SPARC: Sparse Render-and-Compare for CAD model alignment in a Single RGB Image

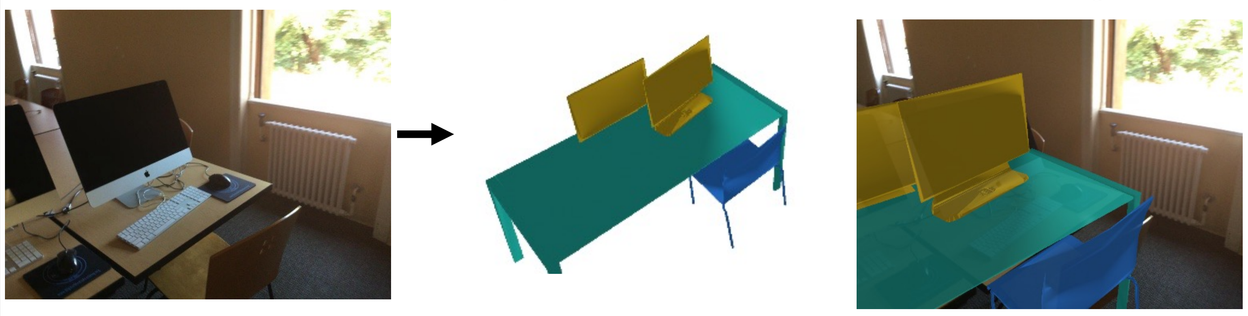
Florian Langer, Gwangbin Bae, Ignas Budvytis, Roberto Cipolla
Department of Engineering, University of Cambridge

BMVC 2022

Problem

From a single RGB image predict aligned 3D shapes that represent the given scene.

RGB Image → **Aligned 3D Shapes** → **Overlays**

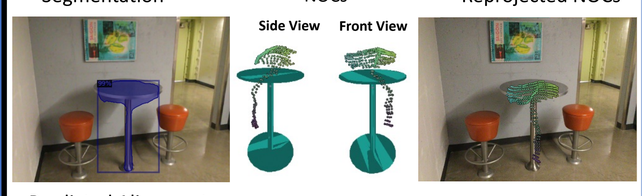


Related Work

Normalised Object Coordinates

ROCA [1]

Segmentation | NOCs (Side View, Front View) | Reprojected NOCs



Predicted Alignment

Limitations:
- NOCs are difficult to predict
- Ambiguity in alignments at train time cause displacements at test time
- Scale is predicted directly from the RGB image

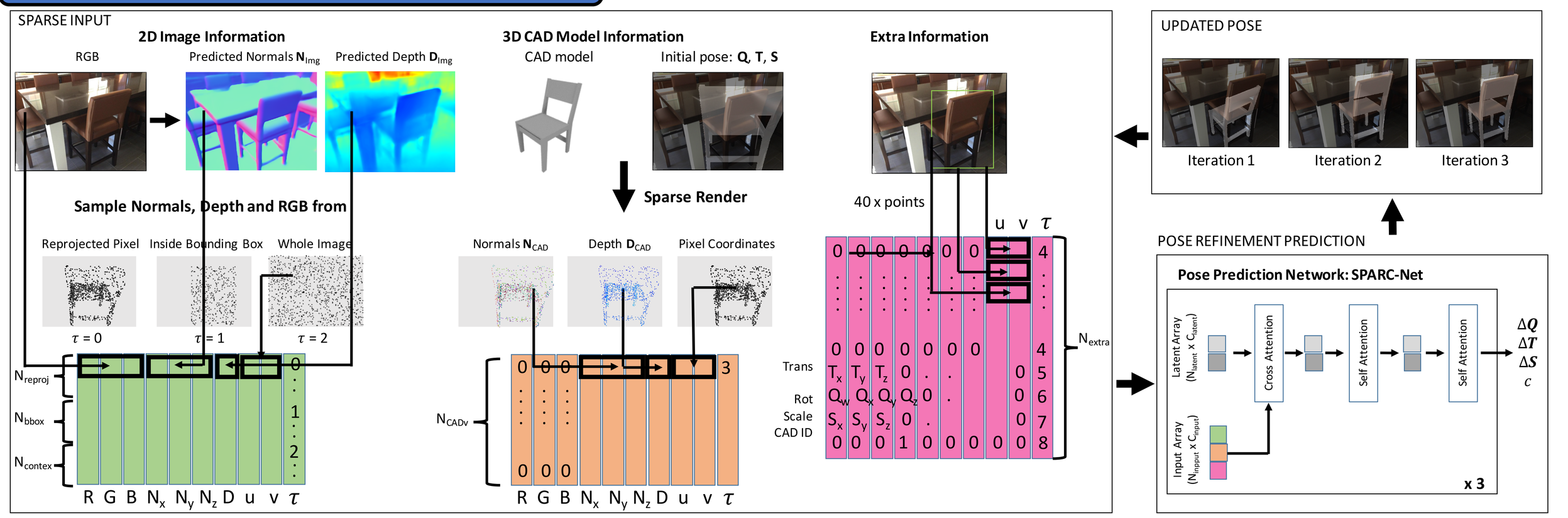
Render-and-Compare

Im2CAD [2]

