FastCAD: Real-Time CAD Retrieval and Alignment from Scans and Videos - Supplementary Material

Anonymous ECCV 2024 Submission

Paper ID #2302

A Run-Time Analysis of Competing Methods for RGB Videos

Due to its efficient design, FastCAD can integrate information from a new frame into its CAD-based reconstruction in just 100 ms (50 ms for running [6] and 50 ms for running FastCAD itself). ODAM [7] requires 166 ms for its object detection and object association and a further 200 ms to optimise each object pose. As different object poses can be optimised in parallel, their total time for integrating information from a new frame is 366 ms. Vid2CADs [8] detection method takes 200 ms per frame, the tracking takes an additional 500 ms per frame, and the optimisation over poses takes 2.5 seconds, which gives it a run-time of 3200 ms. RayTran [10] does not provide any information in terms of run-time and the authors were not able to share any information regarding this with us. However, from their method design and their training requirements (8 x 16 GB GPU using 16-bit float arithmetic), one can infer that their method is extremely compute-intensive and can not integrate new information, except by running it again on all frames.

B Additional Results for the Reconstruction and Shape Accuracy

Tab. 1 shows the alignment, reconstruction and shape accuracy on Scan2CAD for different settings of μ (see Sec. 4.2 of the main paper for the definition of μ). We find that, in general, the trends observed at $\mu = 0.7$ can also be found at $\mu = 0.5$ and $\mu = 0.9$. In addition to providing results at extra settings for μ, we include ablations for the number of input points in the last section of Tab. 1. When reducing the number of input points from its default value of 50 K, we observe a graceful decline in the reconstruction accuracy and alignment accuracy. Note that even using 5 K points for the entire scene still yields good reconstruction and alignment accuracy. We also find that the shape accuracy remains particularly high even for very low numbers of points. This means that FastCAD can still predict shape embeddings well in this data regime and the challenge lies more in accurate CAD alignment predictions.

	Method	Alignment Acc.	Recon. Acc. $\mu = 0.5$	Recon. Acc. $\mu = 0.7$	Recon. Acc. $\mu = 0.9$	Shape Acc. $\mu = 0.5$	Shape Acc. $\mu = 0.7$	Shape Acc. $\mu = 0.9$	time [ms]
Competing Methods - Input RGB-D Scan									
Input RGB-D Scan	ScanNotate* [1]	78.2	78.6	60.1	22.2	95.1	83.5	45.7	660000
	Ours (Scan)	61.7	68.9	41.7	7.0	95.3	83.1	44.1	50
Competing Methods - Input RGB Video									
Input RGB Video	Vid2CAD* [8]	38.6	38.0	22.9	6.2	81.2	76.6	69.5	3200
	Ours (Video)	48.2	52.2	24.7	3.7	94.7	79.8	41.7	100
	Ours (Video, same retrieval Vid2CAD)	48.2	52.9	29.6	7.7	95.3	87.7	71.4	100
Ablation Experiments- Input RGB-D Scan									
Embedding Distillation	2-step retrieval: pred bbox	61.7	43.4	15.6	1.2	77.7	51.0	17.6	104
	2-step retrieval: nearest GT bbox	61.7	61.7	30.6	4.1	93.1	78.1	36.1	104
	Embedding distillation	61.7	68.9	41.7	7.0	95.3	83.1	44.1	50
Auxiliary Tasks for Training Encoder	Triplet	62.3	67.9	38.3	5.5	94.8	81.1	41.1	50
	Tiplet + Chamfer	61.0	67.5	38.7	7.3	95.2	82.0	42.3	50
	Triplet + Segmentation	61.3	68.8	41.5	7.5	95.7	84.3	43.6	50
	Triplet + Chamfer + Segmentation	61.7	68.9	41.7	7.0	95.3	83.1	44.1	50
Encoder Architecture	PointNet [9]	61.5	61.2	29.6	4.0	92.0	74.0	30.9	50
	Perceiver [5]	62.3	67.9	38.3	5.5	94.8	81.1	41.1	50
Different Input Sources	ScanNet (Gray)	60.4	67.7	40.1	6.6	95.6	82.9	43.1	50
	DG Recon [6] (Gray)	48.2	52.2	24.7	3.7	94.7	79.8	41.7	100
Input Points	5 K	51.4	56.8	29.8	4.0	95.3	81.9	43.0	42
	10 K	54.9	60.1	34.2	5.2	95.2	82.0	43.6	46
	20 K	57.8	65.6	37.8	6.3	95.5	83.3	44.4	47
	50 K	61.7	68.9	41.7	7.0	95.3	83.1	44.1	50
	100 K	61.8	69.8	41.7	7.5	95.5	83.6	44.0	50

Table 1: Alignment, reconstruction and shape accuracy on Scan2CAD [2] in comparison to competing methods and for various ablations. All accuracies are percentages and higher is better. Compared to Tab. 2 in the main paper, here we provide additional results for different settings of the threshold μ for computing the reconstruction and shape accuracy as explained in Sec. 4.2. Note that ScanNotate [1] initialises its CAD alignments from their ground-truth poses and Vid2CAD [8] constraints its CAD retrieval to the very small ground-truth scene pool, making some of their results not exactly comparable to ours.

