Learning Local RGB-to-CAD Correspondences for Object Pose Estimation

Georgios Georgakis[1], Srikrishna Karanam[2], Ziyan Wu[2], and Jana Kosecka[1]
George Mason University[1], Siemens Corporate Technology[2]

Object Pose Estimation

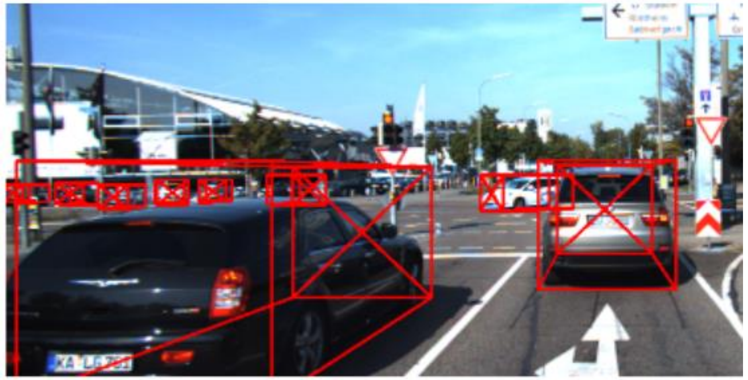
- Given an RGB image of an object, estimate rotation matrix $\boldsymbol{R} \in \boldsymbol{SO}(\mathbf{3})$ and translation vector $\boldsymbol{t} \in \mathbb{R}^{\mathbf{3}}$.
- $\boldsymbol{R}$ is parameterized by the three Euler angles: <u>azimuth</u>, <u>elevation</u>, <u>in-plane rotation</u>.

Object Pose Estimation

- Essentially estimate an object's orientation and position relative to a coordinate system.
- **Challenges:**
  - Estimating geometry from 2D image plane.
  - Large space of possible rotation matrices.
  - Large variety of object shapes.

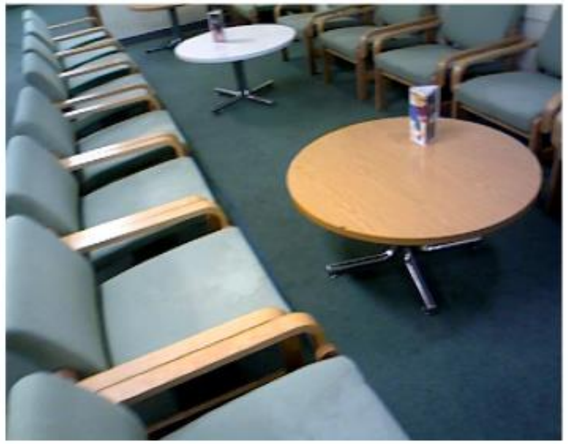
Applications

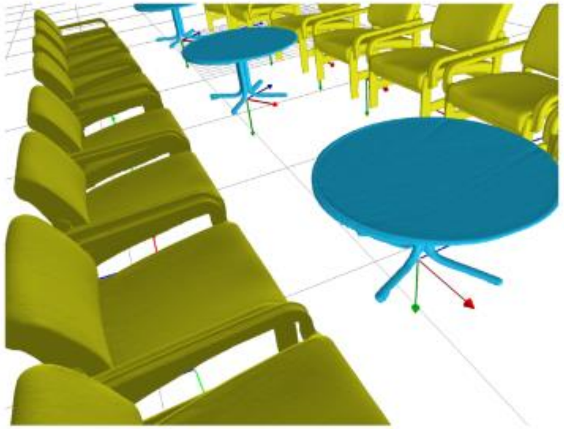
Autonomous driving



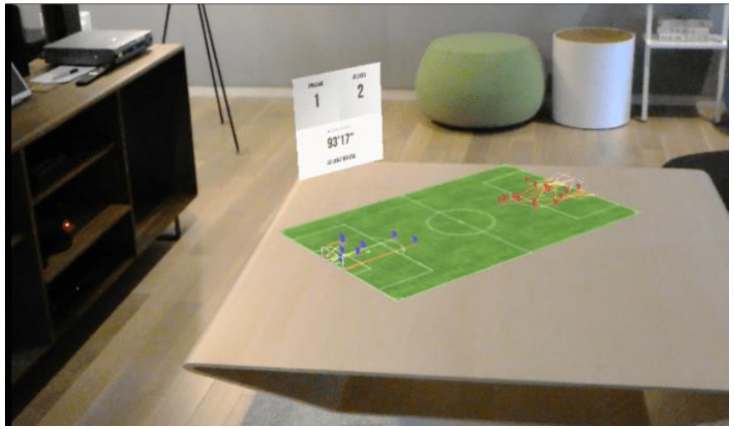
Robotic manipulation

Indoor navigation





Augmented Reality

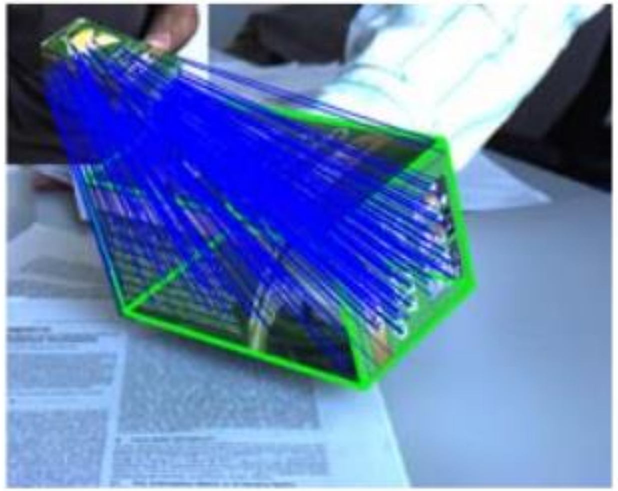


Traditional Pose Estimation

3D Model

Keypoint detection

Reference frame

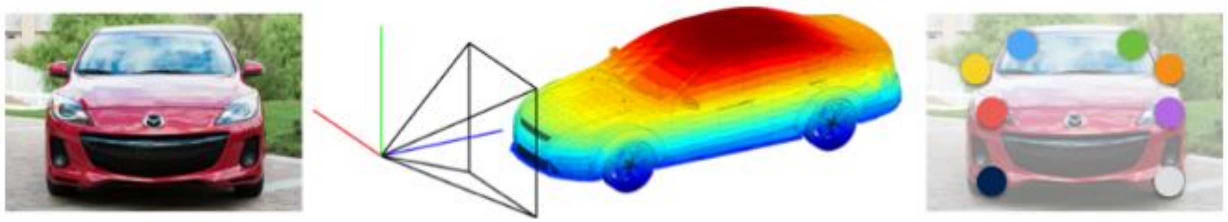


Pose

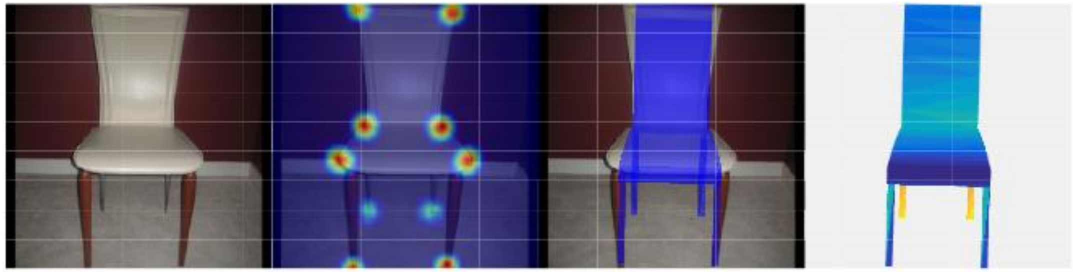
- <u>Keypoints:</u> Interest points which yield discriminative representations and can be matched reliably across images.

- Pose estimated through Perspective-n-Point (PnP) using set of correspondences.

