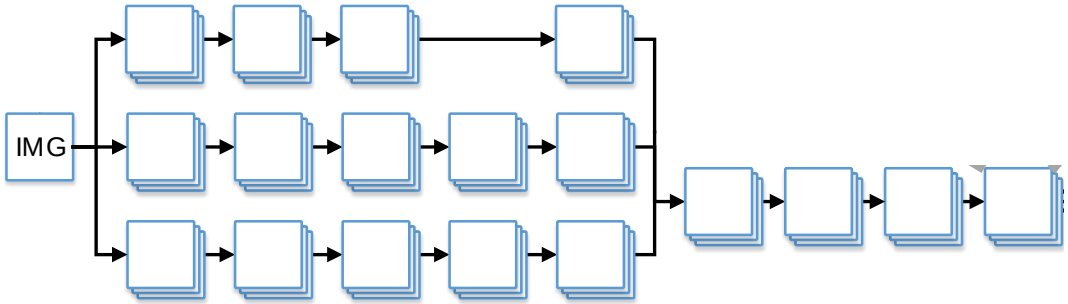
Output: 3D parameters $\{P, R, \vec{\alpha}, T\}$

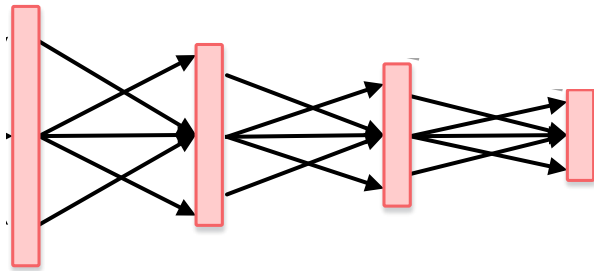
3D-INN: Initial Design



2D Keypoint Estimation

3D Interpreter





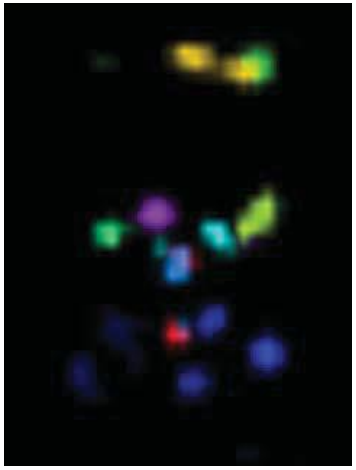
Initial Results

Image

Initial Results

Image

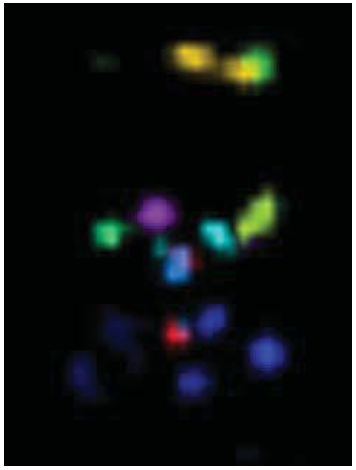
Inferred Keypoint Heatmap



Initial Results

Image

Inferred Keypoint Heatmap

