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

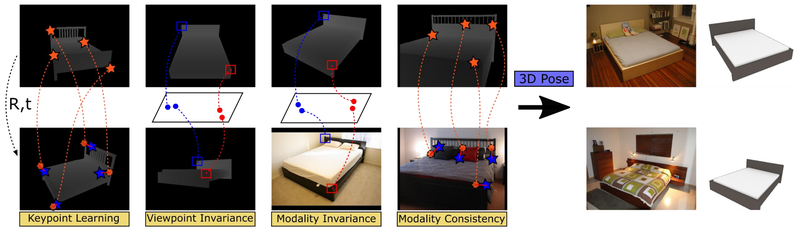
Georgios Georgakis[1], Srikrishna Karanam[2], Ziyan Wu[2], and Jana Kosecka[1]
George Mason University[1], Siemens Corporate Technology[2]

Challenges

- Difficult to obtain RGB images with ground-truth 3D geometry.
- Reliance on accurate annotated images limits generalizability and scalability.
- Large appearance gap between RGB and synthetic data.

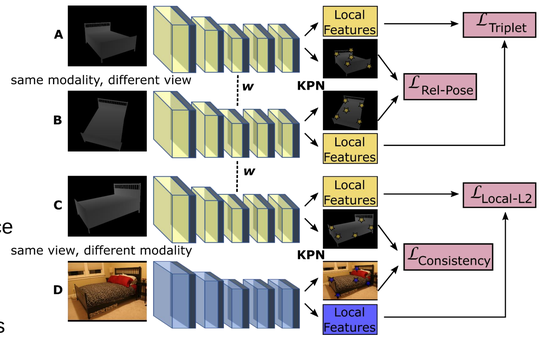
Contributions

- A new framework for 3D object pose estimation using texture-less CAD models without explicit 3D pose annotations for the RGB images.
- An end-to-end learning approach for keypoint selection optimized for the relative pose estimation objective.
- State-of-the-art results in cross-dataset evaluation, and demonstration of the generalization capability of our method to new instances.



Approach Overview

- Learn keypoints and their descriptors from depth images rendered from CAD models (A,B).
- Transfer this knowledge to the RGB domain (C,D).
- Four constraints that enforce viewpoint and modality invariance of local features, and learn how to select keypoints consistently across modalities.



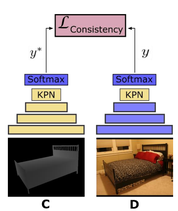
Keypoint Learning

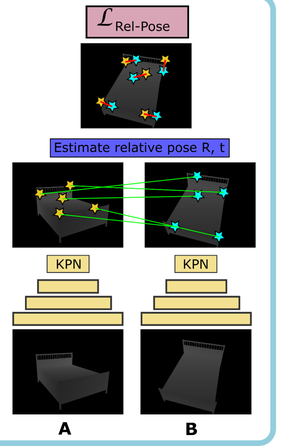
- **Relative pose loss:** For a weighted set of corresponding points $P = \{p_1, p_2, \dots, p_n\}$, $Q = \{q_1, q_2, \dots, q_n\}$ find the rigid transformation for which the re-projection error is minimum.

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

- **Keypoint consistency loss:** Align keypoint predictions on rendered depth and RGB.

$$L_{cons} = -\frac{1}{n}\sum_{i=1}^{n} y_i^c \log y_i^p$$





Descriptor Learning

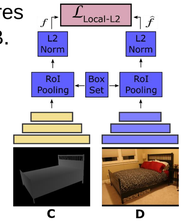
- **Triplet loss:** Enforce viewpoint invariance by sampling triplets from views A and B.

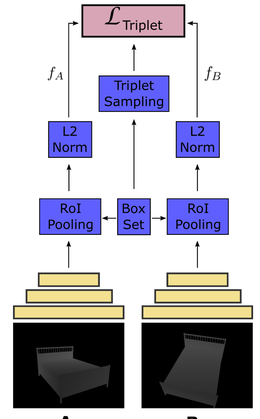
$$L_{tri} = \sum_{i}^{N} max(0, \|f_i^{\alpha} - f_i^{p}\|^2 - \|f_i^{\alpha} - f_i^{n}\|^2 + m)$$

- **Local-L2 loss:** Align local features between rendered depth and RGB.

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

- **Overall loss:** $L = \lambda_1 L_{RP} + \lambda_2 L_{cons} + \lambda_3 L_{tri} + \lambda_4 L_{loc}$





Experimental Results

1) Comparison with supervised approaches

- Training on Pix3D – Testing on Pascal3D+
- Evaluation metric: Geodesic distance: $\Delta(R_1, R_2) = \frac{\|\log R_1^T R_2\|_F}{\sqrt{2}}$

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

2) Model transferability

- Test on category instances not seen during training (Pix3D)

Category	Bed					Chair				
Metric	Az.	El.	Pl.	$Acc_{\frac{\pi}{6}}$	MedErr	Az.	El.	Pl.	$Acc_{\frac{\pi}{6}}$	MedErr
Baseline-A	38.2	39.6	30.6	9.7	1.9	28.6	41.4	20.3	3.7	1.9
Baseline-ZDDA	29.9	39.6	22.2	4.9	2.3	30.1	44.6	21.5	7.6	1.9
Proposed-joint	66.7	50.0	62.5	29.2	0.9	43.7	50.4	31.3	15.1	1.4
Proposed-alternate	**75.7**	**61.1**	**74.3**	**45.1**	**0.6**	**52.0**	**57.4**	**38.0**	**21.2**	**1.2**