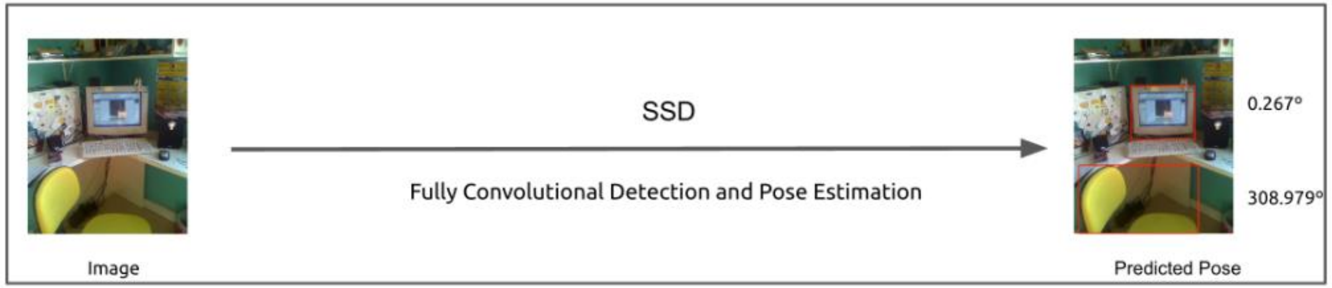
Related work



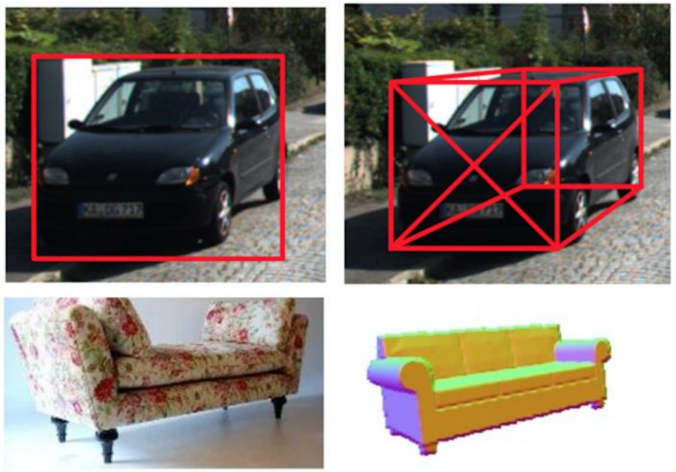
Viewpoint conditioned keypoints
[Tulsiani et al. 2015]



Fixed semantic keypoints
[Pavlakos et al. 2017]

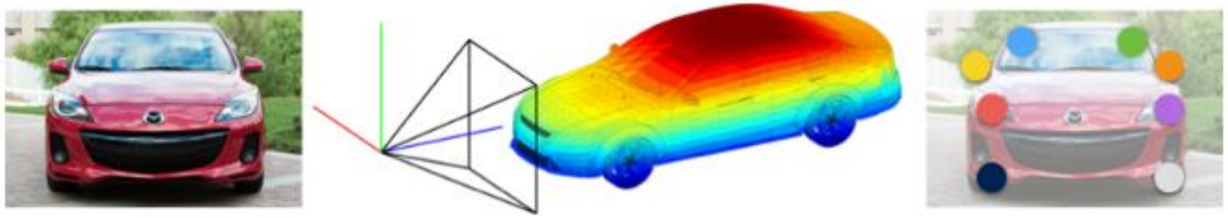


Single-step pose regression
[Poirson et al. 2016]

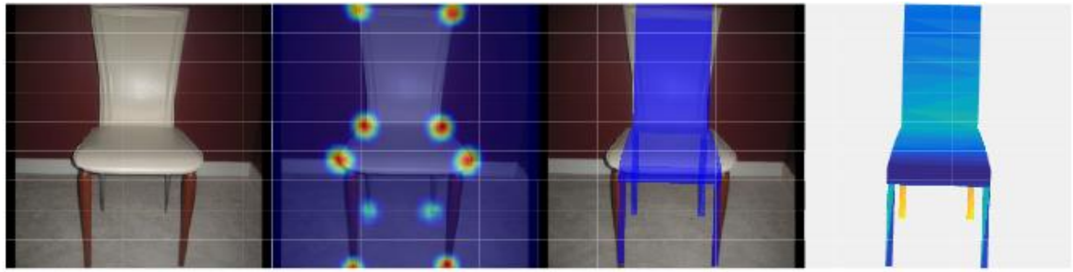


Classification-Regression Hybrid
[Deep3DBox, Mousavian et al. 2017]

Related work



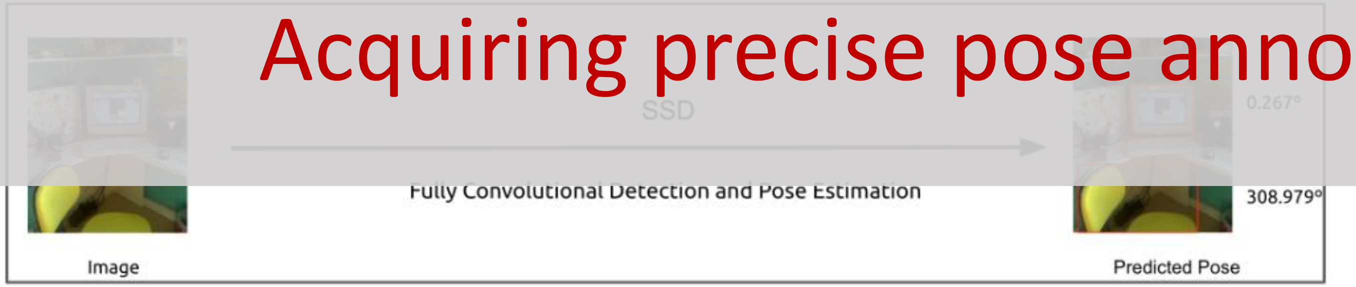
Viewpoint conditioned keypoints
[Tulsiani et al. 2015]



Fixed semantic keypoints
[Pavlakos et al. 2017]

Require 3D pose or keypoint annotations for RGB images.

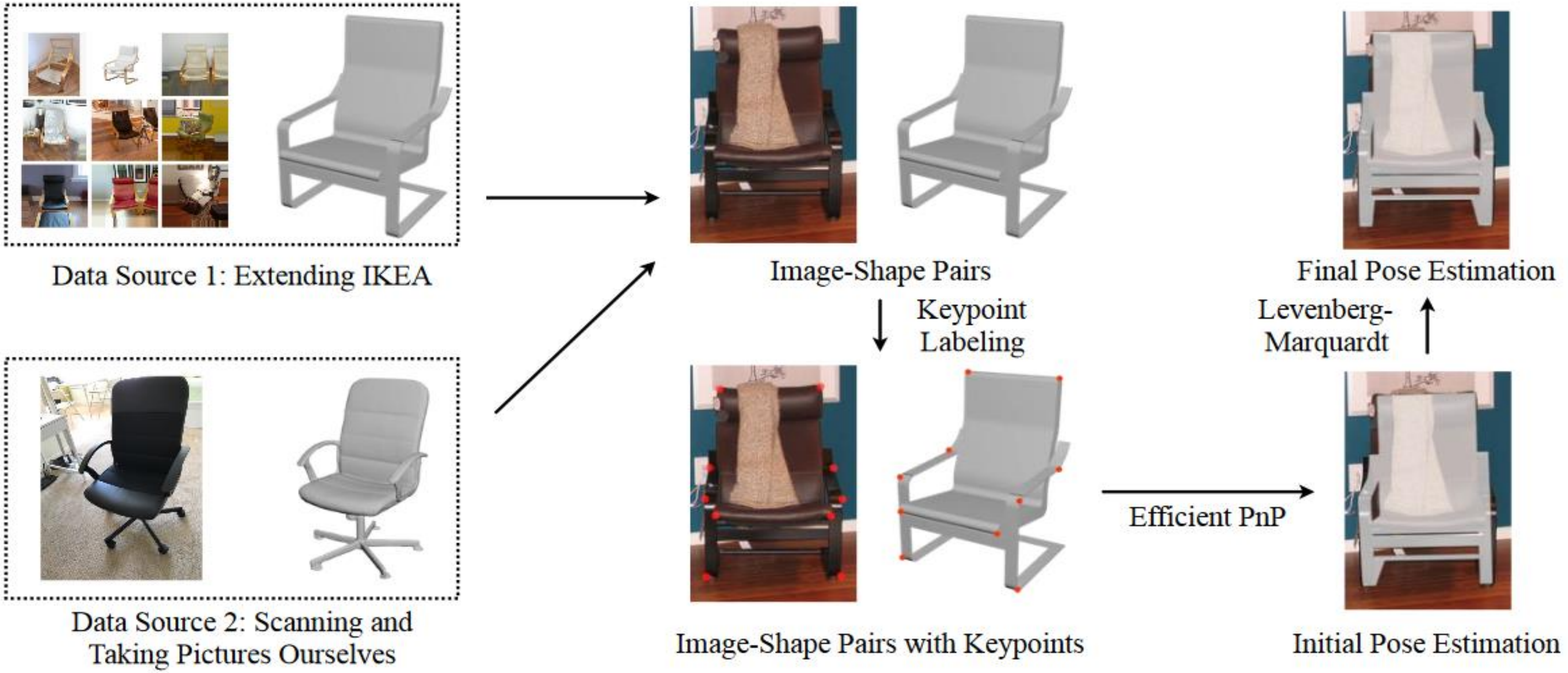
Acquiring precise pose annotations is tedious.