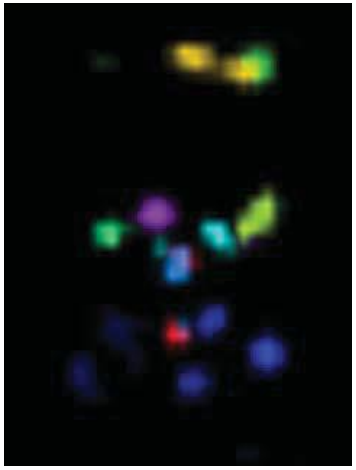


Inferred 3D Skeleton

Initial Results

Image

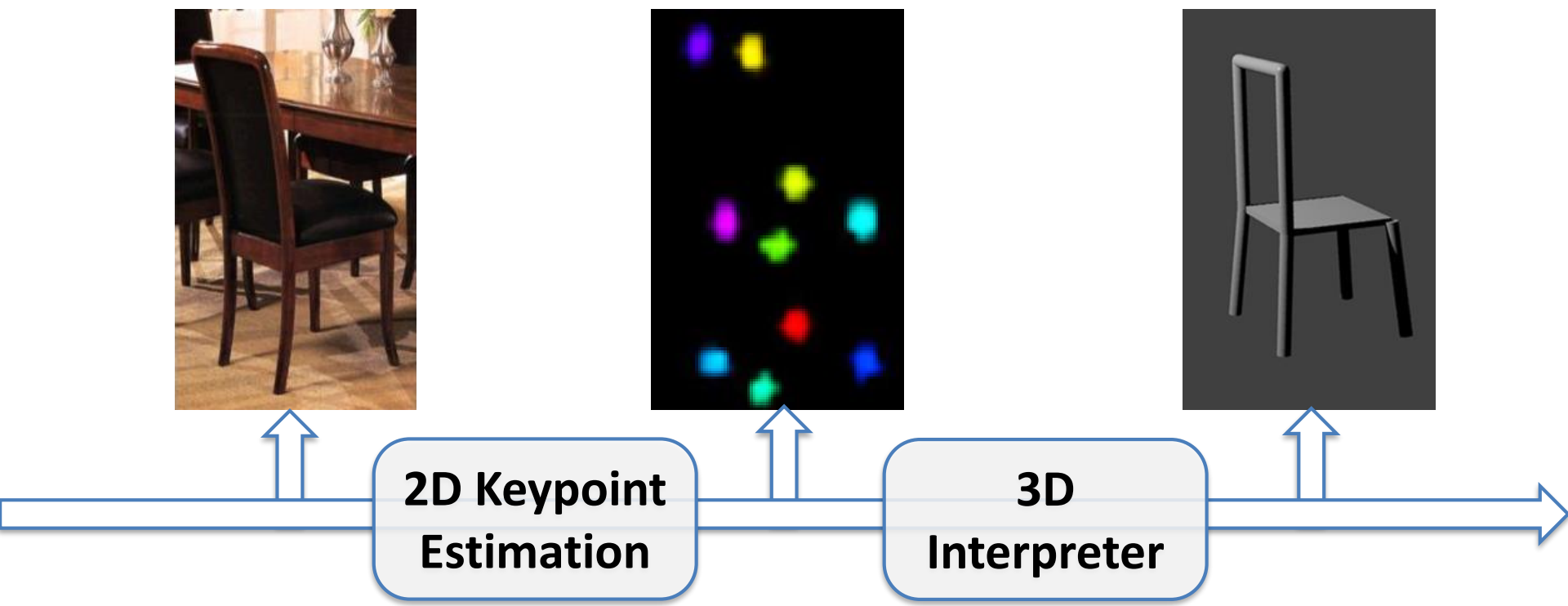
Inferred Keypoint Heatmap



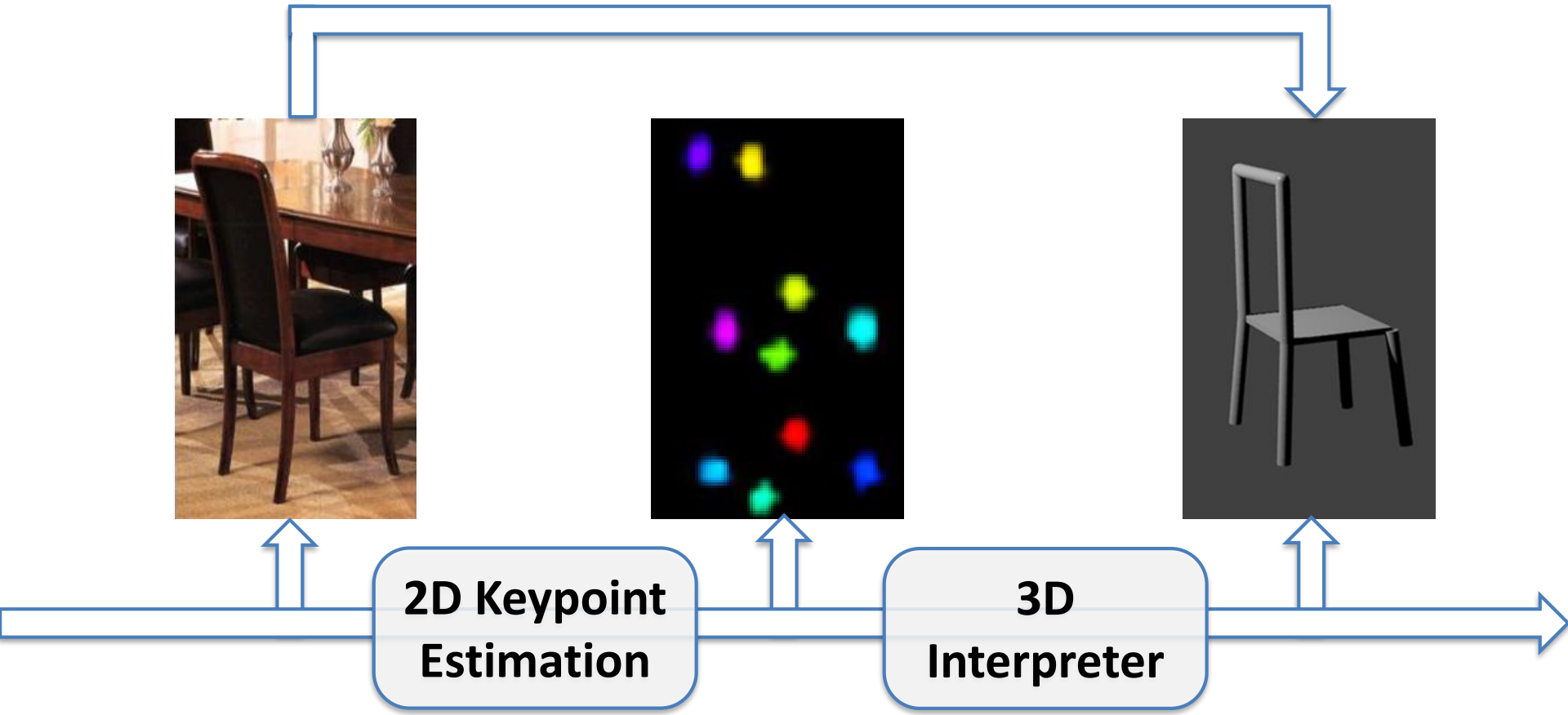
Inferred 3D Skeleton

Errors in the first stage propagate to the second

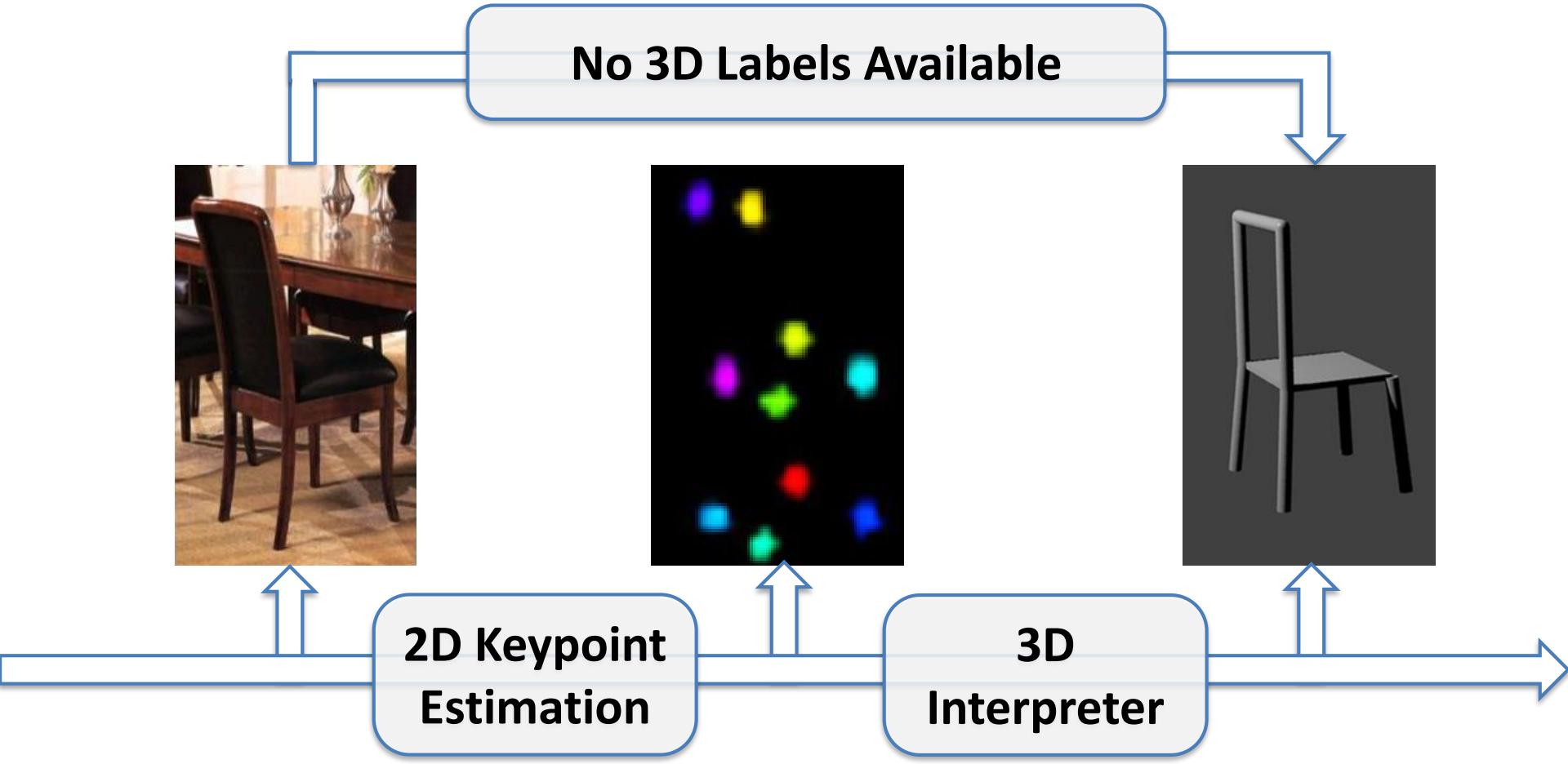
3D INterpreter Network (3D-INN)



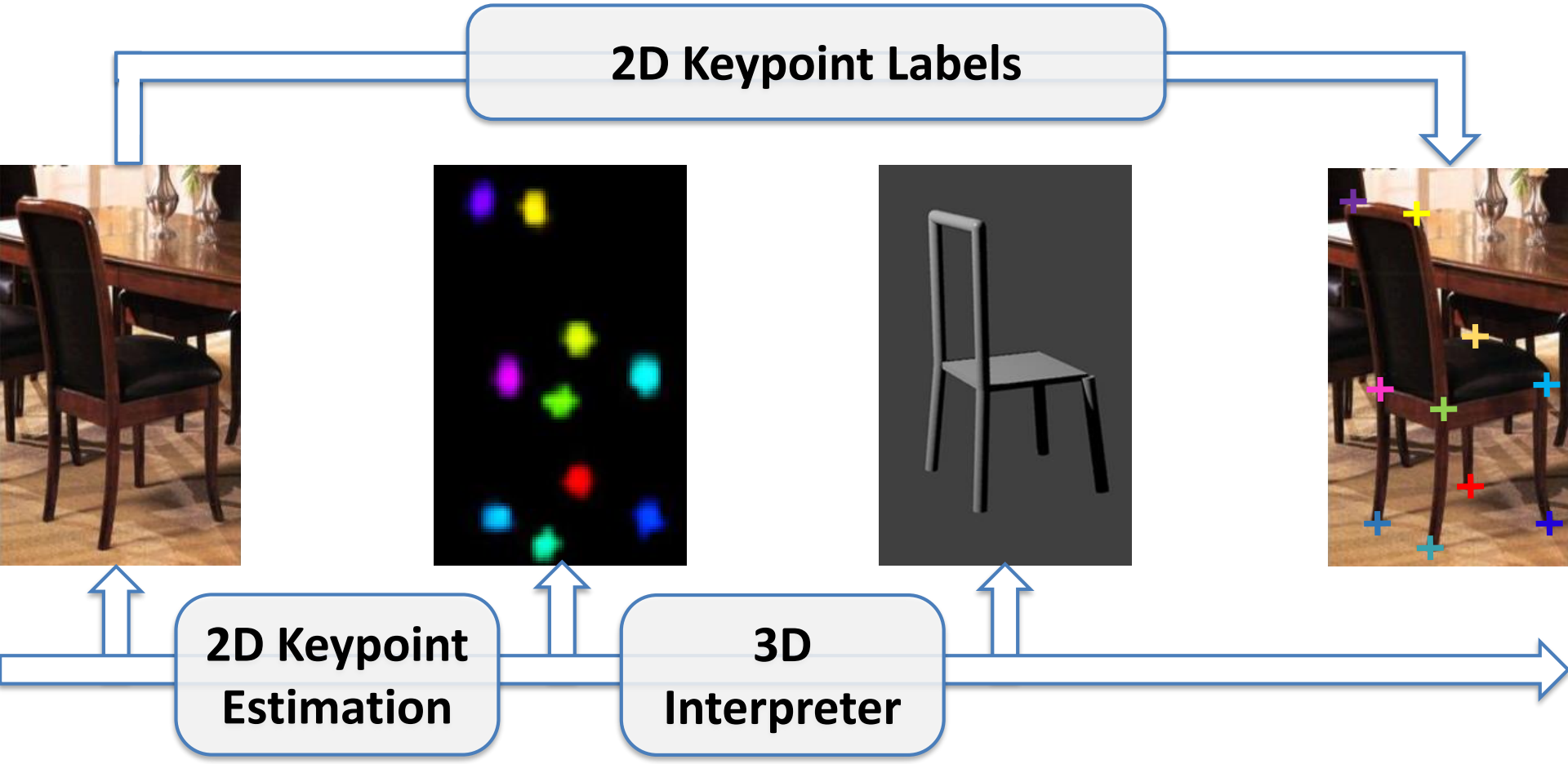
3D-INN: End-to-End Training?



