3D Semantic MapNet: Building Maps for Multi-Object Re-Identification in 3D

Vincent Cartillier[1], Neha Jain[2], Irfan Essa[1,3]
[1]GeorgiaTech [2] Meta [3] Google
vcartillier3@gatech.edu
vincentcartillier.github.io/3d_smnet.html

Abstract

We study the task of 3D multi-object re-identification from embodied tours. Specifically, an agent is given two tours of an environment (e.g. *an apartment) under two different layouts* (e.g. *arrangements of furniture). Its task is to detect and re-identify objects in 3D* – e.g. *a 'sofa' moved from location A to B, a new 'chair' in the second layout at location C, or a 'lamp' from location D in the first layout missing in the second. To support this task, we create an automated infrastructure to generate paired egocentric tours of initial/modified layouts in the Habitat simulator [7, 10] using Matterport3D scenes [3], YCB [2] and Google-scanned objects [5]. We present 3D Semantic MapNet (3D-SMNet) – a two-stage re-identification model consisting of (1) a 3D object detector that operates on RGB-D videos with known pose, and (2) a differentiable object matching module that solves correspondence estimation between two sets of 3D bounding boxes. Overall, 3D-SMNet builds object-based maps of each layout and then uses a differentiable matcher to re-identify objects across the tours. After training 3D-SMNet on our generated episodes, we demonstrate zero-shot transfer to real-world rearrangement scenarios by instantiating our task in Replica [9], and RIO [11] environments depicting rearrangements. On all datasets, we find 3D-SMNet outperforms competitive baselines. Further, we show jointly training on real and generated episodes can lead to significant improvements over training on real data alone.*

1. Multi-Object Re-identification from Tours

We consider a multi-object re-identification problem in 3D environments observed through egocentric tours. Each problem instance (or episode) is defined by a pair of egocentric tours through different layouts of the same environment – an initial layout and a modified layout. Objects in the initial layout may be moved, removed, or unchanged in the modified layout. Further, new objects may be added to the modified layout. The task is to re-identify objects present in both layouts and identify objects which have been removed or added. Notably, we do not constrict the two tours to follow the same path.

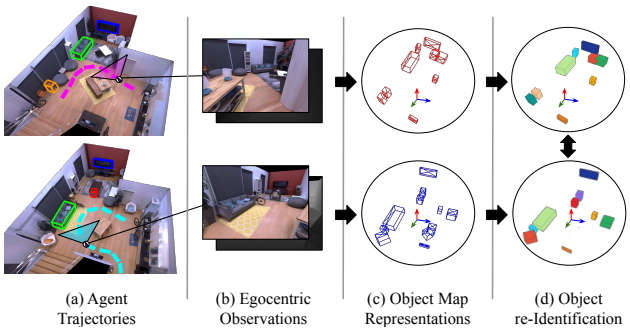

Figure 1. 3D Multi-Object Re-Identification: an agent is provided two tours of an environment (egocentric RGD-D videos with known pose). The two layouts may differ with objects added (red), removed (orange), moved (green) or unchanged (blue). The goal for the agent is to detect and re-identify objects in 3D.

To support this problem definition, we develop a procedure to generate paired egocentric tours of initial/modified layouts in the Habitat simulator [7, 10]. At a high level, we use Matterport3D (MP3D) [3] environments and insert YCB [2] and Google-scanned-objects [5] to create initial and modified layouts. We also develop an iterative sampling procedure to build trajectories and then run a simulated agent through the exploration paths to collect an RGB-D tour.

Dataset Statistics. Following the strategy described, we create 625 episodes split 461/65/126 between train/val/test. This corresponds to a total of 24k, 3k, 7k unique object pairs. Tours consist of 800 steps on average. It provides good space coverage with 78% of objects being actually observed during the tours.

