Supplementary Material
ViewAL: Active Learning With Viewpoint Entropy for Semantic Segmentation

Yawar Siddiqui[1] Julien Valentin[2] Matthias Nießner[1]

[1]Technical University of Munich [2]Google

A. Dataset Statistics

We evaluate our approach on three public datasets, SceneNet-RGBD [8, 7], ScanNet [3], and Matterport3D [1]. SceneNet-RGBD is a synthetic dataset containing large-scale photorealistic renderings of indoor scene trajectories, with around 5M RGBD frames. ScanNet contains around 2.5M views in 1513 real indoor scenes. Matterport3D has around 200K RGBD views for 90 real building-scale scenes. We use a subset of images, as highlighted in Table 1, since active learning iterations on the entire datasets would be too expensive in terms of compute.

Statistic	SceneNet	ScanNet	Matterport3D
Train Sequences	2434	1041	968
Train Frames	72990	23750	25761
Validation Seqs.	500	465	214
Validation Frames	15000	5453	13702
Test Sequences	-	80	370
Test Frames	-	5320	22588
Semantic Classes	13	40	40

Table 1: Statistics of SceneNet-RGBD[7], ScanNet[3] and Matterport3D[1] dataset subsets used in our experiments.

B. Baseline Active Learning Methods

We compare our method against popular uncertainty and diversity based active learning approaches found in the literature. Here, we give a brief overview of these approaches. In terms of notation, D_U is the unlabeled dataset, D_L is the currently labeled dataset, M is the total number of target classes, K is the number of images from D_U requested to be labeled in each active selection iteration, n goes over pixels for image i and θ_{SEG} are the parameters of the segmentation network.

Random Selection (RAND) In random selection, in each active selection iteration, the next query for K samples is composed of randomly selected samples from the unlabeled dataset.

Softmax Confidence (CONF) The least confidence approach discussed in [13] can be adapted to deep convolutional networks by using softmax probability of the most probable class as confidence [14]. This selection strategy then selects the least K confident samples from D_U as the next query. For semantic segmentation, we calculate confidence for each pixel and use the sum across pixels as the confidence for the image. For each image i, the confidence score is therefore given by Eq. 1, and K least scoring samples are selected for label acquisition.

$$S_i^{CONF} = \sum_n \max_j p(y_i^n = j \mid x_i; \theta_{SEG}) \tag{1}$$

Softmax Margin (MAR) Similar to CONF, this approach [9] ranks all the samples in order of the difference of softmax probabilities of the most probable label (j_1) and the second most probable label (j_2), and chooses the K samples which have the least difference (Eq. 2) [14]. The idea is that samples for which the network has a small margin between the top predictions means that the network is very uncertain between the two.

$$S_i^{MAR} = \sum_n (p(y_i^n = j_1 | x_i; \theta_{SEG}) - p(y_i^n = j_2 | x_i; \theta_{SEG})) \tag{2}$$

Softmax Entropy (ENT) In the case of semantic segmentation, the entropy value for each pixel in the image is summed to get the entropy score for the whole image (Eq. 3).

$$S_i^{ENT} = -\sum_n \sum_{j=1}^{M} p(y_i^n \mid x_i; \theta_{SEG}) \log p(y_i^n \mid x_i; \theta_{SEG}) \tag{3}$$

Entropy takes into account probabilities of all classes unlike CONF, which considers most probable class or MAR, which only considers the top two most probable classes.

CEAL Entropy (CEAL) CEAL [14] combines CONF, MAR, ENT methods with pseudo-labeling in their active learning framework. We only compare with their ENT variant since the results are quite identical for all the other measures. At the end of each active selection iteration, they propose not only adding samples labeled by the oracle, but also high confidence samples from D_U for which softmax entropy is less than the threshold δ. For these samples, the assigned labels are the predicted ones by the current model. The idea behind pseudo-labeling is that since the high confidence samples are close to the labeled samples in CNNs feature space, adding them in training is a reasonable data augmentation for CNN to learn robust features. Further, as the active iteration increase, the number of samples selected for pseudo-labeling increases since the network gets more and more confident. To prevent high amounts of pseudo-labeling, the threshold is decreased at the end of each selection iteration. Our implementation of CEAL only assigns pseudo-labels at pixel level instead of image level to account for locality of segmentation task.