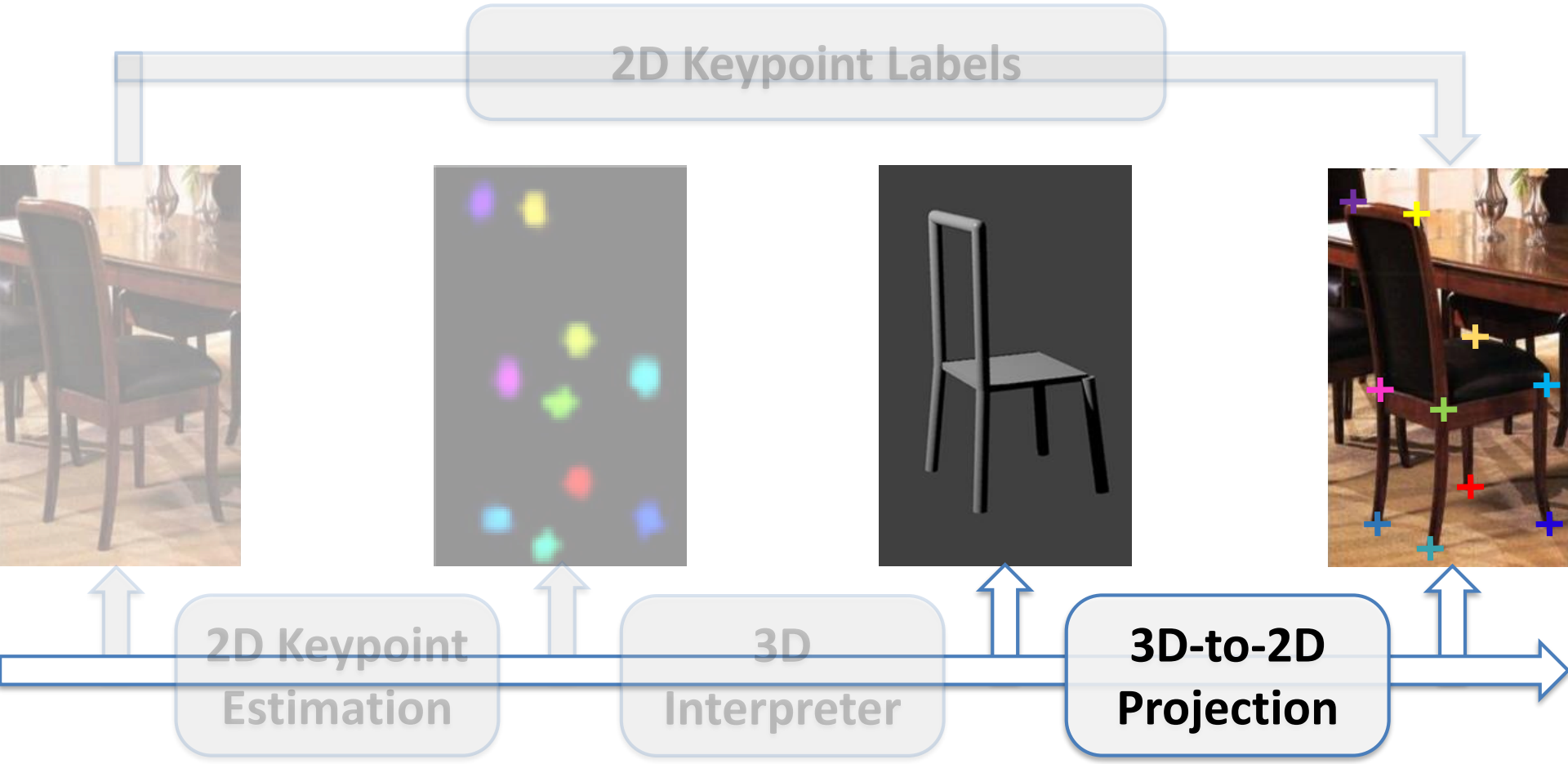
3D-INN: End-to-End Training?