Single-step pose regression
[Poirson et al. 2016]

Classification-Regression Hybrid
[Deep3DBox, Mousavian et al. 2017]

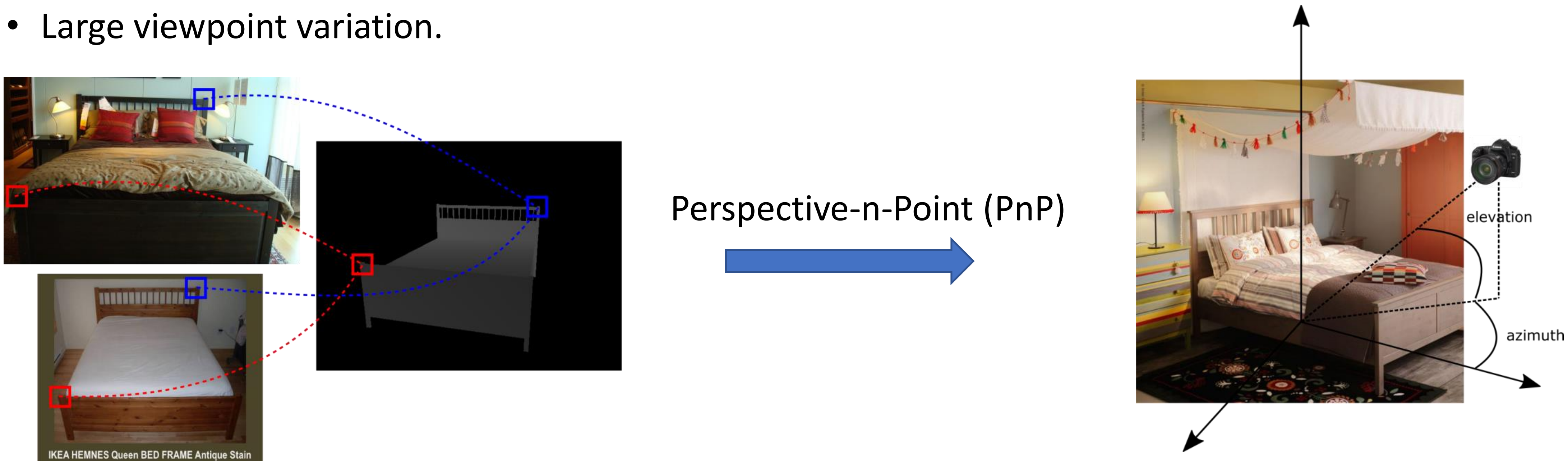
Getting Pose Annotations



Sun et al, Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling, 2018

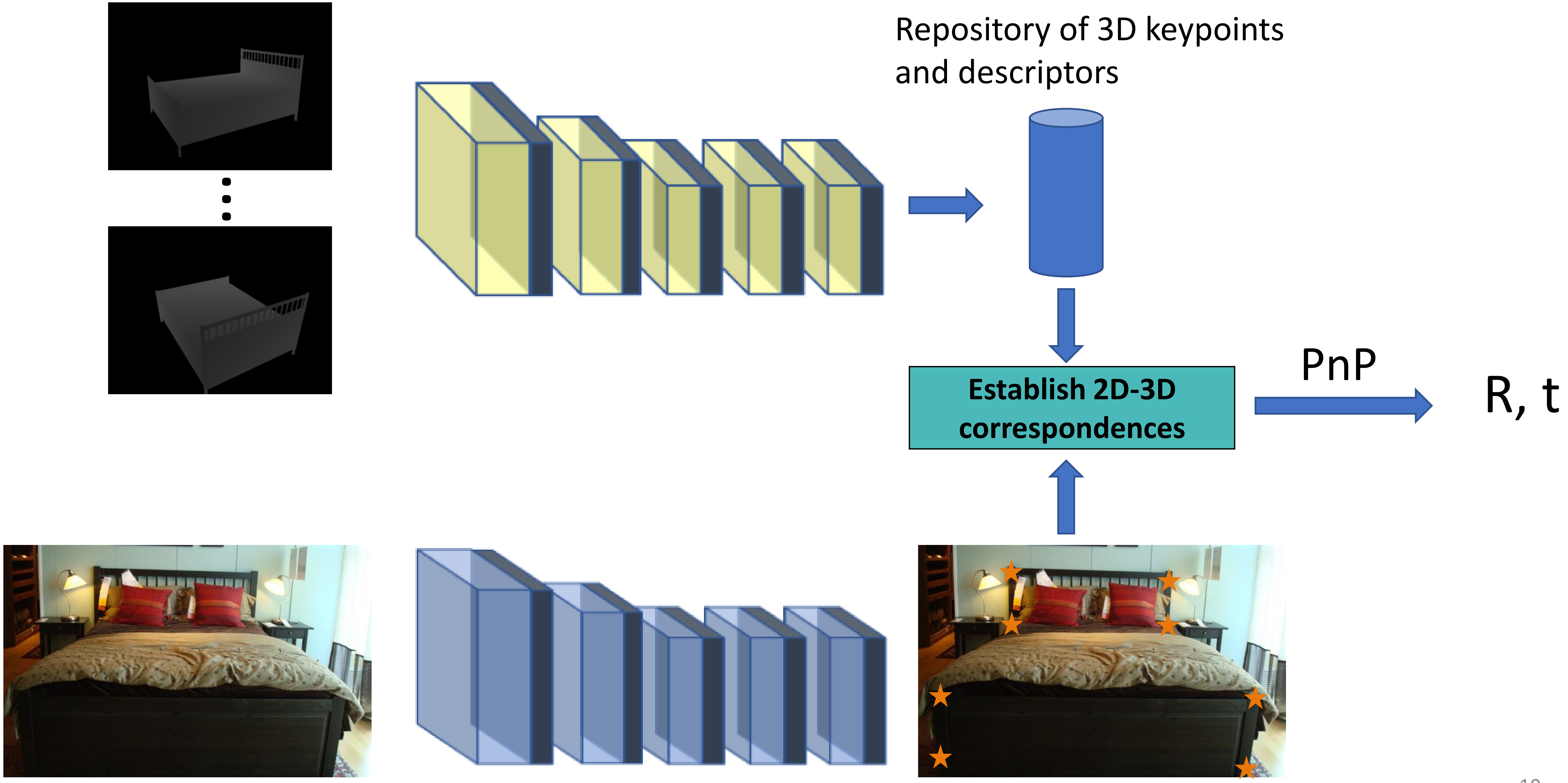
Correspondence Learning for Object Pose Estimation

- We focus on finding the association of parts of objects depicted in RGB images with their counterparts in 3D depth images.
  - Large appearance gap between RGB and depth data.
  - Large viewpoint variation.

