Leveraging Geometry for Shape Estimation from a Single RGB Image - Supplementary Material

Florian Langer
fml35@cam.ac.uk
Ignas Budvytis
ib255@cam.ac.uk
Roberto Cipolla
rc10001@cam.ac.uk

Department of Engineering
University of Cambridge
Cambridge, UK

---

1 Pretrained SuperPoint Network

We use a SuperPoint [1] network for detecting keypoint matches between real RGB images and rendered CAD models. We use an off-the-shelf version of SuperPoint [1] with pre-trained weights and confidence threshold of 0.015. We include a brief summary of the training process below. More information can be found in [1]. SuperPoint [1] was initially trained to detect corners of triangles, quadrilaterals, lines, cubes, checkerboards and stars in synthetic images. The initial baseline detector that was trained on synthetic images was then applied to real images. By performing a set of random homographic adaptations (consisting of random translation, scaling, in-plane rotations and perspective distortions) interest points are detected in the adapted images. These are then accumulated in the original image and serve as pseudo-ground truth interest points which the network is trained to detect. The extensive training for detecting corners both in real and in synthetic images makes SuperPoint [1] suitable for our cross-domain matching task.

2 Ablations

We perform a series of ablation experiments (see Table 1).

- **Pose estimation.** We investigate the accuracy of our system as a function of the number of matches that are used for the pose estimation (see Table 1 a). Similar performance is achieved when using 3, 4 and 5 keypoint matches. This is due to a trade-off between using a low number of matches where more images have at least the given amount of correct matches and using a higher number of matches which allows for preciser poses when the minimum number of correct matches is given. As the number of keypoints is increased to 6 or 7 we note a drop in the AP$^{\text{mesh}}$ score as the large number of images which have less than 6 correct matches (and which therefore will

---

a)

Number of matches	3	4	5	6	7
S1	31.5 (51%)	30.3 (42%)	31.1 (35%)	28.5 (30%)	26.0 (24%)
S2	7.2 (32%)	7.4 (26%)	7.2 (21%)	6.8 (16%)	6.4 (13%)

b)

Number of NN	1	5	10	15	20	GT
S1	31.1 (65 %)	36.3 (80%)	37.9 (85%)	39.3 (88%)	37.6 (91%)	41.2 (100%)
S2	6.9 (24%)	14.2 (43%)	17.1 (55%)	18.4 (62%)	19.5 (67%)	33.4 (100%)

c)

Number of CAD models in database	chair (130)	table (51)	desk (17)	sofa (15)	book-case (14)	bed (13)	misc (9)	tool (5)	war-drobe (2)
10	0.8 (0.8)	9.5 (1.0)	1.4 (0.0)	42.1 (44.0)	8.8 (0.3)	19.8 (16.1)	0.6 (0.0)	3.8 (10.5)	10.1 (0.0)
30	3.2 (2.9)	11.7 (4.2)	3.1 (0.0)	34.8 (45.6)	10.2 (11.0)	21.4 (15.2)	-	-	-
50	3.5 (2.6)	13.0 (4.0)	-	-	-	-	-	-	-
100	4.3 (3.5)	-	-	-	-	-	-	-	-

Table 1: Numbers displayed are AP^{mesh} scores averaged across all classes. a) The ablation for the number of matches was performed with top 1 retrieval results and Swin-Transformer [4] segmentation masks. Numbers in brackets indicate how many of the test images have at least n correct matches. Here a match is considered correct if the reprojection error of the reprojected keypoint under the ground truth pose is less than 5 pixel. b) For the ablation on the number of nearest neighbours numbers in brackets indicate the retrieval accuracy i.e. for how many of the images the correct object is among the top n retrieved. GT indicates ground truth retrieval of the correct object in the most similar pose. c) When investigating the adaptation performance for variable size datasets ground truth masks are used. Numbers in brackets indicate the AP^{mesh} score when no shape adaptation is performed and numbers next to category names indicate the total number of CAD models in the $S2$ train split.

produce inaccurate poses) outweigh the few images for which poses can be computed more accurately.