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

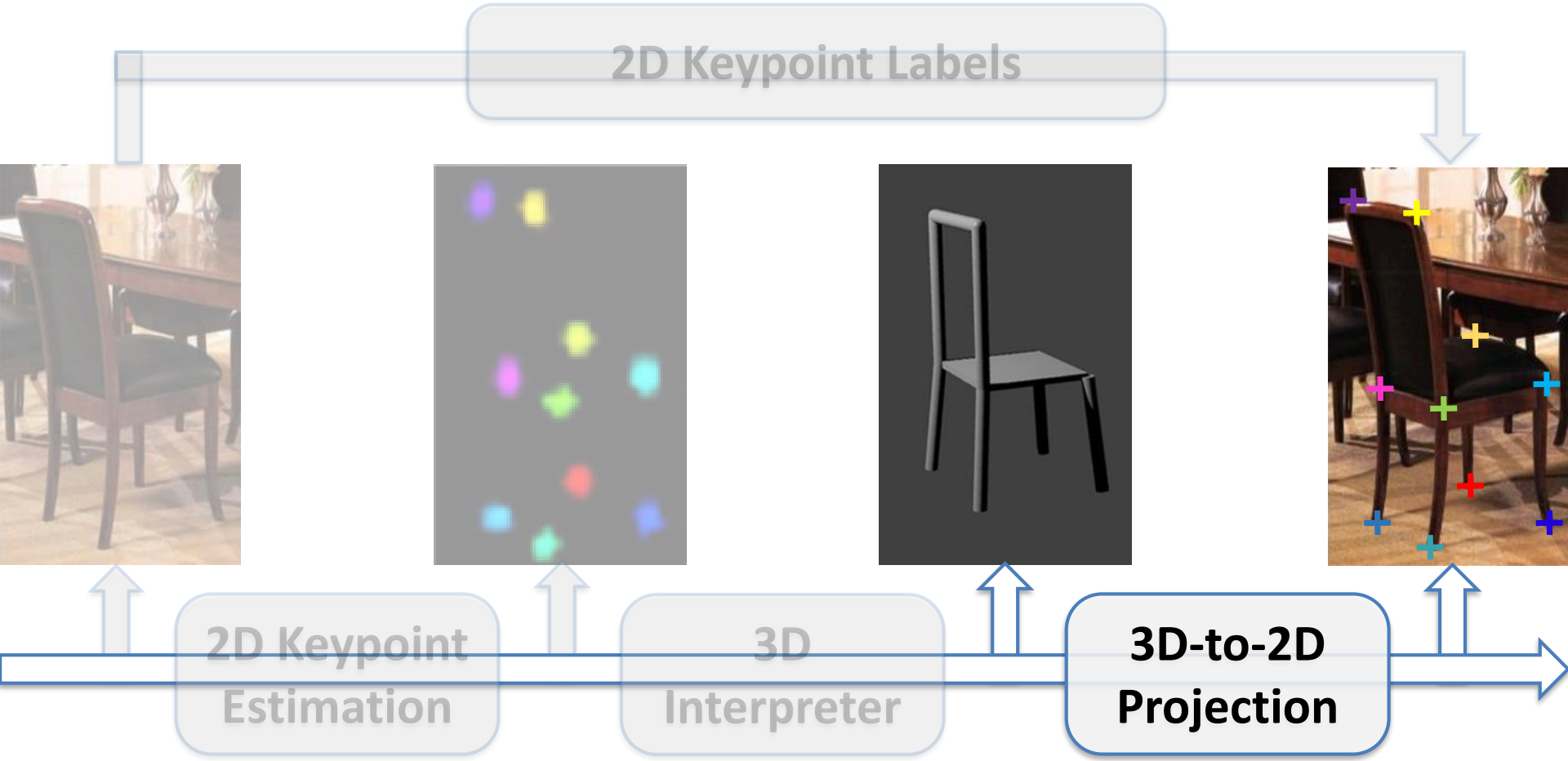


3D-INN: 3D-to-2D Projection Layer



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

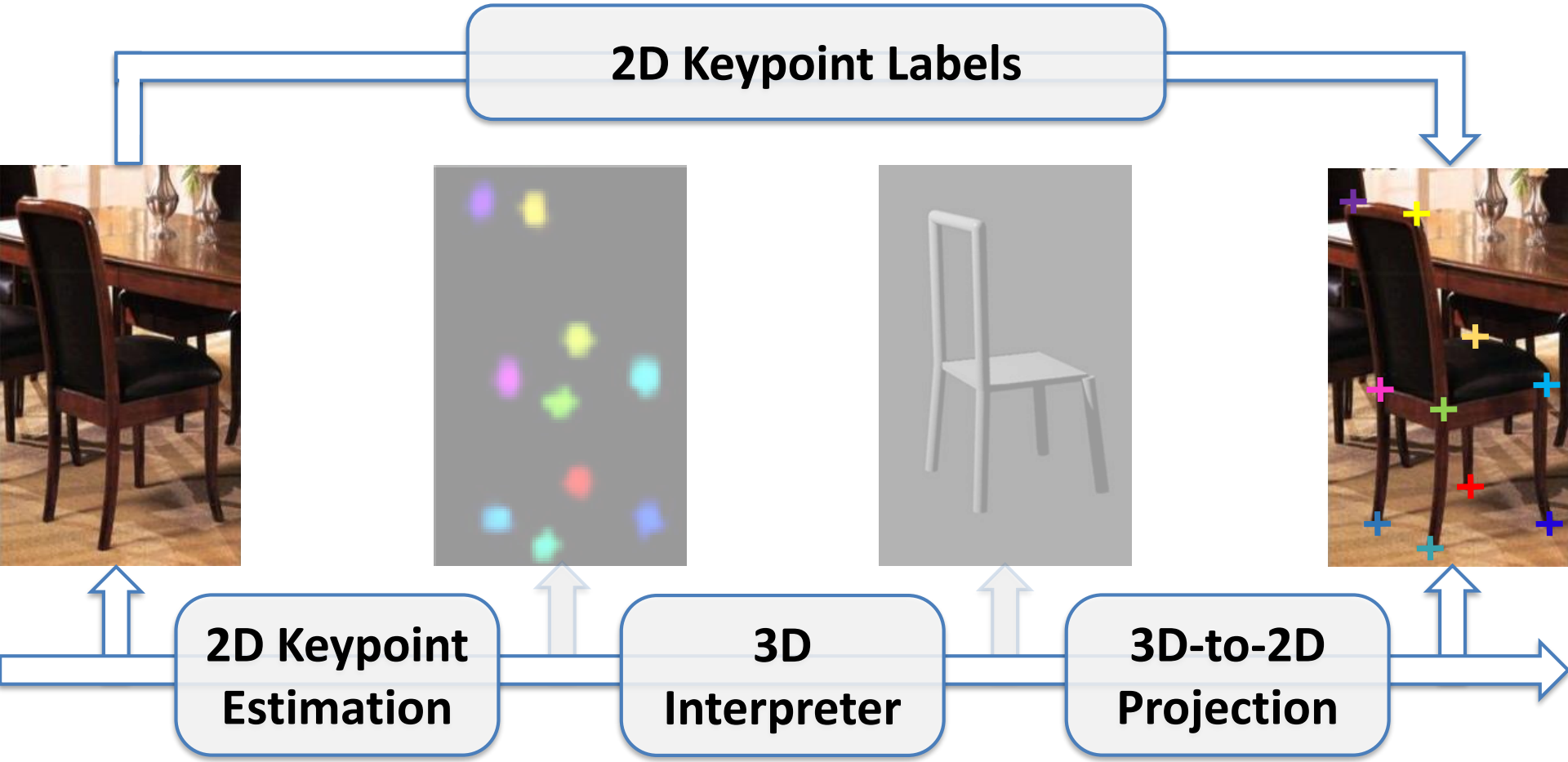
3D-INN: 3D-to-2D Projection Layer



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

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

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

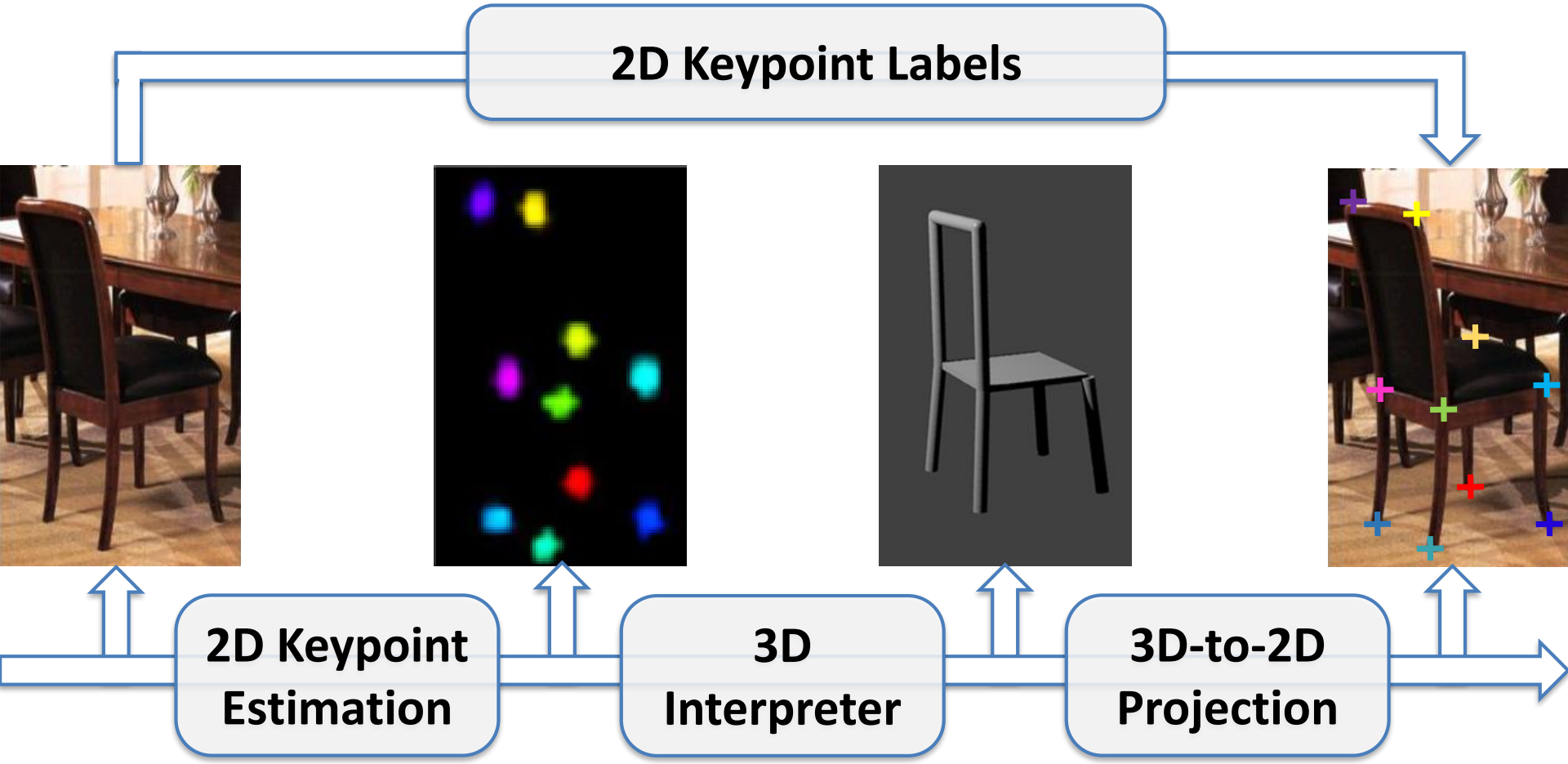


Using 2D-annotated real data

Input: an RGB image

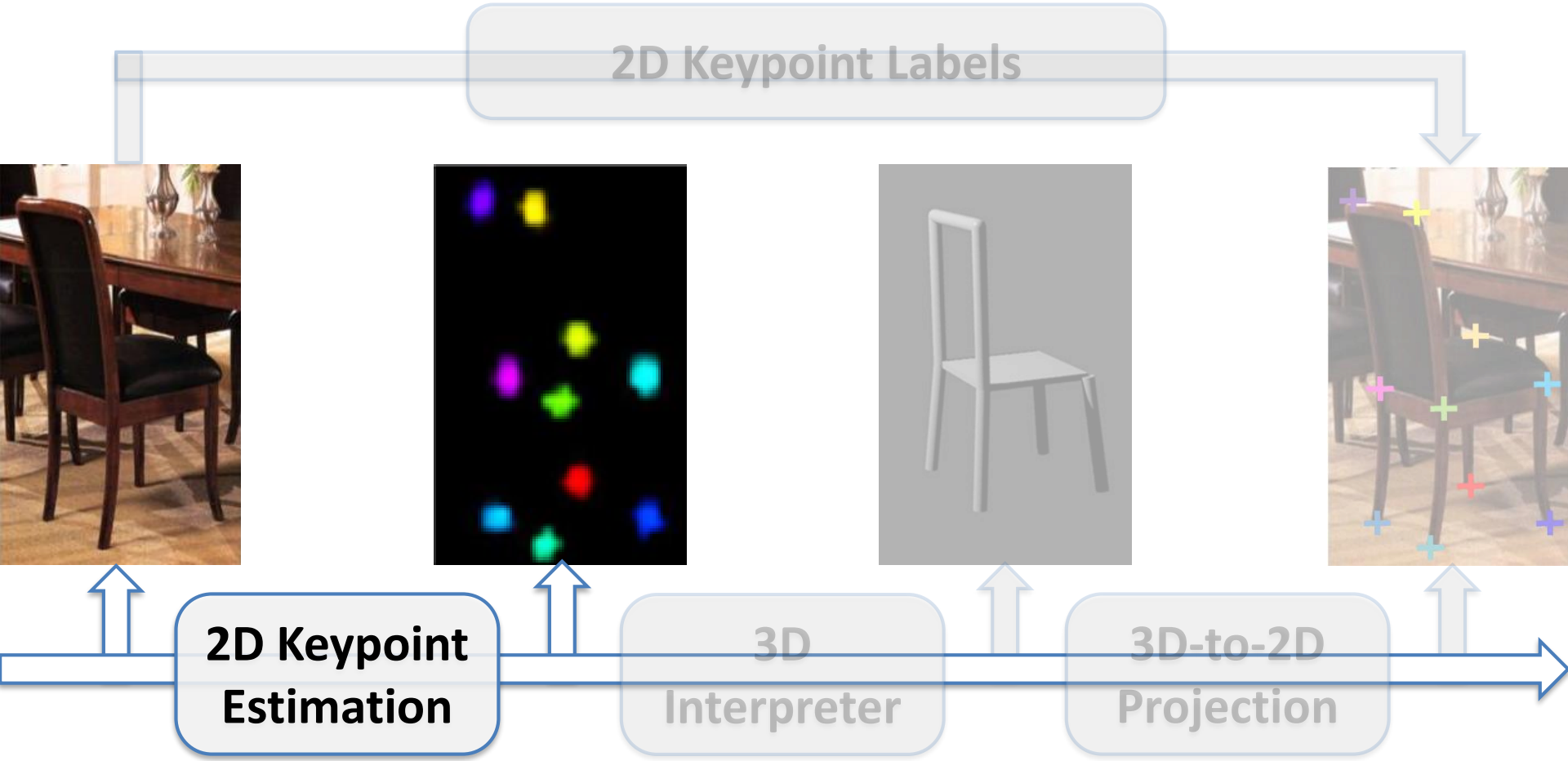
Output: Keypoint coordinates