C Poor Performance on Display Class

When computing the CAD alignment accuracy per class in Tab. 1 in the main paper, we find that FastCAD performs significantly worse on the "display" class compared to other classes. Using scans as input, the alignment accuracy for displays is 24.1% compared to the mean class accuracy of 52.8%. Using videos as inputs, the "display" accuracy is just 4.2% compared to the mean of 39.3%. Visually inspecting the predictions we find that the issue in the majority of cases is a wrong prediction for the object front-facing side $\hat{\boldsymbol{f}}$. Investigating this phenomenon further we found that ShapeNet [3] CAD models of the "display" class are not oriented consistently. Out of 149 different "display" CAD models in the training set, 28 are facing the opposite direction. This results in a confusing training signal and explains the poor performance that is observed for this class. Ignoring this class would make our relative performance compared to competing methods even better.

D Further Visualisations

We include extra qualitative visualisations. Fig. 1 provides additional visualisations of FastCAD when using either the output of [6] or directly the scans from ScanNet [4] as input. While CAD alignments are of high quality in both cases, alignments from scans are consistently more accurate. This is because the reconstructions generated with [6] can be noisy or miss crucial details which poses a challenge for the subsequent CAD retrieval and alignment.

Fig. 2 shows a qualitative comparison of our method compared to Vid2CAD [8]. It can be seen that FastCAD is considerably more accurate in its CAD alignments. We believe that this is largely due to the design decision to perform CAD alignment in 3D as opposed to relying on detecting objects in 2D and matching them across frames. Inaccuracies in 2D detection and wrong associations across frames are likely the cause for many of the misalignments of Vid2CAD observed in Fig 2.

We include additional visualisations for FastCADs CAD retrieval in Fig. 3. In general, the retrieved shapes match the underlying objects very well (e.g. the retrieved tables in the first row). However, for some objects the retrieved CAD models are not closely fitting. We find this is the case particularly for objects with missing scene geometry (e.g. the third CAD retrieval for the chair in the last row) or for objects with a lot of clutter (e.g. the retrieved tables in the last row).

Finally, we provide a visualisation video (screenshot in Fig. 4) showcasing FastCAD's ability to perform accurate CAD-based reconstructions from videos online. From the start of the sequence, the aligned CAD models provide a faithful reconstruction of the underlying scene. Failure modes can include overlapping CAD models in the reconstruction when partially seen objects are revealed further as well as sub-optimal shape retrieval for certain objects.

References

1. Ainetter, S., Stekovic, S., Fraundorfer, F., Lepetit, V.: Automatically annotating indoor images with cad models via rgb-d scans. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (2023) 2, 6
2. Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A.X., Niessner, M.: Scan2cad: Learning cad model alignment in rgb-d scans. In: IEEE Conf. Comput. Vis. Pattern Recog. (2019) 2, 5, 6
3. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 2
4. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017) 2, 5
5. Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., Carreira, J.: Perceiver: General perception with iterative attention (2021) 2
6. Ju, J., Tseng, C.W., Bailo, O., Dikov, G., Ghafoorian, M.: Dg-recon: Depth-guided neural 3d scene reconstruction. In: Int. Conf. Comput. Vis. (2023) 1, 2, 5, 6, 7
7. Li, K., DeTone, D., Chen, S., Vo, M., Reid, I., Rezatofighi, H., Sweeney, C., Straub, J., Newcombe, R.: Odam: Object detection, association, and mapping using posed rgb video. In: Int. Conf. Comput. Vis. (2021) 1
8. Maninis, K.K., Popov, S., Nießner, M., Ferrari, V.: Vid2cad: Cad model alignment using multi-view constraints from videos. IEEE Transactions on Pattern Analysis and Machine Inttelligence (2022) 1, 2, 3, 6
9. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Adv. Neural Inform. Process. Syst. (2017) 2

10. Tyszkiewicz, M.J., Maninis, K.K., Popov, S., Ferrari, V.: Raytran: 3d pose estimation and shape reconstruction of multiple objects from videos with ray-traced transformers. In: Eur. Conf. Comput. Vis. (2022) 1