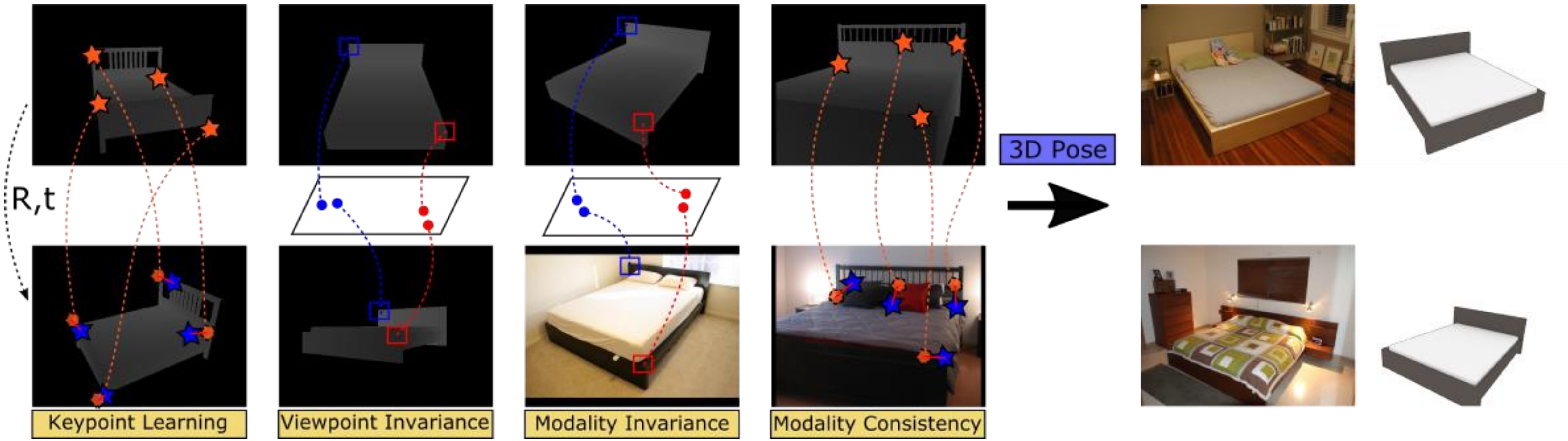


- Learn how to select which parts of objects are most informative (Keypoint detection).
  - We do not rely on keypoint annotations.

Testing


Repository of 3D keypoints
and descriptors
Establish 2D-3D
correspondences
PnP
R, t

Overview

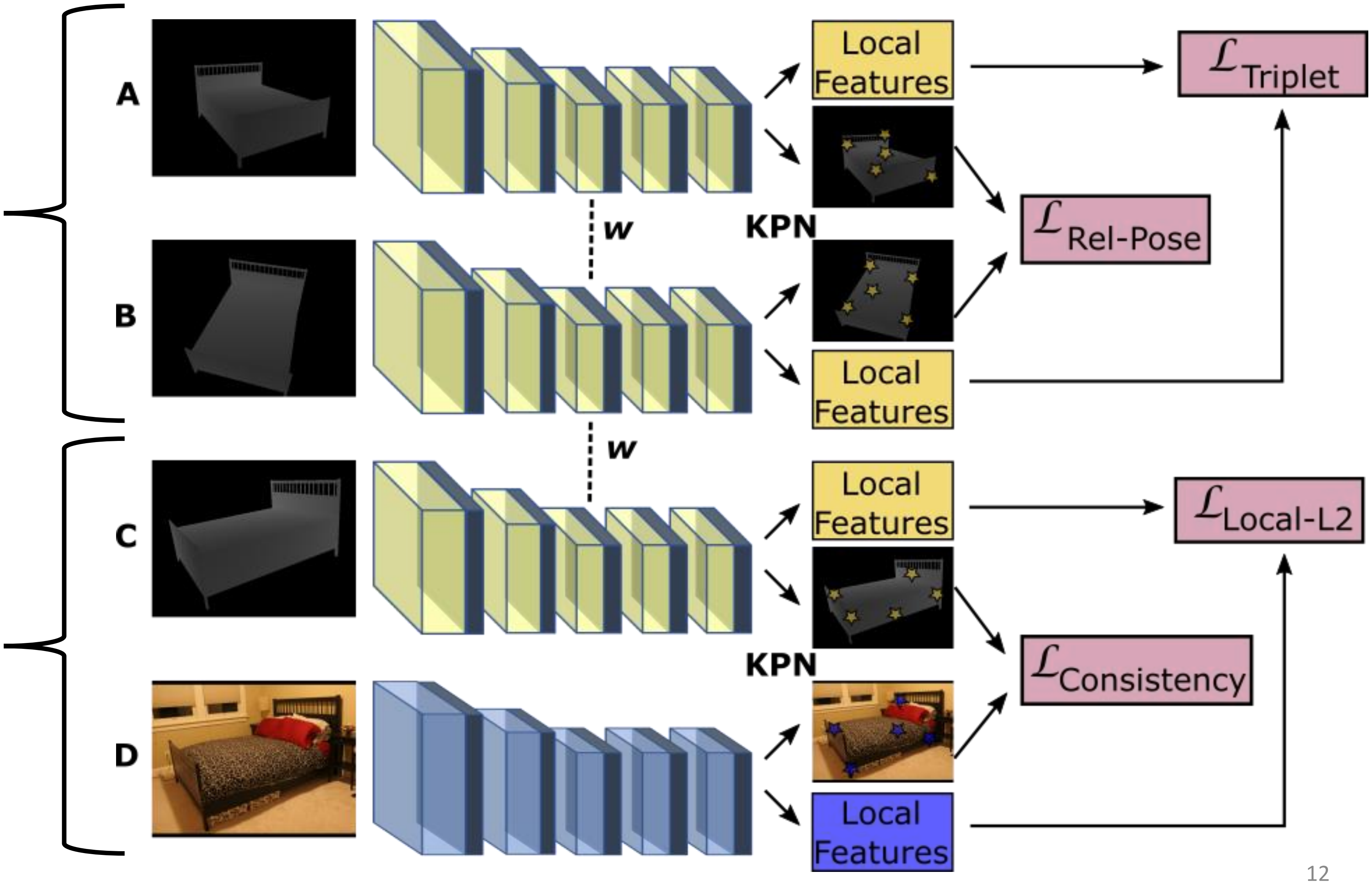


- We do not use:
  - Explicit 3D Pose annotations on RGB images.
  - Textured 3D models.

Architecture outline

Keypoint and descriptor learning.

Cross-modality representation learning.