3D-INN: Training Paradigm



Three-step training paradigm

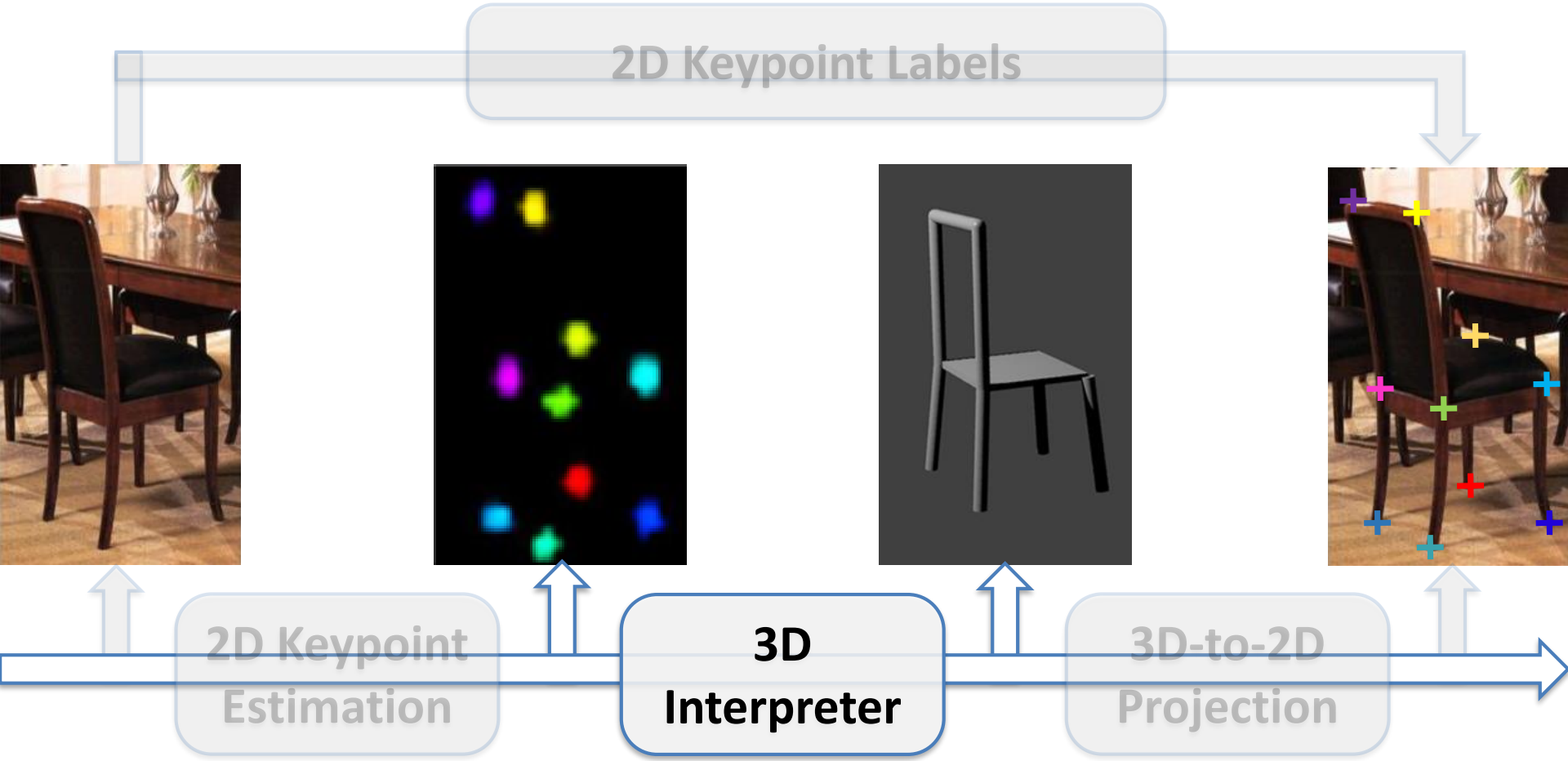
3D-INN: Training Paradigm



Three-step training paradigm

I: 2D Keypoint Estimation

3D-INN: Training Paradigm

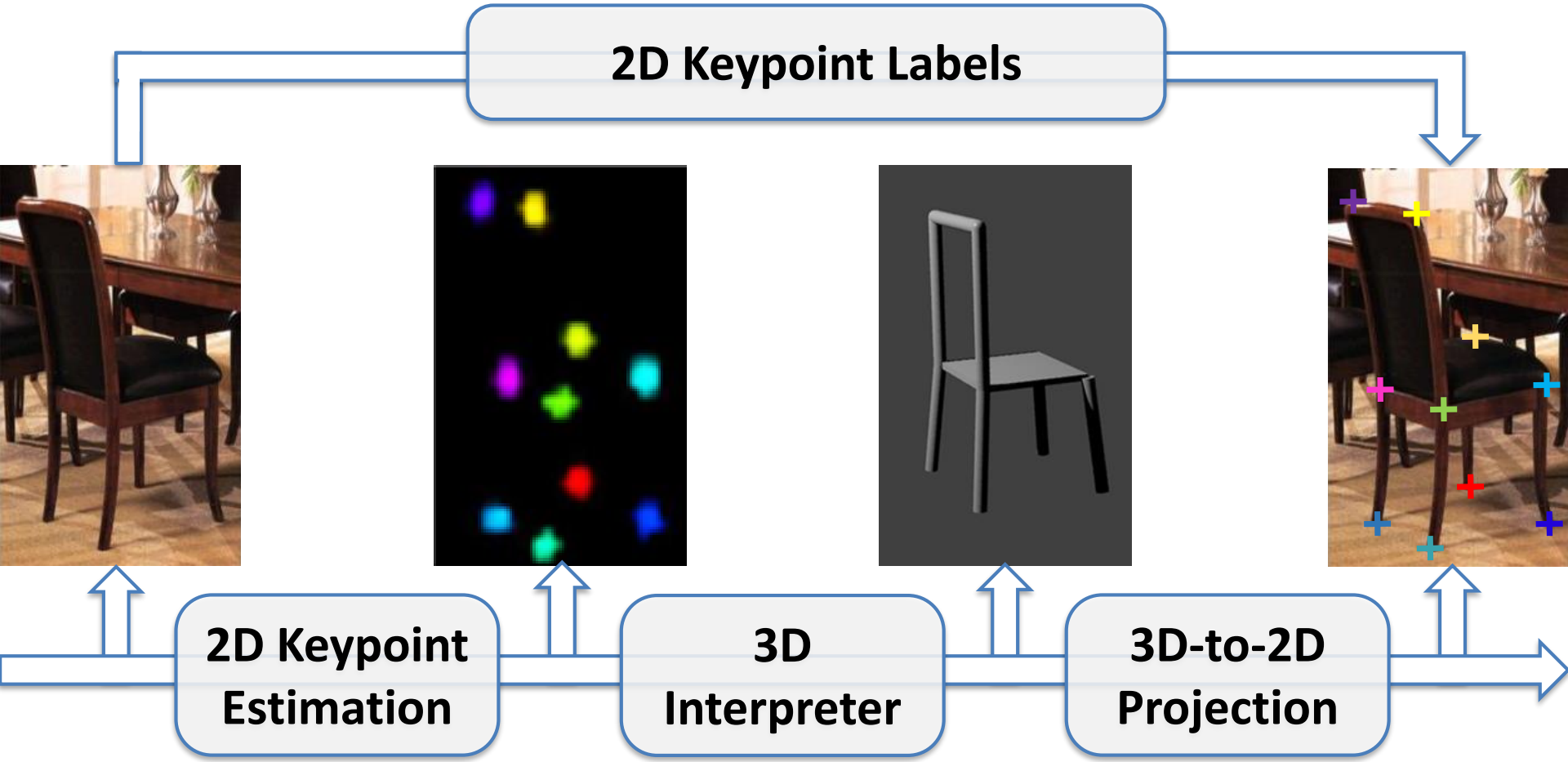


Three-step training paradigm

I: 2D Keypoint Estimation

II: 3D Interpreter

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

Refined Results

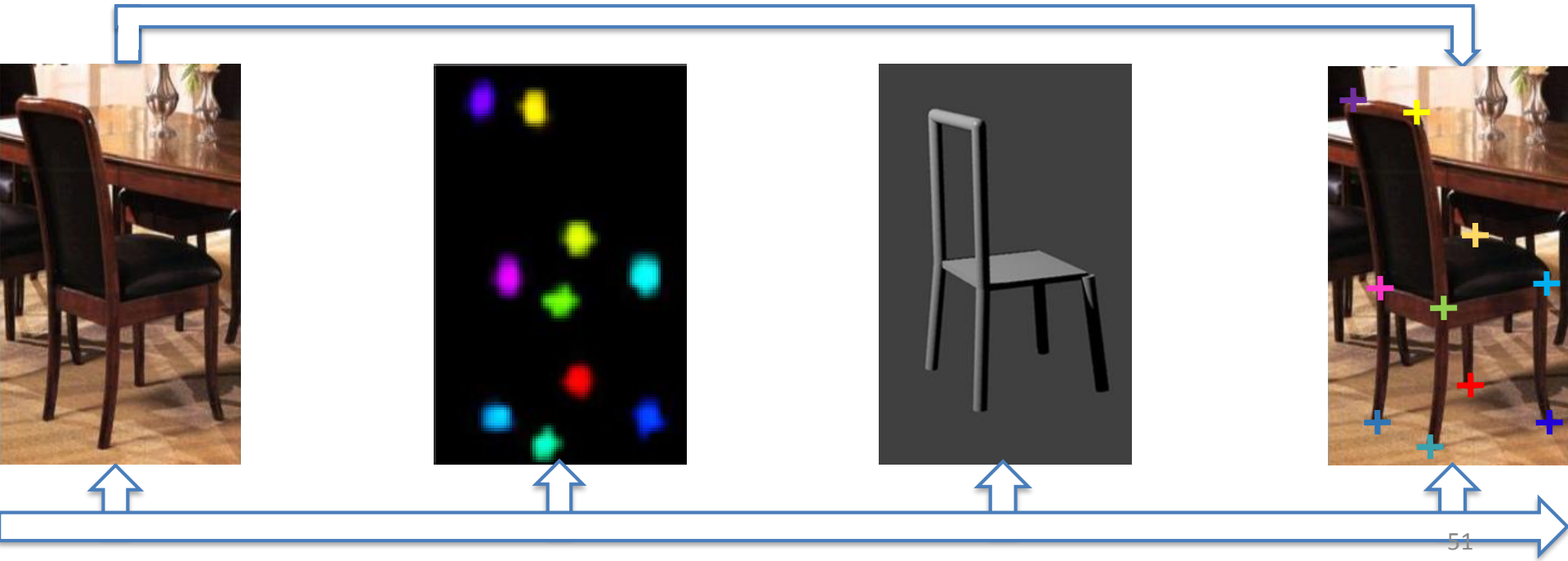
Image

Initial Estimation

After End-to-End Fine-tuning

Contribution II

- Real 2D labels + synthetic 3D models
- Keypoints as intermediate representations