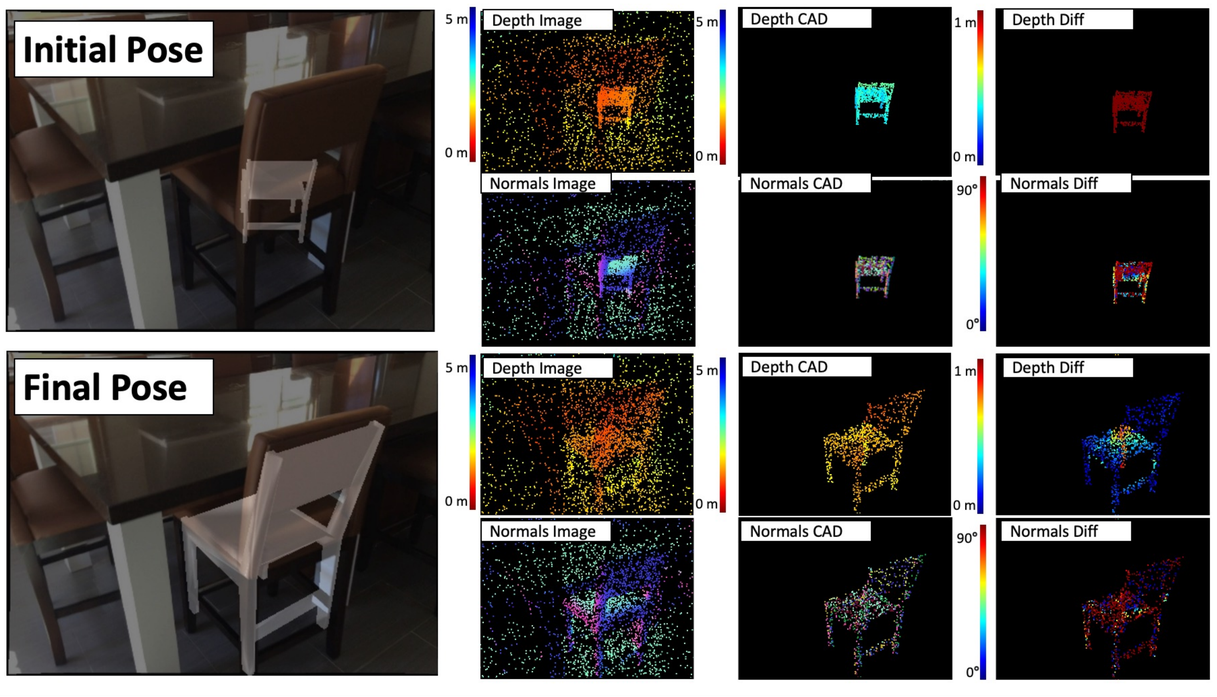
Geometric Correspondence Fields [3]

Limitations:
- Slow as require large number of iterations ([2]: 250 , [3]: 1000)
- Slow as perform and process full render (Total time: [2] 4 min, [3] 36s)
- Require very good initialisation