Monte Carlo Dropout (MCDR) It has been argued in [4] that vanilla deep learning models rarely represent model uncertainty, and softmax entropy is not really a good measure of uncertainty. Instead of softmax probabilities [5] use Monte Carlo (MC) dropout to estimate model uncertainty.

$$S_i^{MCDR} = -\sum_n \sum_{j=1}^{M} p_{MC}(y_i^n \mid x_i; \theta_{SEG}) \log p_{MC}(y_i^n \mid x_i; \theta_{SEG}) \tag{4}$$

where p_{MC} is given by

$$p_{MC}(y_i^n \mid x_i; \theta_{SEG}) = \frac{1}{D} \sum_{d=1}^{D} p_{SM}(y_i^n \mid x_i; \theta_{SEG}^d), \tag{5}$$

with D being the total number of MC Dropout runs.

Regional MC Dropout (RMCDR) Proposed for semantic segmentation in [6], it follows the same approach as MCDR. However, instead of calculating scores for whole images, scores are calculated for fixed-size regions. The selection algorithm is then selecting as many highest entropy regions as it takes to make up K images. The original method of [6] uses Vote Entropy [2], however we use MC Dropout since it gives slightly better results. Further, the method of [6] uses cost estimates regressed from annotator click patterns, which we dont use since these are not available for any of the datasets we evaluate on.

Maximum Representativeness (MREP) Unlike the other approaches discussed until now which were only uncertainty based, MREP is a mixed approach that combines uncertainty and diversity. This method, proposed in [15] first choose points that are highly uncertain. From among these points, it further chooses points that best represent the rest of distribution based on some similarity measure. In our implementation, vote entropy is first used to select $2K$ samples, and then K most representative samples amongst those are selected to be labeled. We use the Euclidean norm for the similarity measure.

Core-Set Selection (CSET) Core-Set [12] is a purely diversity-based approach. The method aims to select a subset of K points such that the model trained on a subset of the points is competitive for the rest of the points. The K samples selected are the ones that have the smallest δ for the δ cover of the set. This means that the algorithm seeks to minimize the maximum distance between sample x_i in the remaining unlabeled dataset and its closest neighbor x_j in the selected subset. We use the simple greedy selection strategy proposed in [12] as it performs only slightly worse than the robust version.

C. Performance with Imperfect Depth and Pose

Here we evaluate the performance of our method when the ground truth depth and pose are not available, i.e. only RGB frames are available. In such a case, one alternative for making associations between pixels across frames is to use structure from motion/multi-view stereo methods. We use COLMAP [10, 11] to first reconstruct the scenes from RGB frames and obtain depth and camera parameters. We use 5 scenes from ScanNet [3] for this. We keep to just 5 scenes as the time taken to reconstruct a scene using COLMAP is quite long, and since here we only want to compare the performance using ground truth depth and pose against reconstructions, these should be sufficient. We use 1000 frames from each scene, and split the total 5000 frames into 2000 training (unlabeled), 1000 validation and 2000 test frames. The seed set has 100 fully labeled frames. Each selection iteration chooses 100 more frames (or equivalent superpixels) from the training set to be labeled. We compare against random selection (**RAND**) and the variant of our method that uses true pose and depth (ViewAL(TRUE)).

Fig. 1 shows the results for this experiment. We observe that our method which uses reconstructed depth and pose still outperforms the **RAND** baseline and performs only slightly worse than the variant using true depth and poses.

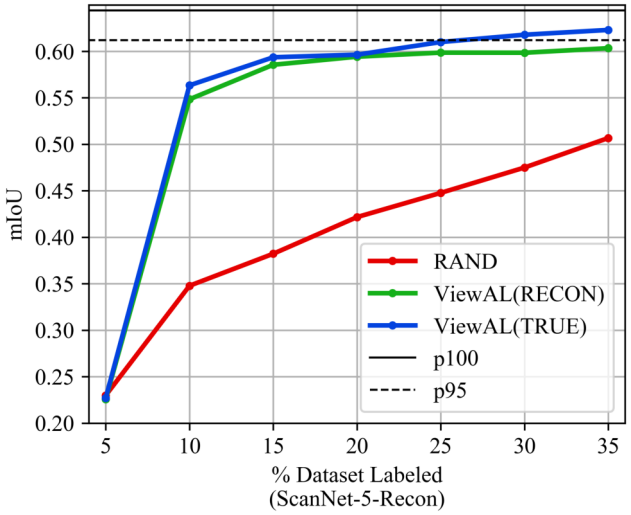

Figure 1: Performance with imperfect depth and pose. Our method using reconstructed depth and pose, ViewAL(RECON), outperforms the RAND baseline and performs only slightly worse than the variant using true depth and poses, ViewAL(TRUE).