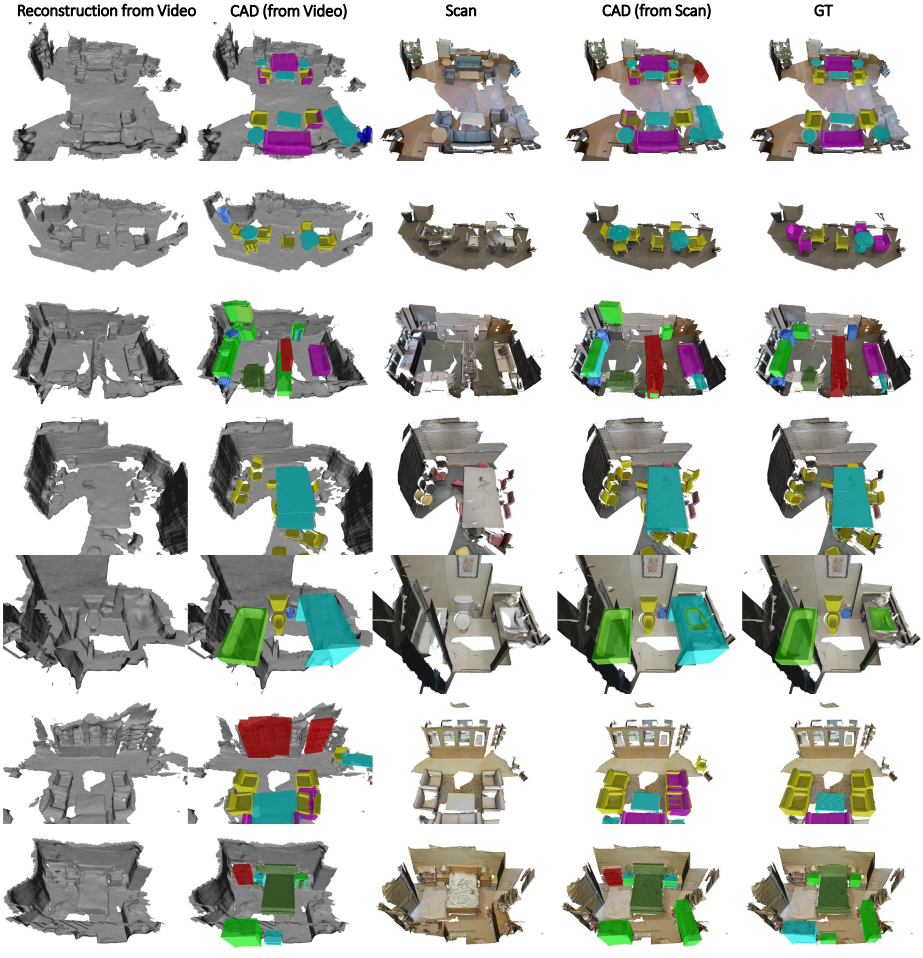

Fig. 1: Further Qualitative Visualisation on ScanNet. Column 1 shows the reconstruction generated by applying [6] to the input video. Column 2 shows the CAD retrieval and alignments predicted by FastCAD when operating on the reconstruction in column 1. Columns 3 and 4 show the input scan from ScanNet [4] and the CAD alignments FastCAD predicts for it. Column 5 shows the ground-truth CAD alignments from Scan2CAD [2].