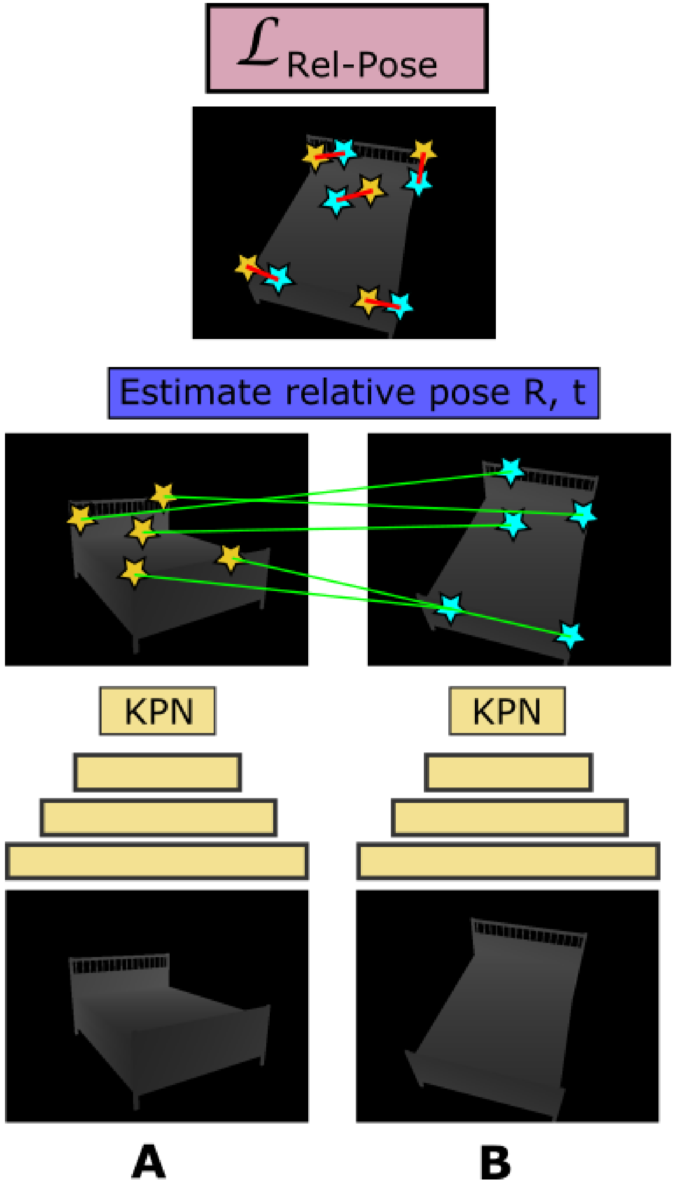


Keypoint learning by relative pose estimation

- **Relative pose loss:** For a weighted set of corresponding points find the rigid transformation for which the re-projection error is minimum.

$$(R, t) = arg \min_{R \in SO(3), t \in \mathbb{R}^3} \sum_{i=1}^{n} w_i \|(Rp_i + t) - q_i\|^2$$

- Given the correspondences, there exists a differentiable SVD-based closed form solution for estimating R, t.

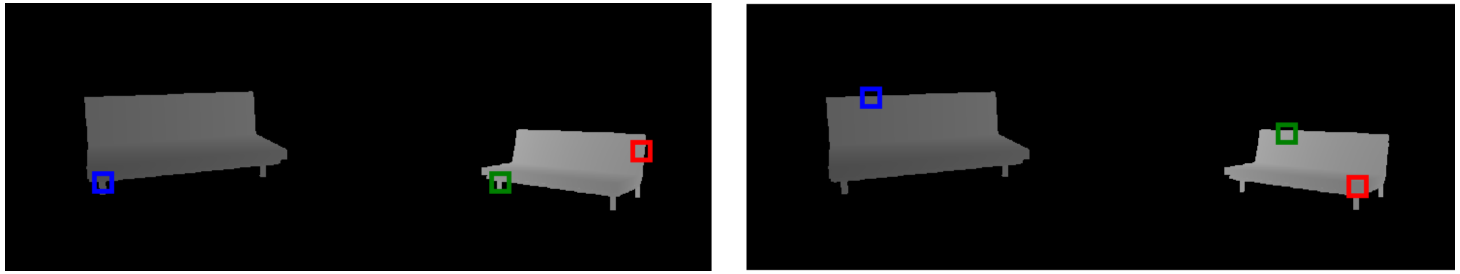


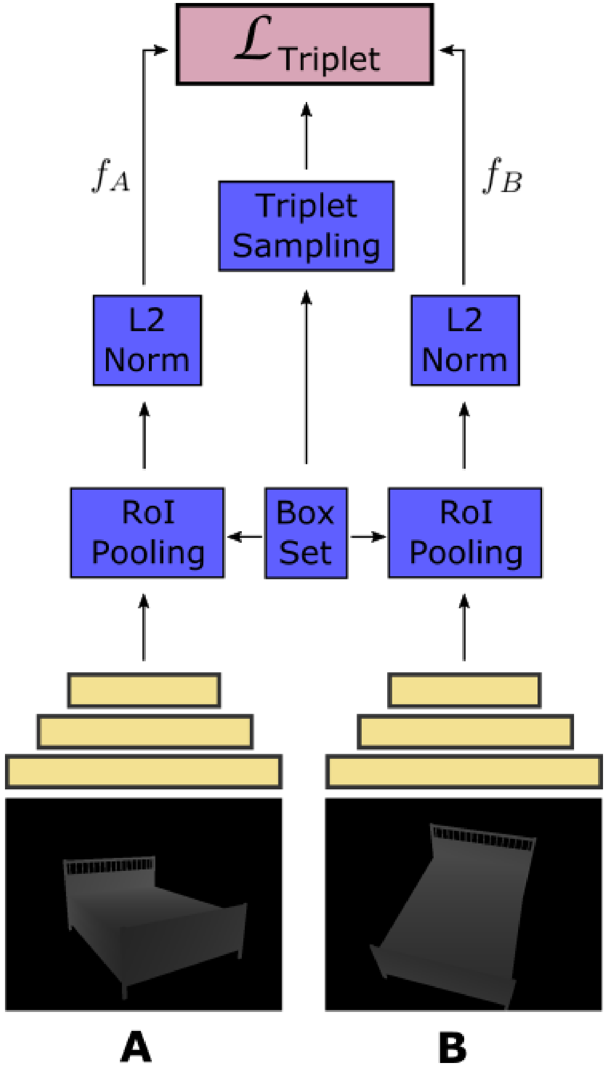
Learning keypoint descriptors – Triplet loss

$$L_{Triplet} = \sum_{i}^{N} \max(0, ||f_i^a - f_i^p||^2 - ||f_i^a - f_i^n||^2 + m)$$

- $f_i^a \; f_i^p \; f_i^n$: Local features of anchor, positive, and negative from triplet i.

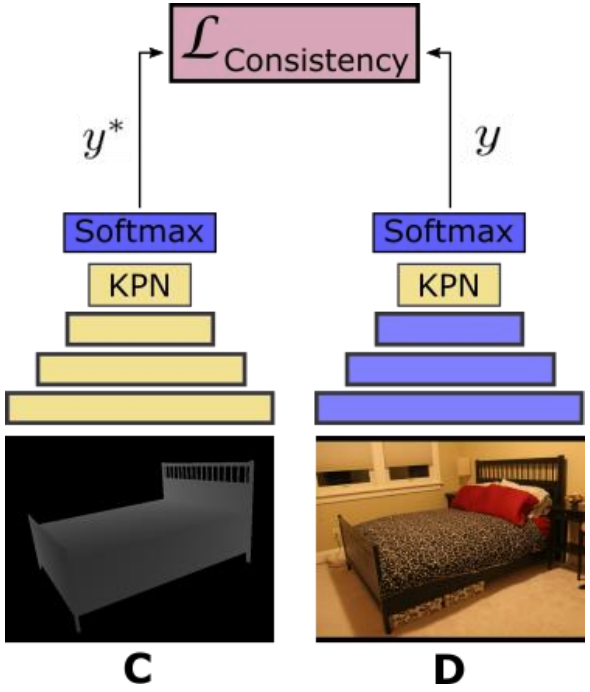
- Triplet examples:

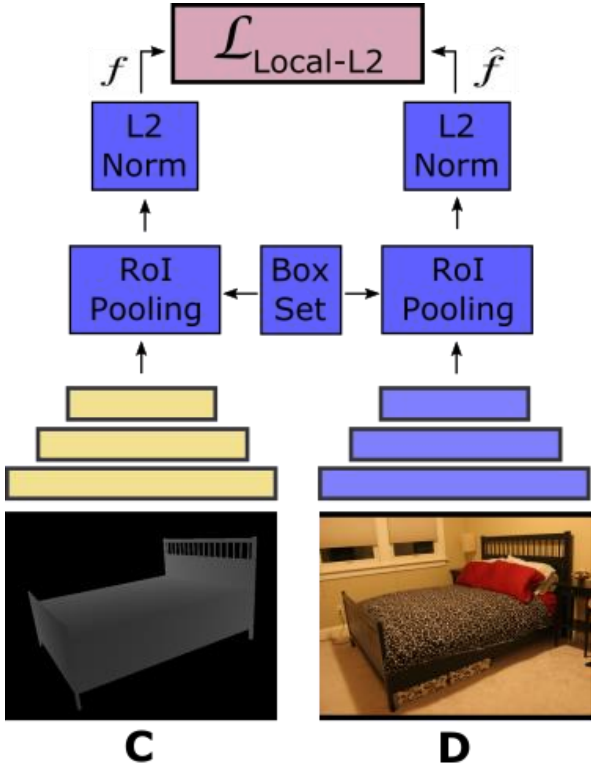




Cross-modality representation learning

- Transfer learned features and keypoints from branches A,B,C to branch D.
- Inspired by knowledge distillation.