2. 3D Semantic MapNet (3D-SMNet)

Our approach, named 3D Semantic MapNet (3D-SMNet) and illustrated in Fig. 2, consists of two broad components: (1) a 3D object detector that operates on RGB-D videos with known pose, and (2) an object matching module that solves correspondence estimation between two sets of 3D bounding

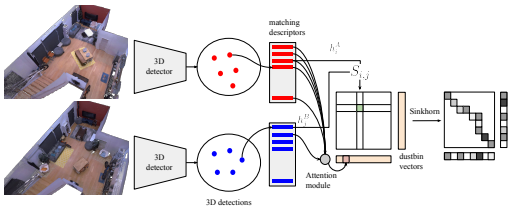

Figure 2. 3D-SMNet consists of a 3D object detector and a matching module. The 3D object detector [4] takes as input a textured point-cloud representation of the scene and outputs a set of 3D detections along with feature descriptors. The matching module computes similarity scores from the pairwise descriptors and then extends the score matrix with dustbin vectors estimated from an attention mechanism over the two sets of features to capture added/removed objects. The Sinkhorn algorithm [8] is then applied to solve the partial assignment problem.

boxes.

3. Experiments

Evaluation Metrics: We report the entire range of evaluation metrics for object re-identification in a query-to-gallery setup: Cumulative Matching Characteristics (CMC) [12] and mean Average Precision (mAP) [13]. CMC-k (or rank-k) is the probability that the correct matched object of a given query object appears in the top-k ranked object list.

Experiments with different matchers: We compare 3D-SMNet with different matchers. We experiment with the Hungarian algorithm with different score functions. We tested the L_2 and Mahalanobis distances, and learning a 1-layer mapping of the descriptors prior to using the L_2 distance to compute the score matrix. We train this one linear layer using a triplet loss function [1]. In addition, we compare our model to a Sinkhorn matcher (S) without the attention model to estimate the dustbin vectors $\boldsymbol{z}^A$ and $\boldsymbol{z}^B$. Instead we set the dustbin vectors values to a single trained parameters as in [6]. Table 1 shows the matching performances of 3D-SMNet compared to different baselines.

Zero-shot experiments on photorealistic environments. Next, we test our method on the photorealistic Replica dataset [9]. We select the 6 FRL apartment scenes and create 15 episodes by combining pairwise layouts. The results are shown in Tab. 1. On this zero-shot experiment we observe that our method performs the best in terms of matching accuracy with +3.3% increase compared to other baselines (see line 5 *vs*. 1-4). However, 3D-SMNet is outperformed on all other metrics by H-L2 and H-M. We explain this result because, first the Replica scenes do not have many objects added or removed in the scene and therefore the use of dustbins scores becomes obsolete, and second because the three matchers H-1L, S and 3D-SMNet are trained on Matterport scenes and objects.

Using simulated episodes as data augmentation. We conduct experiments using the RIO dataset [11]. RIO has 1335 point clouds of houses that we select to create 903 episodes split into train and val. We create ground-truth object pairs for each episodes using the instance level annotations of the dataset keeping a subset of categories: *chair, bed, couch, TV, plant and toilet*. We train 3D-SMNet on both the RIO and simulated episodes and compare the performances when the network is trained on RIO or simulated episodes separately. We find from Tab. 2 that using simulated episodes as additional data helps with an increase in performances of $+10\%$ on matching accuracy and $rank@1$ and $+5\%$ on mAP.

Detection and re-Identification measured jointly. We also measure the overall task performance (detection and re-ID) with accuracy, precision and recall metrics in Table 3. We find that 3D-SMNet outperforms all baselines on both the Matteport and Replica datasets (rows 5 compared to 1-4). We also report experiments with ground-truth detections (GTbox) and and an oracle matcher (GTmatch). We notice a gap in performance comparing 3D-SMNet with GTmatch and GTbox (lines 5 and 6).

	MP3D				Replica			
	rank@1	rank@5	mAP	Acc	rank@1	rank@5	mAP	Acc
H-L2	42.32 ± 0.12	70.94 ± 0.13	55.51 ± 0.11	28.82 ± 0.09	41.59 ± 0.23	**100.00 ± 0.00**	**62.54 ± 0.15**	24.91 ± 0.11
H-M	38.18 ± 0.13	58.74 ± 0.13	48.24 ± 0.12	31.44 ± 0.10	**41.83 ± 0.20**	95.84 ± 0.08	61.21 ± 0.13	21.42 ± 0.12
H-1L	62.18 ± 0.10	90.53 ± 0.06	74.45 ± 0.08	41.70 ± 0.07	33.28 ± 0.22	95.77 ± 0.08	55.86 ± 0.16	25.02 ± 0.09
Sinkhorn	68.83 ± 0.07	94.42 ± 0.04	79.75 ± 0.06	58.33 ± 0.05	21.12 ± 0.14	87.54 ± 0.12	47.43 ± 0.12	30.51 ± 0.10
3D-SMNet	**72.85 ± 0.08**	**94.84 ± 0.04**	**82.36 ± 0.06**	**64.35 ± 0.06**	29.29 ± 0.14	95.82 ± 0.08	53.63 ± 0.11	**33.88 ± 0.10**
GTbox	87.74 ± 0.06	98.83 ± 0.01	92.49 ± 0.04	81.30 ± 0.05	65.66 ± 0.06	97.23 ± 0.02	79.92 ± 0.04	52.63 ± 0.06