- **Shape retrieval.** We ablate our system in terms of the number of nearest neighbour CAD model renderings that are retrieved (see Table 1 b). Note that two renderings of the same CAD model, but rendered from different orientations, are considered two separate retrievals for which keypoint matching and pose estimation is performed independently. For the S1 split the class-averaged AP^{mesh} score for just the top 1 retrieval is 31.1 which is already high. It can be further improved by increasing the number of nearest neighbours that are considered. This effect is even more pronounced on the S2 split where the class-averaged AP^{mesh} score increases from 6.9 for top 1 retrieval to 19.5 for top 20 retrieval. The reason for this is that the worse segmentation mask quality on the S2 split compared to the S1 split leads to worse retrieval results (24% accuracy compared to 65 % accuracy). At this low retrieval accuracy the benefits of considering additional retrieved shapes is larger compared to the case when the retrieval accuracy is already high.

- **CAD model database.** In order to investigate the dependency of the system on the available database size we vary the number of CAD models to which the network has access at test time (see Table 1 c). Results on the table, bookcase or wardrobe class demonstrate that when performing shape adaptation even extremely small sets of CAD models can be used to estimate object shapes. Note that the total number of CAD models in the Pix3D [5] dataset is very small. For realistic settings one can easily use hundreds or thousands of CAD models per category therefore allowing for even more precise object shape predictions.

- **Pose selection.** Finally, we compare the accuracy of the poses when selecting poses based on the minimum distance of reprojected keypoints compared to the approximated silhouette overlap. We obtain an AP^{mesh} score of 17.1 (compared to 37.8) on

	S1	AP	AP 50	AP 75	bed	book-case	chair	desk	misc	sofa	table	tool	ward-robe
Average of all pairwise distances	Ours (Mesh R-CNN) Top 1	29.2	46.1	27.0	31.7	15.7	30.6	22.5	33.6	45.5	24.7	22.2	36.4
	Ours (Mesh R-CNN) Top 10	34.4	52.5	33.2	27.0	16.3	34.1	27.0	49.2	47.6	33.2	43.7	32.7
	Ours (Swin) Top 1	31.1	49.1	29.1	21.2	19.9	29.9	25.1	35.5	41.6	25.9	43.9	36.9
	Ours (Swin) Top 10	33.6	50.9	32.8	20.9	16.1	37.2	26.8	42.2	43.2	34.2	45.7	36.6
	Ours (GT) Top 1	54.4	67.8	53.2	65.0	37.8	60.7	50.2	63.9	60.9	57.6	42.4	54.5
	Ours (GT) Top 10	60.5	72.9	60.5	67.4	35.1	65.1	60.5	68.0	62.9	67.9	71.9	45.9
Average of 20 % largest pairwise distances	Ours (Mesh R-CNN) Top 1	30.9	46.2	30.0	32.0	15.3	30.6	25.6	37.0	47.3	27.1	21.0	42.5
	Ours (Mesh R-CNN) Top 10	37.1	55.4	35.3	32.3	25.4	37.8	29.5	48.9	53.7	37.7	29.7	39.0
	Ours (Swin) Top 1	31.1	49.6	29.4	20.6	19.6	30.2	23.6	37.4	41.7	24.7	43.3	38.9
	Ours (Swin) Top 10	37.8	56.1	36.8	24.3	25.0	40.8	28.4	51.4	50.0	36.9	39.6	44.3
	Ours (GT) Top 1	57.6	72.1	55.9	60.1	49.6	54.5	64.7	68.7	62.7	63.1	41.7	53.0
	Ours (GT) Top 10	71.5	82.5	71.7	66.9	53.7	63.8	85.1	86.5	69.3	78.2	71.0	69.0