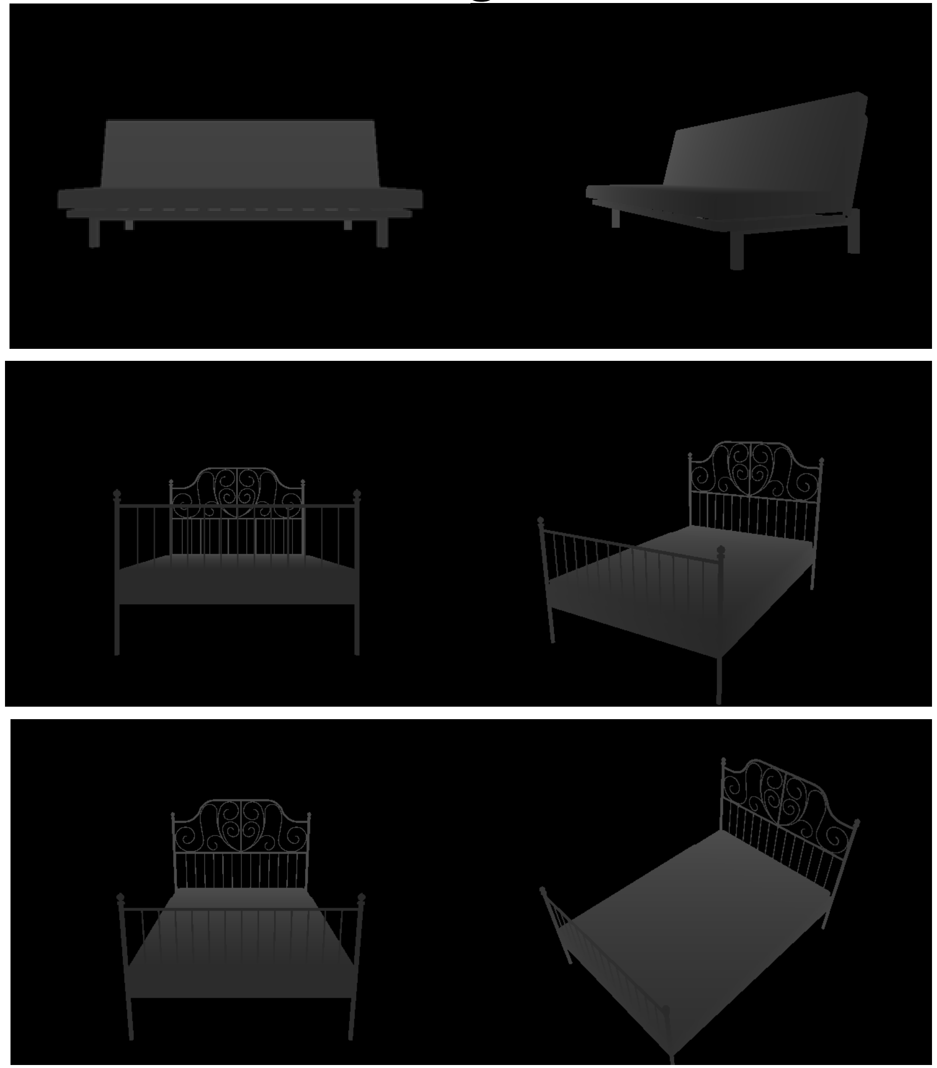


$$L_{consistency} = -\frac{1}{n}\sum_{i=1}^{n} y_i^C \log y_i^D$$

$$L_{local-l2} = \frac{1}{k}\sum_{i=1}^{k} \|\hat{f}_i - f_i\|$$

Examples of learned representations

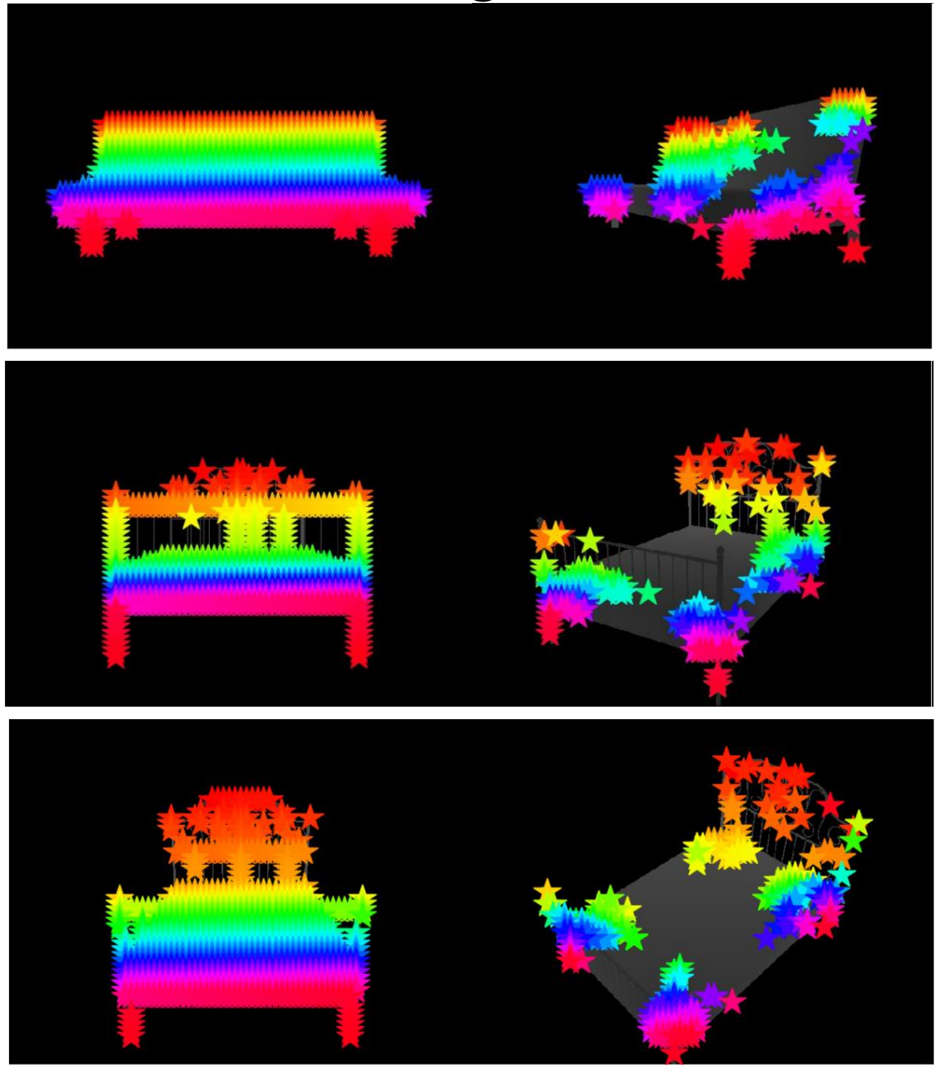
View alignment



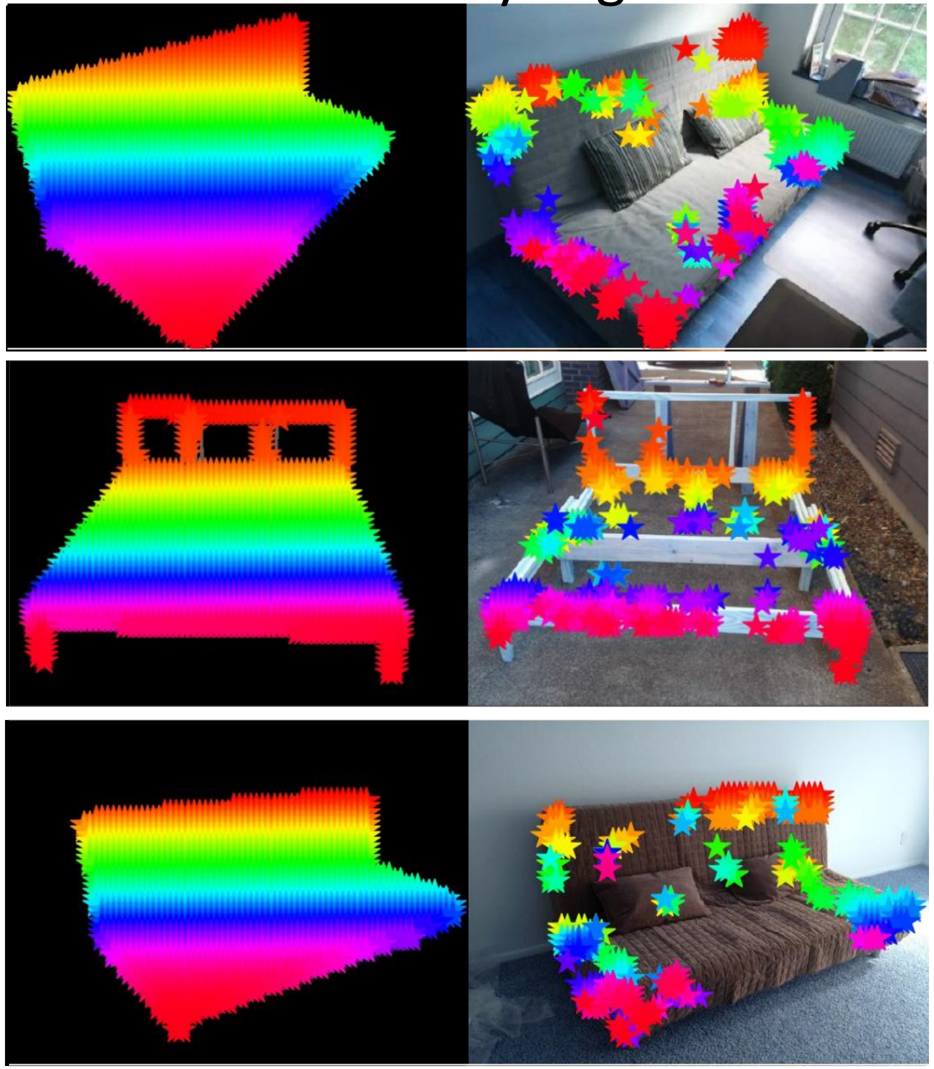
Cross-modality alignment

Examples of learned representations

View alignment

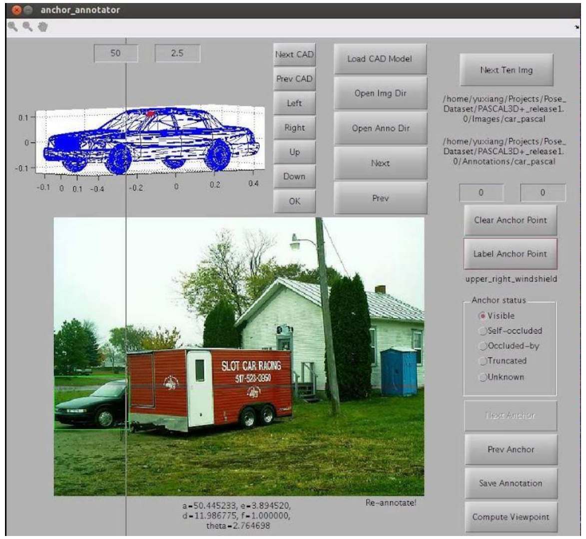


Cross-modality alignment



Experiments

- Pascal3D+: Manual alignment of generic 3D models on images



- Pix3D: Precise 3D pose annotations for object instances

Results