Table 1. 3D-SMNet test-set matching results on the Matterport scenes [3] with YCB [2] and Google-scanned [5] assets and on the zero-shot experiments on Replica scenes [9]. The GTbox experiment reports numbers working with ground-truth detections setting up an upper bound for our study.

training dataset	rank@1	rank@5	mAP	Acc
MP3D	49.46 ± 0.50	**100.00 ± 0.00**	70.00 ± 0.29	51.49 ± 0.31
RIO	51.68 ± 0.49	**100.00 ± 0.00**	71.36 ± 0.28	51.40 ± 0.29
RIO + MP3D	**62.76 ± 0.52**	97.38 ± 0.14	**76.53 ± 0.32**	**61.13 ± 0.33**

Table 2. 3D-SMNet matching results on the validation set of RIO [11] when trained with different datasets. 3D-SMNet performs best when trained jointly on real (RIO) and simulated (MP3D) episodes.

	MP3D			Replica			RIO		
	Acc	Precision	Recall	Acc	Precision	Recall	Acc	Precision	Recall
H-L2	10.29 ±0.04	21.66 ±0.07	16.39 ±0.05	2.63 ±0.01	6.88 ±0.04	4.09 ±0.02	2.82 ±0.02	6.31 ±0.06	4.86 ±0.04
H-M	10.84 ±0.03	22.70 ±0.06	17.18 ±0.05	2.70 ±0.02	7.08 ±0.05	4.18 ±0.03	2.94 ±0.02	6.50 ±0.03	4.95 ±0.03
H-1L	15.19 ±0.03	30.62 ±0.06	23.16 ±0.04	2.68 ±0.01	7.00 ±0.04	4.15 ±0.02	2.89 ±0.02	6.41 ±0.05	5.00 ±0.04
Sinkhorn	21.08 ±0.03	39.48 ±0.05	31.15 ±0.04	2.82 ±0.01	7.30 ±0.03	4.40 ±0.02	3.07 ±0.03	6.78 ±0.06	5.34 ±0.04
3D-SMNet	**24.59 ±0.04**	**44.86 ±0.06**	**35.23 ±0.05**	**3.31 ±0.01**	**8.55 ±0.04**	**5.12 ±0.02**	**3.64 ±0.03**	**8.01 ±0.06**	**6.26 ±0.05**
GTmatch	40.38 ±0.05	65.09 ±0.06	51.54 ±0.06	10.32 ±0.03	24.41 ±0.08	15.16±0.04	5.77 ±0.04	12.40 ±0.08	9.76 ±0.07
GTbox	69.29 ±0.07	81.86 ±0.05	81.86 ±0.05	29.58 ±0.05	45.65 ±0.06	45.65±0.06	35.77 ±0.18	52.64 ±0.19	52.64 ±0.19

Table 3. 3D-SMNet detection and re-ID performances on Matterport [3], Replica [9] and RIO [11] scenes. 3D-SMNet (line 5) outperforms the baselines (lines 1-4) on all metrics. The GTbox and GTmatch rows report numbers working with ground-truth detections and an oracle matcher, setting up an upper bound for our experiment.

References

[1] Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *BMVC*, 2016. 2

[2] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015. 1, 2

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 2

[4] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019. 2

[5] Google Research. Google-scanned-objects, 2020. 1, 2

[6] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[7] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *ICCV*, 2019. 1

[8] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967. 2

[9] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1, 2

[10] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *NeurIPS*, 2021. 1

[11] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 1, 2

[12] Xiaogang Wang, Gianfranco Doretto, Thomas Sebastian, Jens Rittscher, and Peter Tu. Shape and appearance context modeling. In *2007 ieee 11th international conference on computer vision*, pages 1–8. Ieee, 2007. 2

[13] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 2