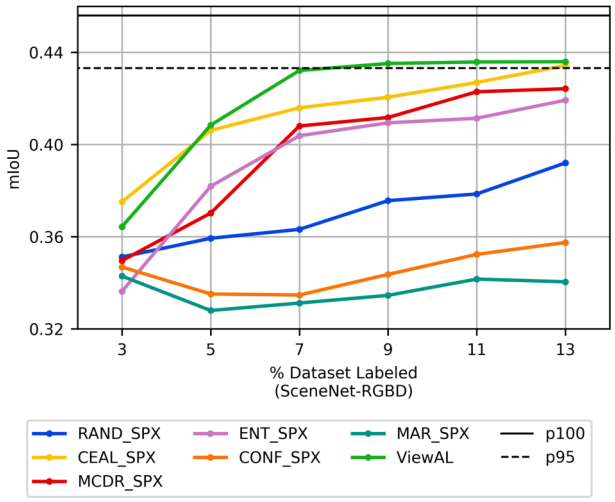

Figure 2: Active learning performance for our method and other baselines when all baselines use superpixels.

D. Comparison with baselines allowed to select superpixels

Fig. 2 shows the scenario where other methods are allowed to use superpixel selection instead of window / image selection. It can be observed that most methods do benefit from superpixel selection.

E. Handling non-static data

For computation of view entropy and divergence scores, we need to associate superpixels between frames. In our experiments, we use frame depth and pose to get these associations. However, this can be done only in case of static scenes, i.e. when objects do not change positions across frames. A promising future direction could be to extend this work for the dynamic setting using, for instance, optical flow estimates or keypoint descriptor matching to achieve superpixel association across frames.

F. Result Tables

Due to limited space in the main paper, we present the experimental results here in tabular form. Table 2, Table 3, Table 4 list results for all the methods we compared on SceneNet-RGBD [7], ScanNet [3] and Matterport3D [1] datasets. Table 5 reports results for the ablation study.

References

[1] 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, 3, 5

[2] Ido Dagan and Sean P Engelson. Committee-based sampling for training probabilistic classifiers. In *Machine Learning Proceedings 1995*, pages 150–157. Elsevier, 1995. 2

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1, 2, 3, 4

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016. 2

[5] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1183–1192. JMLR. org, 2017. 2

[6] Radek Mackowiak, Philip Lenz, Omair Ghori, Ferran Diego, Oliver Lange, and Carsten Rother. Cereals-cost-effective region-based active learning for semantic segmentation. *arXiv preprint arXiv:1810.09726*, 2018. 2

[7] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J.Davison. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. 2016. 1, 3, 4

[8] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J.Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? 2017. 1

% Labeled Data	RAND	RMCDR	MCDR	ENT	CONF	CSET	MAR	MREP	CEAL	ViewAL (Images)	ViewAL
1	0.2245	0.2124	0.2160	0.2261	0.2257	0.2259	0.2254	0.2168	0.2255	0.2159	0.2125
3	0.2612	0.3524	0.2427	0.2586	0.2584	0.2509	0.2558	0.2650	0.2624	0.2585	**0.3643**
5	0.2791	0.3776	0.2768	0.2864	0.2868	0.2767	0.2882	0.2980	0.3101	0.2854	**0.4084**
7	0.2991	0.4026	0.3038	0.3082	0.3029	0.3001	0.3038	0.3165	0.3376	0.3094	**0.4321**
9	0.3173	0.4092	0.3278	0.3292	0.3208	0.3234	0.3194	0.3345	0.3542	0.3385	**0.4352**
11	0.3290	0.4187	0.3409	0.3395	0.3334	0.3346	0.3313	0.3451	0.3580	0.3541	**0.4358**
13	0.3405	0.4226	0.3583	0.3541	0.3510	0.3467	0.3459	0.3644	0.3639	0.3649	**0.4359**
15	0.3509	0.4337	0.3716	0.3616	0.3630	0.3285	0.3522	0.3755	0.3781	-	**0.4383**
17	0.3587	0.4340	0.3737	0.3726	0.3731	0.3432	0.3688	0.3845	0.3807	-	**0.4412**

Table 2: Semantic segmentation performance in terms of mIoU when labeled data is selected using baseline active learning methods and our method on SceneNet-RGBD [7] dataset.