- Evaluation metric: Geodesic distance $\Delta(R_1, R_2) = \frac{||\log(R_1^T R_2)||_F}{\sqrt{2}}$

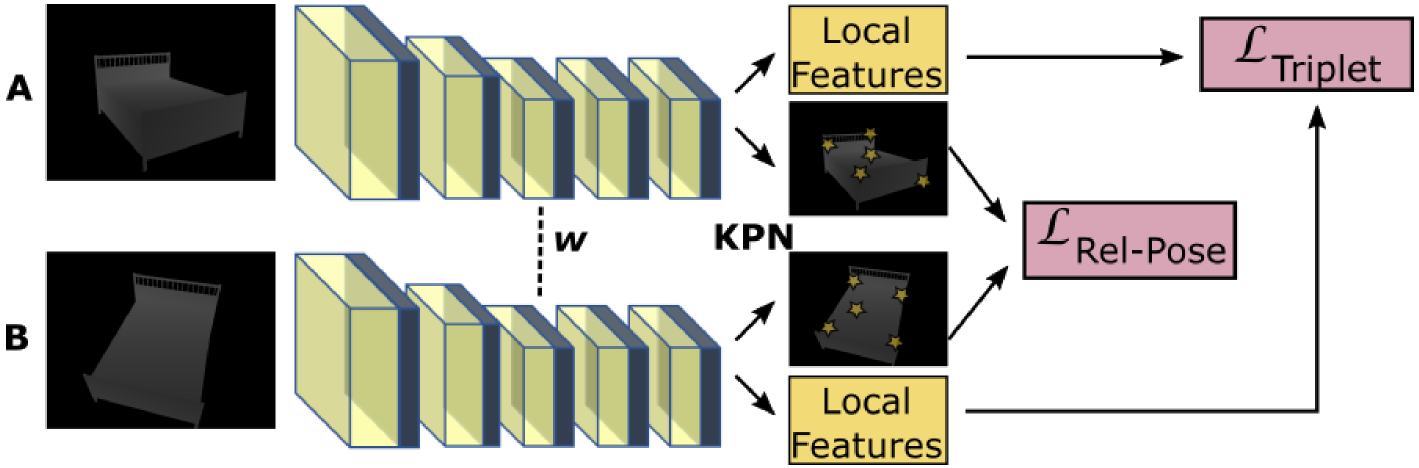
Comparison with supervised approaches

- Training on Pix3D – Testing on Pascal3D+

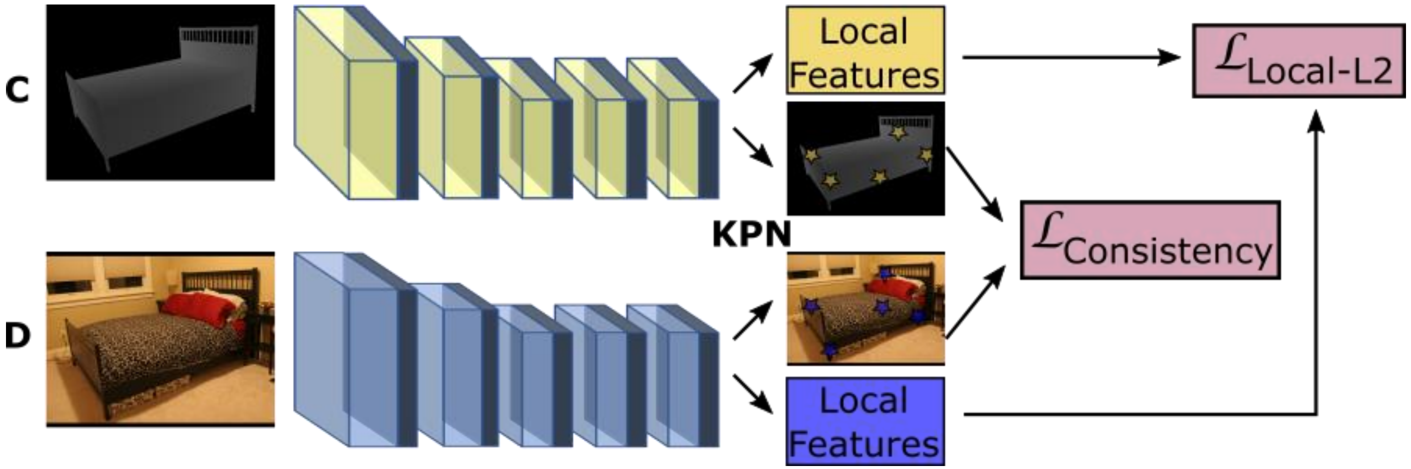
Category	Chair		Sofa	
Metric	Acc$\frac{\pi}{6}$	MedErr	Acc$\frac{\pi}{6}$	MedErr
Render for CNN [33]	4.3	2.1	11.6	1.2
Vps & Kps [39]	10.3	1.7	23.3	1.2
Deep3DBox [25]	10.8	1.9	25.6	**1.0**
Proposed	**13.4**	**1.6**	**30.2**	1.1

Ablation Study - Baselines

- Baseline-A: Assess the importance of the cross-modality representation learning.