S2	AP	AP 50	AP 75	bed	book-case	chair	desk	misc	sofa	table	tool	ward-robe
Ours (Mesh R-CNN) Top 1	7.4	16.8	5.2	9.6	0.5	18.5	1.6	1.5	25.7	1.6	6.8	0.5
Ours (Mesh R-CNN) Top 10	14.0	26.4	12.1	23.4	1.4	35.2	4.3	0.4	39.8	4.9	10.5	5.7
Ours (Swin) Top 1	7.0	15.9	4.3	7.3	3.2	18.7	0.6	2.1	24.3	3.6	1.6	1.2
Ours (Swin) Top 10	15.2	30.7	10.8	20.6	4.9	38.5	4.1	6.9	34.6	7.7	9.6	10
Ours (GT) Top 1	37.0	53.3	34.4	43.4	32.7	71.6	25.3	45.8	41.7	32.8	31.0	8.9
Ours (GT) Top 10	60.7	74.1	59.9	54.0	50.8	91.2	70.0	74.5	58.5	59.9	41.0	46.2
Ours (Mesh R-CNN) Top 1	7.9	17.1	5.8	9.4	1.5	19.2	1.6	3.4	27.0	1.7	6.3	1.1
Ours (Mesh R-CNN) Top 10	16.6	31.0	14.6	25.3	2.6	41.1	6.2	4.5	38.5	6.2	16.8	8.4
Ours (Swin) Top 1	6.9	15.8	4.4	7.0	2.6	18.9	0.4	1.3	25.2	3.5	1.4	1.5
Ours (Swin) Top 10	17.1	32.2	13.8	24.1	7.1	44.5	4.1	5.9	40.0	8.3	9.9	11.9
Ours (GT) Top 1	43.7	59.7	41.3	58.8	38.0	74.0	40.8	43.6	42.1	44.9	40.2	10.5
Ours (GT) Top 10	70.7	83.7	70.4	74.7	56.1	94.1	83.4	75.6	69.2	76.9	56.2	50.3

Table 2: We compare the results we obtain when using different metrics for selecting the final pose. For the results in the top halves of the tables we compute the average over all pairwise distances used to approximate the silhouette overlap of the reprojected CAD model with the predicted segmentation mask. For the results in the lower halves we compute the average over only the 20% furthest pairwise distances. This is a stronger signal for a good pose as it weighs object outlines (along which pairwise distances are large for bad poses) more compared to area overlaps.

the S1 split and 8.1 (compared to 17.1) on the S2 split. This demonstrates the need for using the estimated silhouette overlap for pose prediction.

3 Estimating the Silhouette Overlap

In order to estimate object poses all possible quadruplets of keypoint matches are sampled and their corresponding poses are computed. Choosing the final pose based on the minimum distance of reprojected keypoints often results in sub-optimal poses (see Section 2). Pixel-level inaccuracies of the matches can lead to poses which minimise the reprojection error of the keypoints, but which are very inaccurate. To avoid selecting these poses, poses are chosen based on the estimated silhouette overlap of the reprojected CAD model and the predicted segmentation mask. For this purpose 1000 points are sampled from the retrieved CAD model and reprojected for the current pose estimate. These points are compared to 1000 points sampled inside the predicted segmentation mask by computing the minimum pairwise distance from the reprojected point to points sampled in the segmentation mask and vice versa. Rather than taking the average of all pairwise distances we found it beneficial to only take the average of the 20% of the largest distances. This poses a stronger signal for matching object outlines and leads on average to more accurate poses (see Table 2).

4 Similarity of Train and Test CAD Models

We evaluate our proposed model on the Pix3D [5] dataset for which [2] introduced two data splits. For the S1 split the 10,069 images are randomly split into 7539 train images and 2530 test images. Under this split all CAD models are seen during training. For the S2 split train and test images are split such that the test images contain CAD models that were not present in the training images. The challenge is therefore to construct a system that given an input image is able to retrieve an unseen CAD model and precisely predict its pose.