% Labeled Data	RAND	RMCDR	MCDR	ENT	CONF	CSET	MAR	MREP	CEAL	ViewAL (Images)	ViewAL
1	0.0998	0.0957	0.0950	0.0961	0.0958	0.0961	0.0999	0.0934	0.1001	0.0957	0.0953
6	0.1746	0.2158	0.1821	0.1686	0.1672	0.1741	0.1662	0.1843	0.1598	0.1895	**0.2365**
12	0.1976	0.2525	0.2083	0.1989	0.1972	0.2077	0.2003	0.2128	0.2035	0.2214	**0.2663**
17	0.2128	0.2619	0.2327	0.2167	0.2146	0.2286	0.2167	0.2349	0.2284	0.2353	**0.2757**
22	0.2298	0.2719	0.2480	0.2350	0.2321	0.2378	0.2291	0.2483	0.2437	0.2490	**0.2808**
27	0.2333	0.2739	0.2558	0.2423	0.2407	0.2444	0.2355	0.2523	0.2524	0.2580	**0.2823**
33	0.2390	0.2812	0.2654	0.2517	0.2469	0.2531	0.2470	0.2581	0.2619	0.2648	**0.2874**

Table 3: Semantic segmentation performance in terms of mIoU when labeled data is selected using baseline active learning methods and our method on ScanNet [3] dataset.

[9] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001. 1

[10] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518. Springer, 2016. 2

[11] Steven M Seitz, Brian Curless, James Diebel, Daniel Scharstein, and Richard Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 519–528. IEEE, 2006. 2

[12] Ozan Sener and Silvio Savarese. Active Learning for Convolutional Neural Networks: A Core-Set Approach. 08 2017. 2

[13] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009. 1

[14] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016. 1, 2

[15] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 399–407. Springer, 2017. 2

% Labeled Data	RAND	RMCDR	MCDR	ENT	CONF	CSET	MAR	MREP	ViewAL (Images)	ViewAL
1	0.0754	0.0797	0.0825	0.0765	0.0762	0.0778	0.0781	0.0807	0.0815	0.0802
5	0.1086	0.1589	0.1250	0.1141	0.1207	0.1053	0.1159	0.1254	0.1157	**0.1693**
9	0.1310	0.1831	0.1443	0.1424	0.1387	0.1254	0.1343	0.1512	0.1496	**0.1920**
13	0.1429	0.1905	0.1659	0.1590	0.1544	0.1481	0.1478	0.1644	0.1708	**0.2005**
17	0.1564	0.1991	0.1735	0.1692	0.1616	0.1609	0.1614	0.1749	0.1750	**0.2026**
20	0.1609	0.1994	0.1802	0.1787	0.1703	0.1680	0.1673	0.1845	0.1813	**0.2092**
24	0.1660	0.2007	0.1903	0.1836	0.1796	0.1826	0.1769	0.1945	0.1925	**0.2140**
27	0.1766	0.2042	0.1947	0.1826	0.1839	0.1850	0.1777	0.1971	-	**0.2148**
31	0.1823	0.2112	0.2032	0.1960	0.1915	0.1902	0.1869	0.2019	-	**0.2159**

Table 4: Semantic segmentation performance in terms of mIoU when labeled data is selected using baseline active learning methods and our method on Matterport3D [1] dataset.

% Labeled Data	ViewAL(VE)	ViewAL(VE+Spx)	ViewAL(VE+Spx+MCDR)	ViewAL(VE+Spx+MCDR+VD)
1	0.1004	0.1001	0.0952	0.0952
6	0.1795	0.2280	0.2345	**0.2365**
12	0.2033	0.2502	0.2587	**0.2663**
17	0.2247	0.2590	0.2708	**0.2757**
22	0.2380	0.2637	0.2754	**0.2807**
27	0.2445	0.2675	0.2801	**0.2822**
33	0.2556	0.2680	0.2804	**0.2873**

Table 5: Ablation Study Results. ViewAL(VE) is our method without superpixels, MC dropout, and view divergence. When superpixels are used for selection over entire images, we see significant improvements as shown by the curve ViewAL(VE+Spx). Adding MC dropout improves performance further as indicated by ViewAL(VE+Spx+MCDR). Our final method, ViewAL(VE+Spx+MCDR+VD) improves over this further by adding view divergence.