- Baseline-ZDDA: Assess the importance of learning the keypoints and their view-invariant representations.



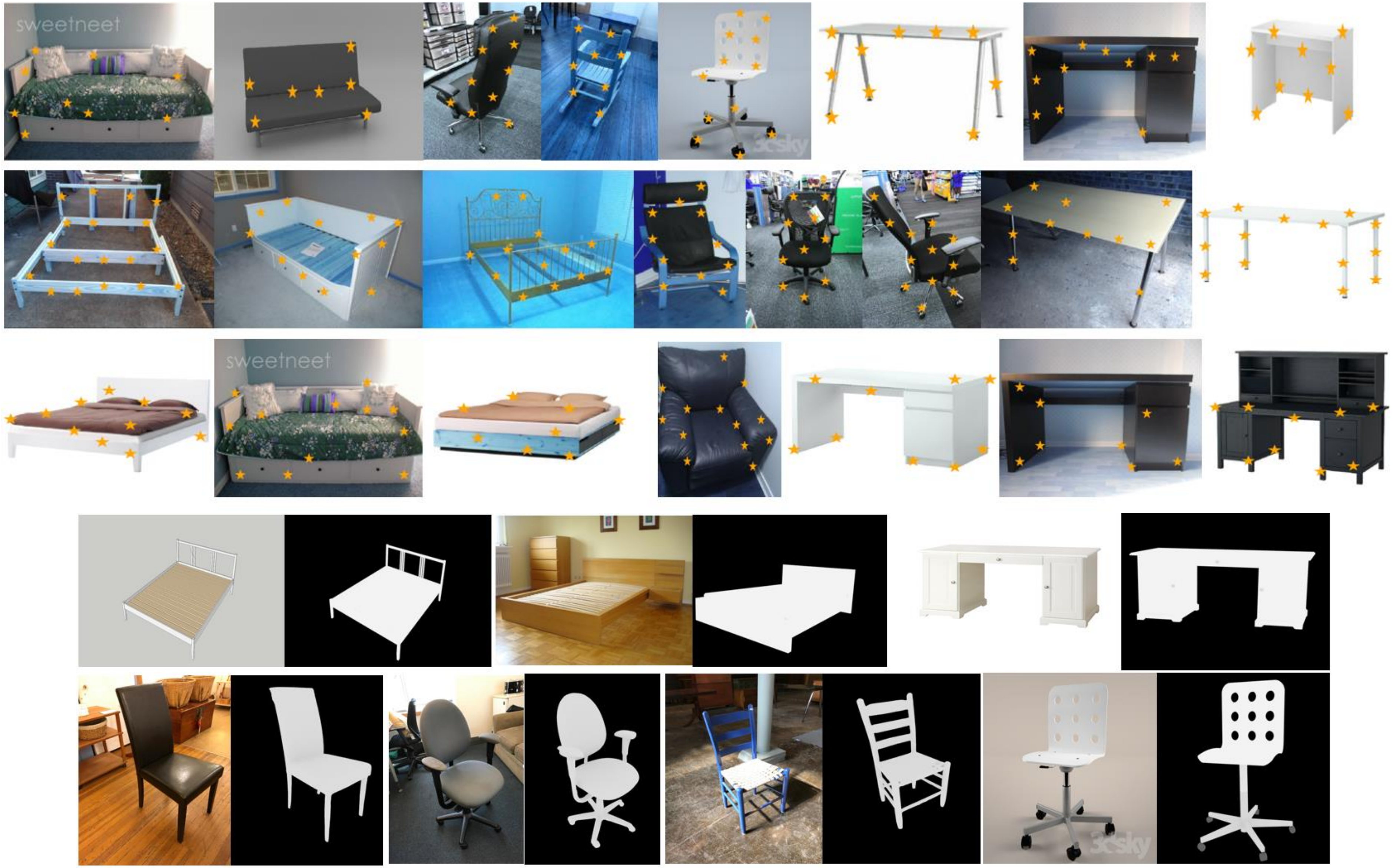
Ablation Study – Results

- Training and testing on Pix3D
  - Both baselines underperform by a large margin.

Category	Bed		Chair	
Metric	$Acc_{\frac{\pi}{6}}$	MedErr	$Acc_{\frac{\pi}{6}}$	MedErr
Baseline-A	7.3	1.7	3.3	2.0
Baseline-ZDDA	21.8	1.5	11.5	1.7
Proposed	**50.8**	**0.5**	**31.2**	1.0

- Test on category instances not seen during training
  - In practice impossible to have 3D models for all object instances.
  - Our method shows robustness to unseen instances.

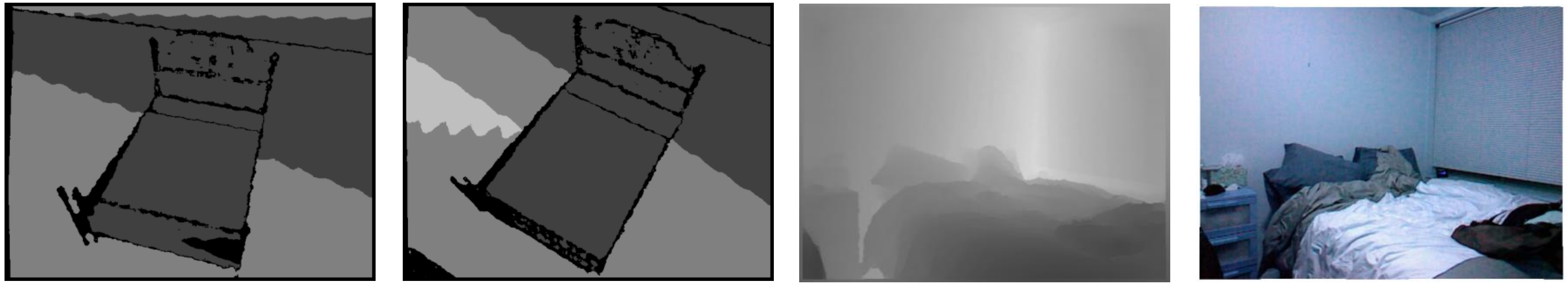
Category	Bed		Chair	
Metric	$Acc_{\frac{\pi}{6}}$	MedErr	$Acc_{\frac{\pi}{6}}$	MedErr
Baseline-A	9.7	1.9	3.7	1.9
Baseline-ZDDA	4.9	2.3	7.6	1.9
Proposed	**45.1**	**0.6**	**21.2**	**1.2**



***G. Georgakis**, S. Karanam, Z. Wu, J. Kosecka,*
Learning Local RGB-to-CAD Correspondences for Object Pose Estimation, ICCV 2019

Limitation

- We wish to demonstrate how our method can be trained with any available auxiliary data.
- **Problem:** There no suitable large-scale datasets to train our method, without relying on annotations to align RGB and depth (for branches C,D).
- We collected limited number of quadruplet training examples from NYUv2 dataset and got poor performance.



A B C D

Metric	$\text{Acc}_{\frac{\pi}{6}}$	MedErr
Bed	24.0	1.0
Chair	15.2	1.6

Conclusion

- 3D object pose estimation framework which uses textureless 3D models and does not require explicit 3D annotations in RGB images.
- End-to-end formulation for discovering the keypoints through a relative pose estimation objective.
- Learning general representation that can be matched between RGB and depth images helps our model to generalize better to new datasets.