- A 3D-to-2D projection layer for end-to-end training



3D Estimation: Qualitative Results

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

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*]

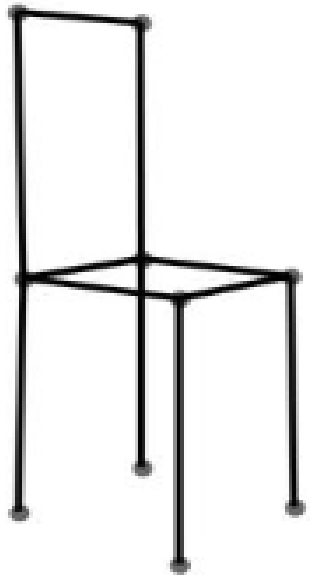
3D Estimation: Qualitative Results

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



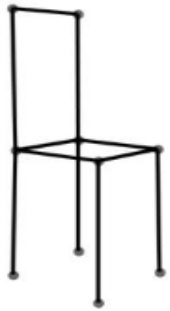
SUN Database [*Xiao et al, '11*]

3D Structure Estimation

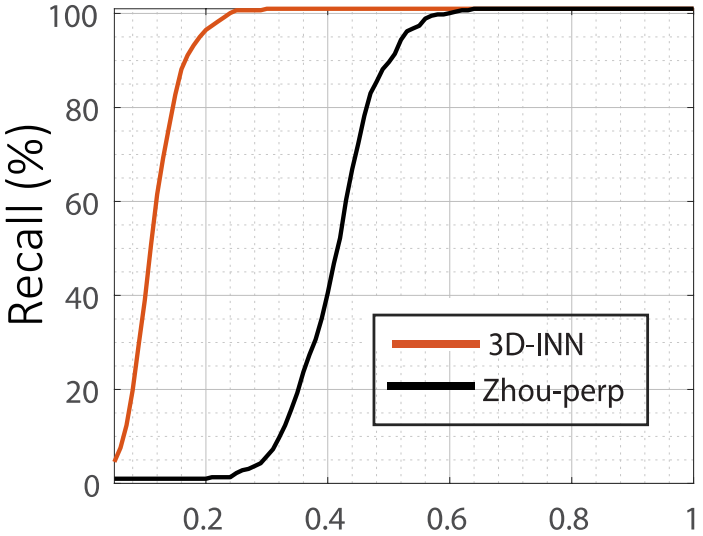
Images Results

3D Structure Estimation

Images Results



IKEA dataset [*Lim et al, '13*]