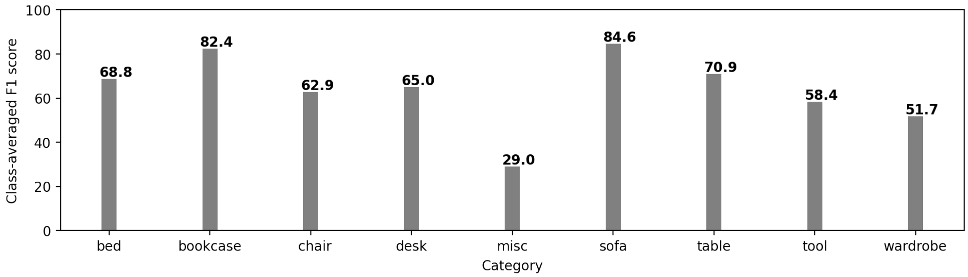

Figure 1: Grey bars provide an indicator for the similarity between CAD models seen during training and unseen CAD models used for testing under the S2 split of Pix3d [5]. Quantitatively grey bars show the class average when the F1 score is computed between every unseen CAD model and its closest matching CAD model (in terms of the F1 score) from the seen ones.

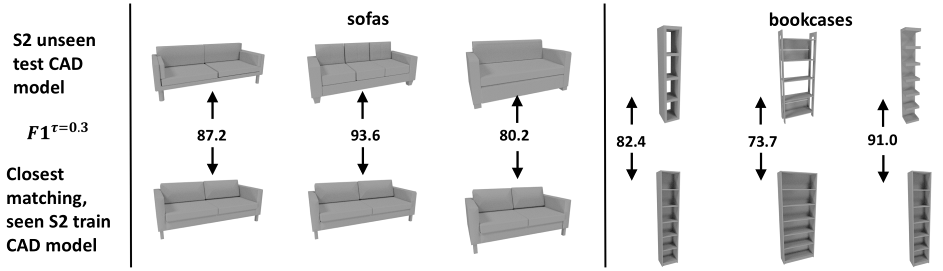

Figure 2: Visualisation of selected test CAD models and their closest matching train CAD models in terms of the F1 score. We note the strong similarity (both visually and in terms of the F1 score) between sofas in the test split and sofas from the train split. While bookcases in the test split also have close matching CAD models in the train split in terms of the F1 score, they differ significantly in their visual appearance. This increases the difficulty for retrieval at test time and explains the poor performance of [3] on bookcases compared to sofas.

We have demonstrated in our main work that the geometric approach that we follow is more accurate compared to directly predicting object poses [3] on the S2 split of Pix3D [5]. Further, we will show here that the good performance [3] achieves on sofas does not require it to retrieve unseen CAD models as for every unseen CAD model in the test images there is a closely fitting CAD model among the seen training CAD models. We quantify this by computing the F1 score at $\tau = 0.3$ between unseen test CAD models and their closest matching CAD models (in terms of F1 score) from the seen ones. We perform this calculation for all unseen CAD models and compute the mean to obtain class averages. These are plotted in gray in Figure 1. Note here that the averaged best-possible F1 score for sofas is 84.6 which is exceptionally high compared to other class averages. This strong similarity (see Figure 2 for selected test CAD models and their closest-matching CAD models from the train set) allows [3] to make accurate shape predictions without retrieving unseen CAD models. We also note that [3] performs poorly on bookcases despite a high class-averaged F1 score between test CAD models and their closest matching train CAD models. The reason for this is that while good candidate CAD models exist in the seen train CAD models in terms of the

F1 score, their different visual appearance (see right side Figure 2) makes them difficult to retrieve at test time.

References

[1] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.

[2] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh R-CNN. In *Proc. IEEE Int. Conf. on Computer Vision*, Seoul, Korea, October 2019.

[3] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2CAD: 3D Shape Prediction by Learning to Segment and Retrieve. In *Proc. 16th European Conference on Computer Vision*, Glasgow, UK, August 2020.

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proc. IEEE Int. Conf. on Computer Vision*, Montreal (Virtual), October 2021.

[5] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B Tenenbaum, and William T Freeman. Pix3D: Dataset and Methods for Single-Image 3D Shape Modeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, Utah, USA, June 2018.