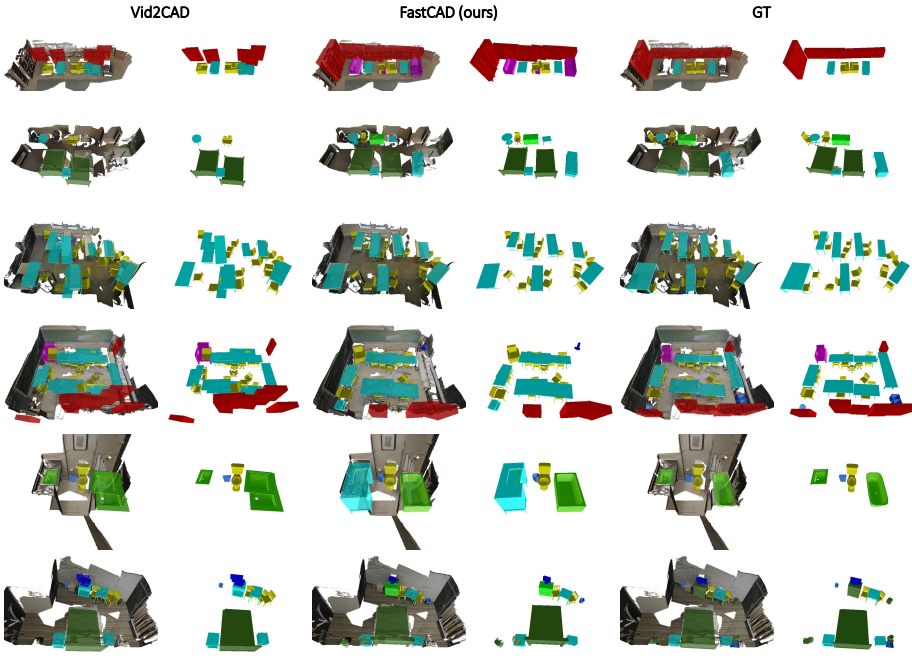

Fig. 2: Qualitative Comparison to Vid2CAD. To obtain CAD alignments, we apply FastCAD to the output of [6], which uses the video of the scene as input. For a given scene, Vid2CAD [8] limits its retrieval to the small ground truth scene pool, whereas we retrieve from all CAD models in the Scan2CAD [2] training set. We find that CAD alignments produced with Vid2CAD [8] are often noisy, not necessarily matching the actual object alignments in the scene. In contrast, CAD alignments produced with FastCAD are accurate, explaining the input scene well.

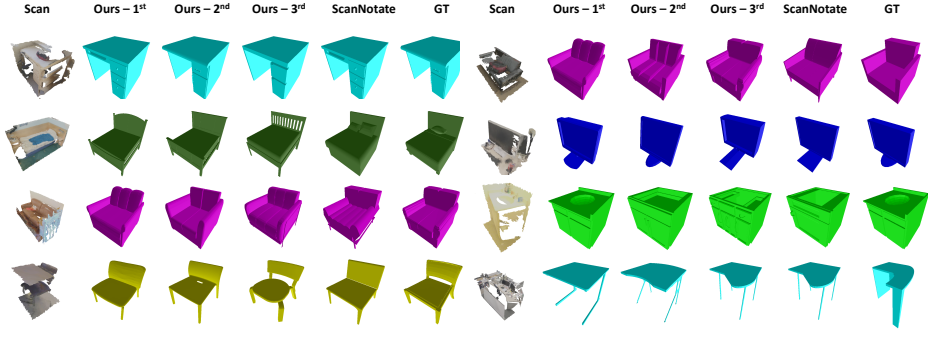

Fig. 3: Qualitative Visualisation of CAD retrievals. Note that the input to FastCAD from which a shape embedding $\hat{\boldsymbol{w}}$ is predicted is the scan of the entire scene. However, for clearer visualisation, we only show the cropped part of the scan for which a CAD model is retrieved. FastCADs CAD retrievals are of high quality, in many cases as good as those obtained with the pseudo-label generation method ScanNotate [1] or from the annotations from Scan2CAD [2] themselves.

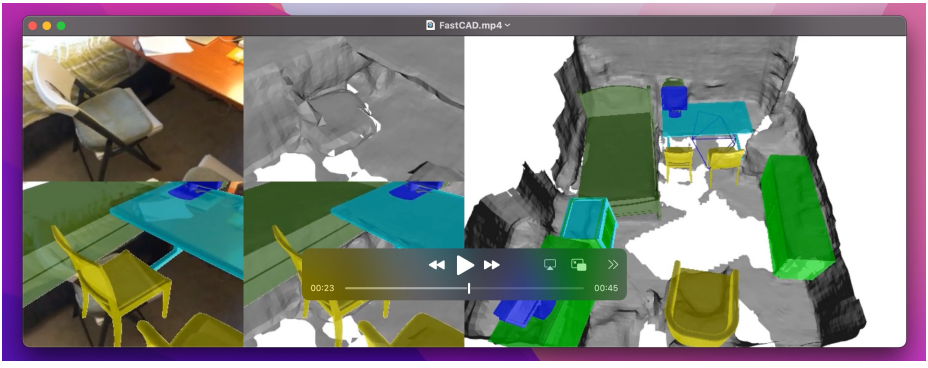

Fig. 4: Visualisation Video. We visualise our online CAD retrieval and alignments in a visualisation video. The top left images show the input RGB video and the reconstruction generated with [6]. The bottom left images show the aligned CAD models overlayed to the input video and the reconstruction from [6]. The image on the right shows the global scene reconstruction and current camera pose. One can see that FastCAD is able to create accurate CAD-based scene reconstructions for a diverse set of input scenes.