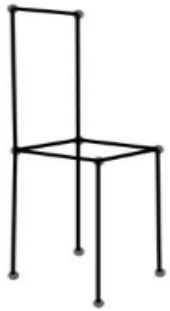


RMSE of estimated 3D keypoints

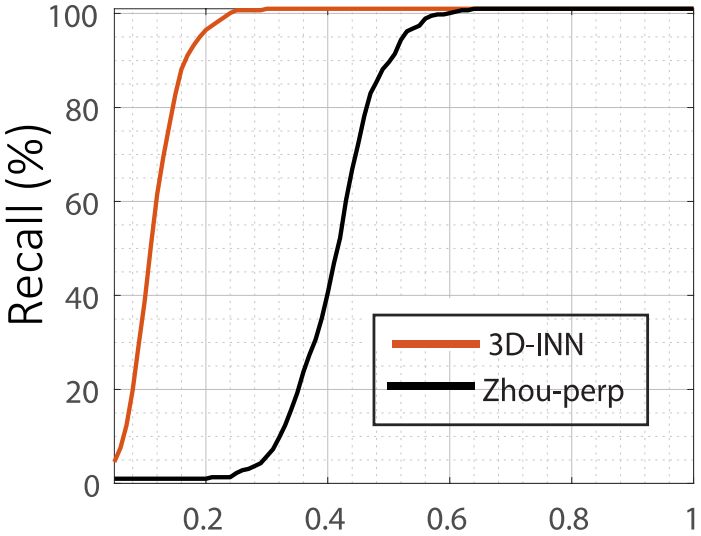
3D Structure Estimation

Images Results



IKEA dataset [*Lim et al, '13*]



RMSE of estimated 3D keypoints

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 (%)

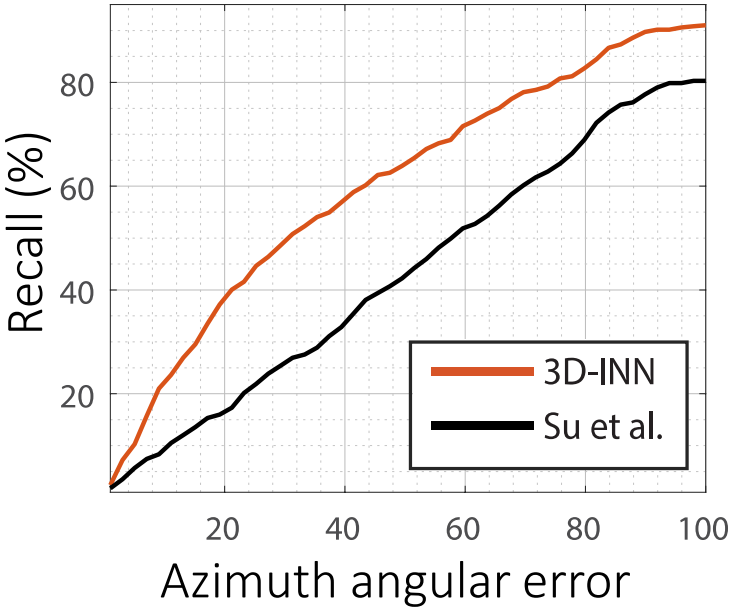
Viewpoint Estimation

Images Results

Viewpoint Estimation

Images Results

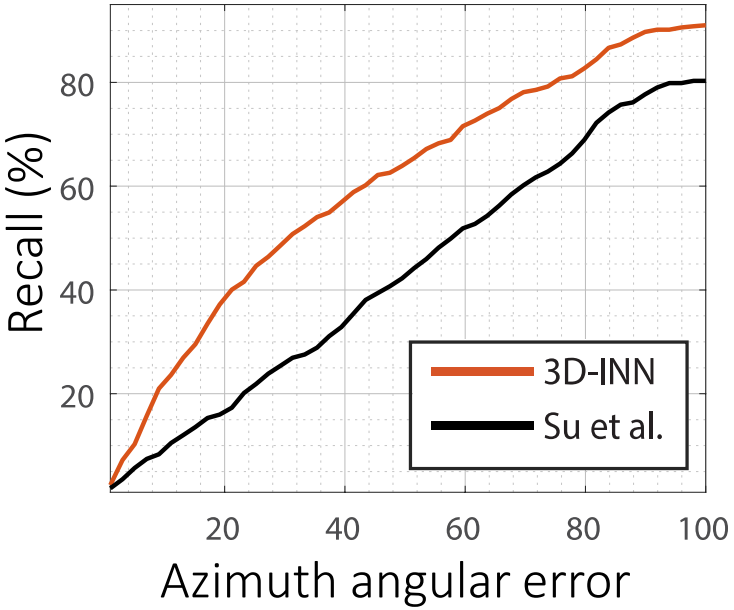
IKEA dataset [*Lim et al, '13*]



Viewpoint Estimation

Images Results

IKEA dataset [*Lim et al, '13*]



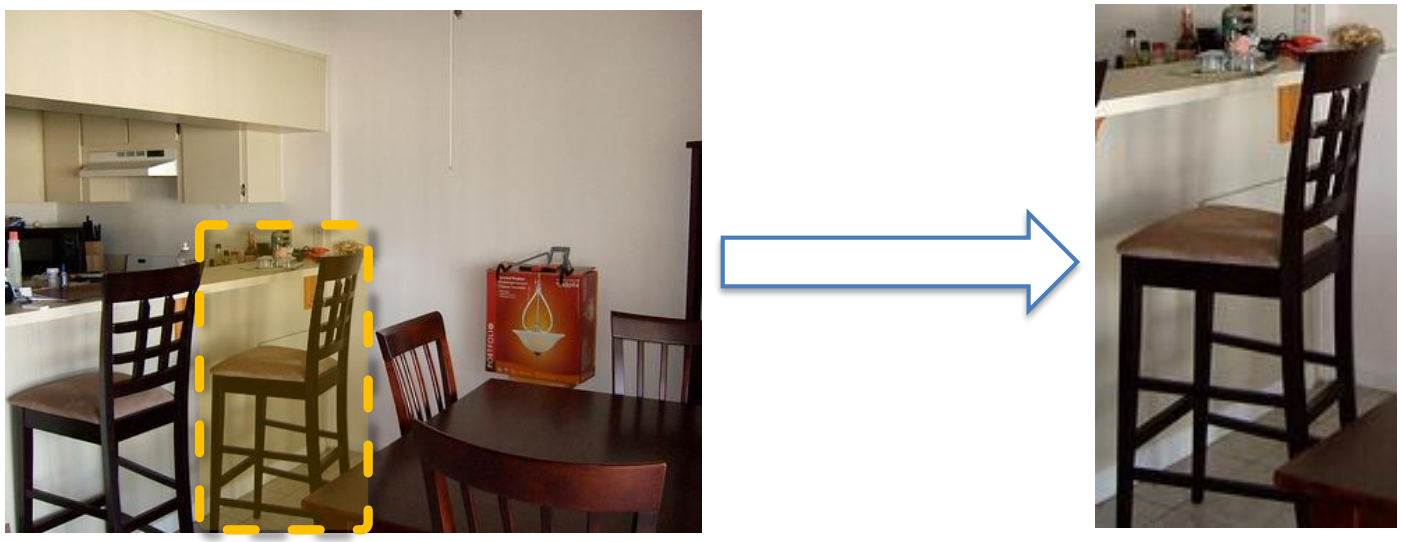
Method	Table	Sofa	Chair	Avg.
3D-INN	55.0	64.7	63.5	60.3
Su, '15	52.7	35.7	37.7	43.3

Average recall (%)