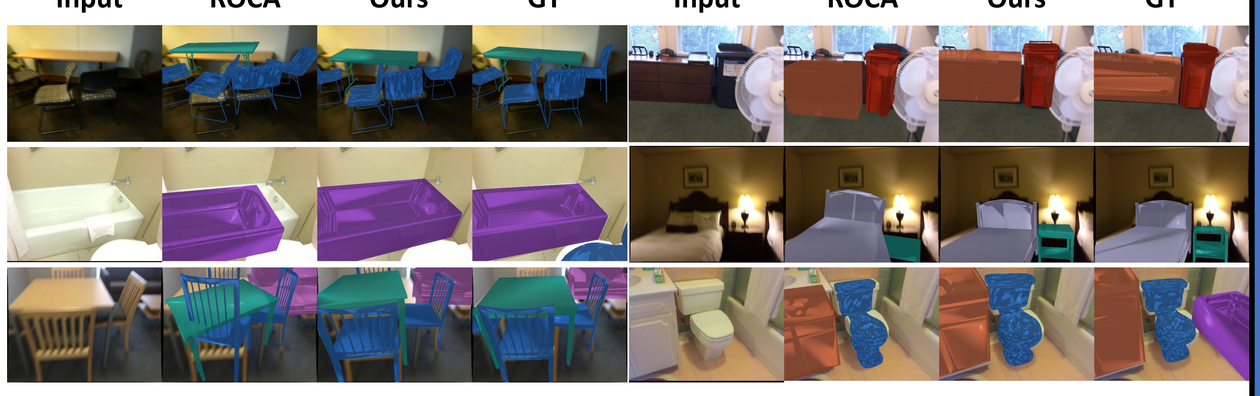
Method



Visualising Inputs



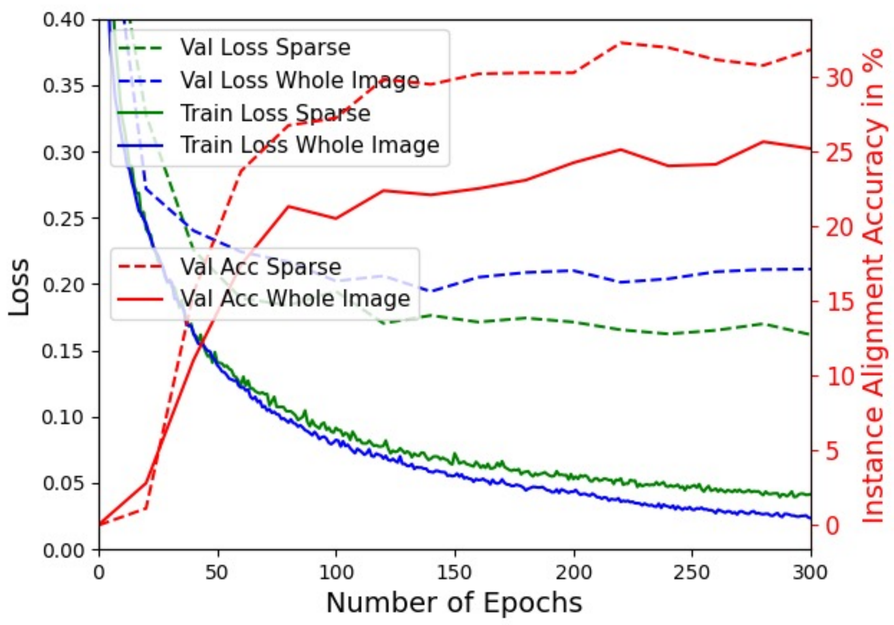
Qualitative Results - ScanNet



Quantitative Results - ScanNet

Method	bathtub	bed	bin	bkshlf	cabinet	chair	display	sofa	table	**class**	**instance**
Number of Instances #	120	70	232	212	260	1093	191	113	553	9	2844
Total3D-ODN [4]	10.0	2.9	16.8	2.8	4.2	14.4	13.1	5.3	6.7	8.5	10.4
Mask2CAD-b5 [5]	7.5	2.9	24.6	1.4	5.0	29.9	13.1	5.3	5.6	10.6	16.7
ROCA [1]	20.8	8.6	26.3	9.0	13.1	39.9	**24.6**	10.6	12.7	18.4	25.0
SPARC-Net (ours)	**25.8**	25.7	24.6	**14.2**	**20.8**	51.5	17.8	**28.3**	15.4	24.9	31.8
SPARC-Net + ROCA rot init	25.0	**30.0**	**36.2**	**14.2**	19.2	**52.3**	20.4	**28.3**	**20.1**	**27.3**	**34.1**

Sparse Inputs Reduce Overfitting



- Using a fraction of available pixels (here 3 %) reduces gap between training and validation loss
- Validation accuracy is increased from 25 % to 32 %

Discussion

+ Sparse Render-and-Compare produces **more accurate** alignments **compared to NOCs.**
+ Sparse Inputs can be **rendered very fast** and **processed efficiently due to cross attention**.
+ Predicting **pose update steps reduces the number of iterations** needed to just **3** leading to a total runtime of **110 ms** (compared to traditional render and compare requiring **100s** of update steps with run times of ca. **30s**).
— SPARC-Net does not seem able to make full use of all available information (particularly precise normal estimates) and does not improve when using more than 3 refinement steps.

References

[1] Gümeli, C., Dai, A., Nießner, M. *ROCA: Robust CAD Model Retrieval and Alignment from a Single Image.* CVPR (2022)

[2] Izadinia, H., Shan, Q., Seitz, S. *Im2CAD*. CVPR (2017)

[3] Grabner, A. Wang, Y. Zhang, P., Guo, P., Xiao, T., Vajda, P., Roth, P., Lepetit, V. *Geometric Correspondence Fields: Learned Differentiable Rendering for 3D Pose Refinement in the Wild*. ECCV (2020)

[4] Nie. Y, Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J. *Total3DUnderstanding: Joint Layout, Object Pose and Mesh Reconstruction for Indoor Scenes from a Single Image.* CVPR (2020)

[5] [14] Kuo, W., Angelova, A., Lin, T., Dai, A. *Mask2CAD: 3D Shape Prediction by Learning to Segment and Retrieve.* ECCV (2020)