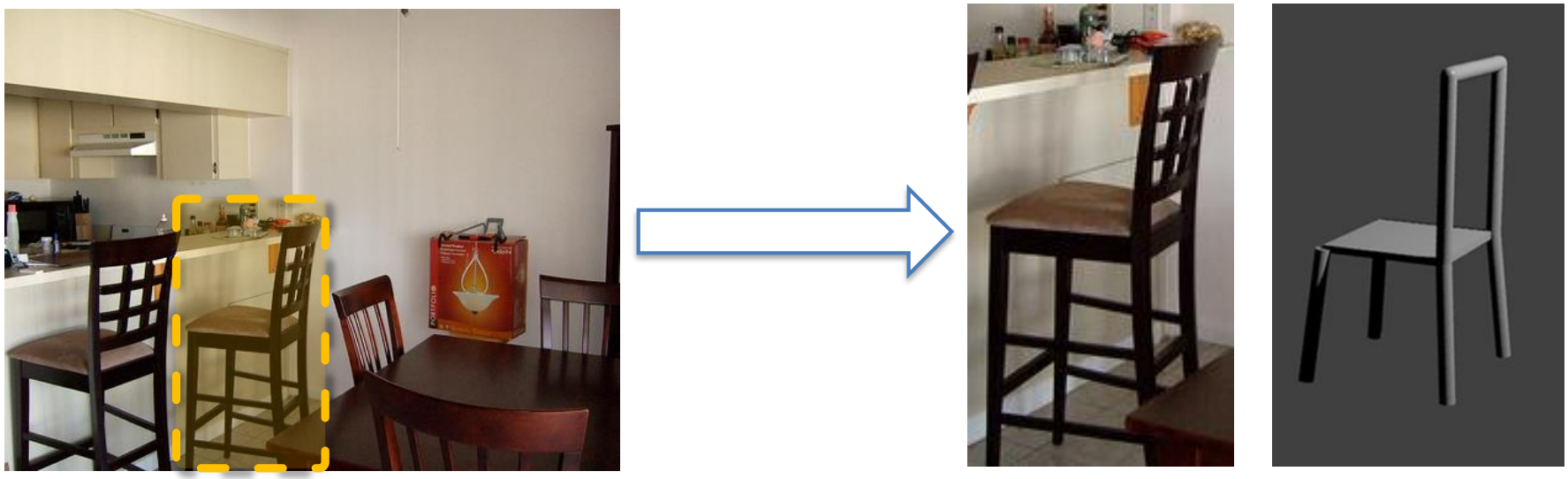
Localization and 3D Estimation

Localization and 3D Estimation

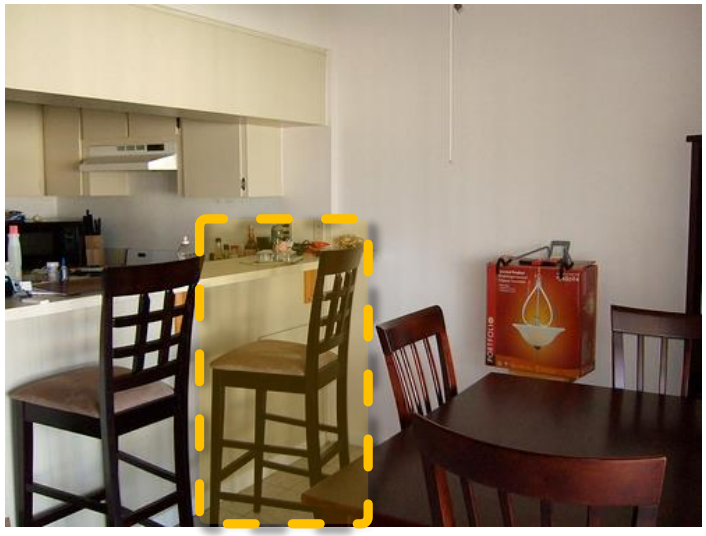
Localization and 3D Estimation



Localization and 3D Estimation



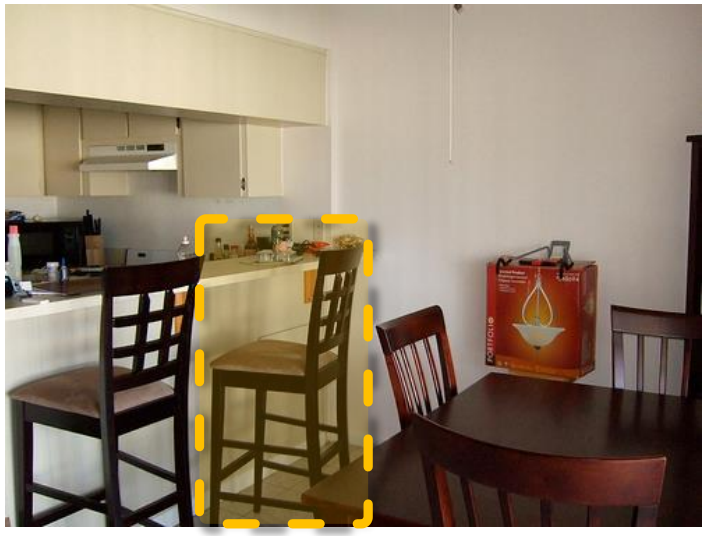
Localization and 3D Estimation



R-CNN

Girshick et al, ’14

Localization and 3D Estimation



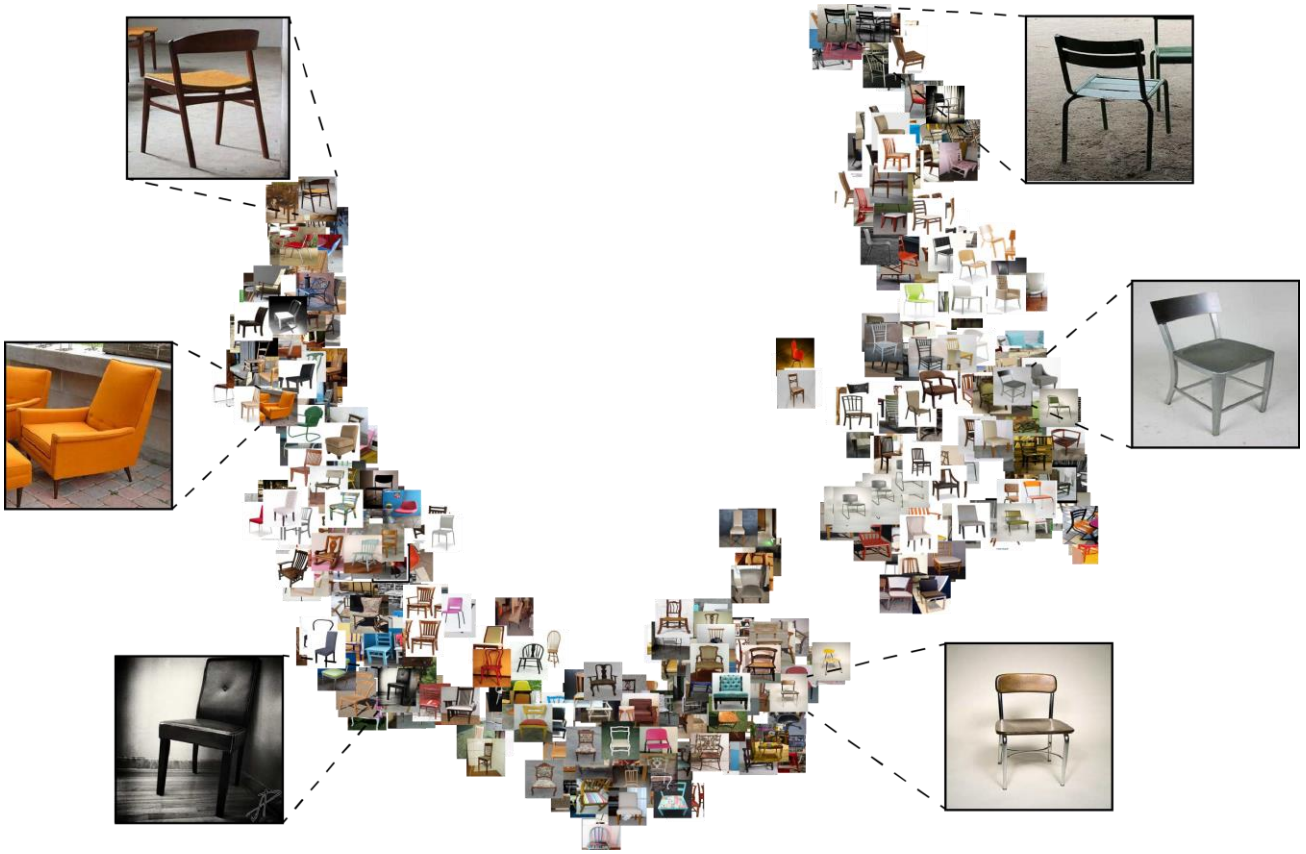
R-CNN

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 pose**

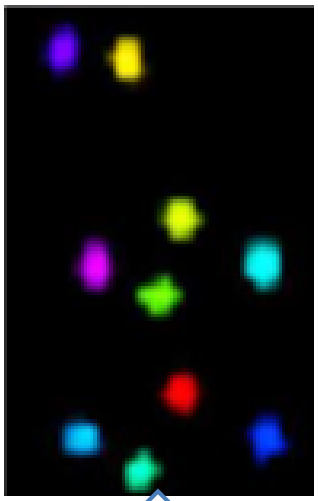
Images retrieved by pose

Contributions

- Single image 3D perception

Contributions

- Single image 3D perception
  - Real 2D labels + synthetic 3D models
  - Keypoints as intermediate representations



Contributions

- Single image 3D perception
  - Real 2D labels + synthetic 3D models
  - Keypoints as intermediate representations
  - A 3D-to-2D projection layer for end-to-end training