Real-time Progressive 3D Semantic Segmentation for Indoor Scenes

Supplementary Materials

1. Temporal accuracy

In our experiments, the main source of fluctuation is usually come from new section of the scene is being scanned. We consider this as a trade-off between stability in accuracy and the capability to progressively fix segmentation errors. In the case if we only adhere to the temporal consistency, the system might stuck in the same segmentation over and over, even if the segmentation is deemed incorrect. In our system, we actually represented this trade-off by a weighted sum in the CRF unary term of time t. Here is a plot that shows accuracy over time when we change the weighting parameter τ for a typical scene. In our experiment, we set $\tau = 0.5$ since it gives a good trade-off between temporal consistency and segmentation accuracy.

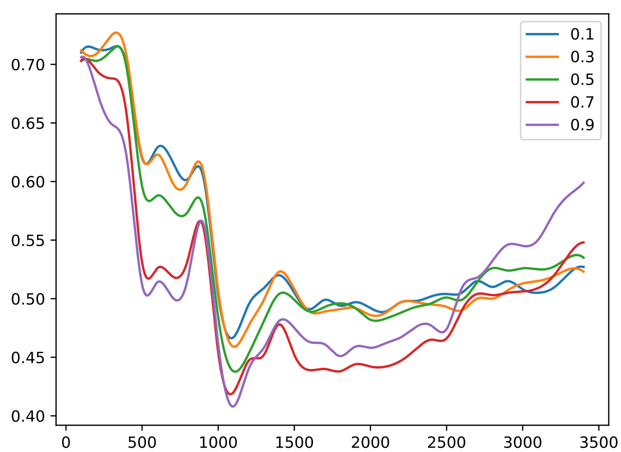


Figure 1: Accuracy over time in SceneNN/061 when changing the temporal consistency weighting parameter τ. When τ is small, the system is subject to less fluctuation, but lose the ability to fix initial segmentation errors.

2. Additional online results

Here we present the full results of our online semantic segmentation experiments. The results consists of 80 scenes from SceneNN (Table 1) and unique scenes from ScanNet (Table 2). Figure 2 shows the accuracy differences between our method and SemanticFusion on those 400 scenes. On average, we gain 4.4% in accuracy and 7.2% in IoU. Specifically, out of a total of 480 scenes, we outperform SemanticFusion in 376 scenes, in which 176 for over 5%, and 62 for over 10% in accuracy. On the other hand, SemanticFusion only outperforms our method in 67 scenes, 21 for over 5%, and 6 for over 10% in accuracy.

We adopt two common metrics from 2D semantic segmentation for our 3D evaluation, namely vertex accuracy (A) and frequency weighted intersection over union (*wIoU*). Let v_{ij} be the number of vertices of class i predicted to be class j, and $t_i = \sum_j v_{ij}$ be the total number vertices of class i. The two metrics can be computed as follows:

$$A = \frac{\sum_i v_{ii}}{\sum_i t_i} \tag{1}$$

$$wIoU = \frac{1}{\sum_k t_k} \sum_i \frac{t_i v_{ii}}{t_i + \sum_j v_{ji} - v_{ii}} \tag{2}$$

We also show additional overtime accuracy evaluation for some selected scenes from SceneNN.

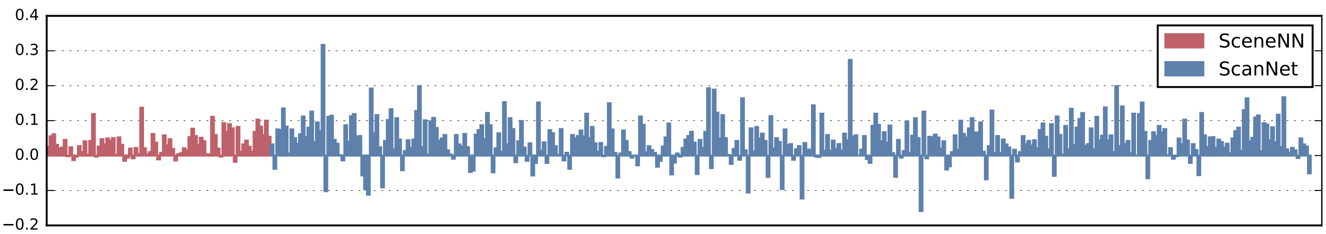


Figure 2: Differences in accuracy between our method and SemanticFusion on SceneNN (80 scenes) and ScanNet (400 unique scenes) dataset. The horizontal axis is scene ID sorted in increasing order.

ID	**Direct**		**SF**		**Ours**		
	A	wIoU	A	wIoU	A	wIoU	mAP@50%
005	0.486	0.386	0.492	0.388	0.539	0.467	0.077
011	0.770	0.654	0.776	0.664	0.800	0.699	0.522
014	0.490	0.391	0.508	0.414	0.539	0.501	0.118
015	0.535	0.366	0.571	0.408	0.620	0.482	0.152
016	0.607	0.476	0.625	0.497	0.680	0.566	0.342
021	0.559	0.427	0.634	0.510	0.676	0.581	0.197
025	0.643	0.519	0.648	0.533	0.698	0.610	0.213
030	0.584	0.450	0.597	0.466	0.658	0.559	0.568
032	0.795	0.682	0.811	0.710	0.813	0.703	0.311
036	0.575	0.431	0.603	0.476	0.655	0.556	0.162
038	0.608	0.499	0.628	0.527	0.659	0.590	0.394
041	0.554	0.441	0.579	0.481	0.564	0.482	0.246
045	0.649	0.515	0.673	0.552	0.667	0.549	0.443
047	0.537	0.441	0.550	0.455	0.570	0.497	0.156
052	0.578	0.448	0.530	0.423	0.552	0.433	0.437
054	0.489	0.335	0.597	0.470	0.602	0.489	0.379
057	0.742	0.668	0.517	0.366	0.654	0.539	0.246
061	0.751	0.606	0.777	0.641	0.809	0.691	0.591
062	0.497	0.353	0.757	0.693	0.778	0.727	0.377
065	0.563	0.412	0.524	0.374	0.528	0.393	0.237
066	0.564	0.479	0.561	0.420	0.571	0.442	0.336
069	0.578	0.467	0.584	0.510	0.646	0.557	0.490
071	0.557	0.435	0.588	0.483	0.626	0.514	0.219
073	0.501	0.397	0.568	0.452	0.558	0.442	0.596
074	0.540	0.387	0.507	0.404	0.515	0.430	0.391
078	0.497	0.365	0.515	0.388	0.535	0.454	0.349
080	0.634	0.503	0.570	0.431	0.628	0.510	0.393
082	0.523	0.364	0.658	0.543	0.689	0.589	0.242
084	0.737	0.636	0.544	0.390	0.591	0.455	0.401
086	0.622	0.519	0.646	0.551	0.668	0.587	0.350
087	0.192	0.151	0.731	0.617	0.751	0.658	0.434
089	0.558	0.449	0.573	0.466	0.618	0.529	0.236
092	0.549	0.424	0.207	0.168	0.193	0.158	0.501
093	0.507	0.376	0.553	0.432	0.558	0.469	0.383
096	0.659	0.536	0.668	0.549	0.666	0.565	0.265
098	0.601	0.497	0.600	0.497	0.623	0.557	0.321
109	0.553	0.461	0.505	0.362	0.512	0.383	0.267
201	0.763	0.619	0.565	0.478	0.587	0.514	0.194
202	0.520	0.403	0.779	0.640	0.794	0.678	0.250

ID	Direct		SF		Ours		
	A	wIoU	A	wIoU	A	wIoU	mAP@50%
205	0.635	0.573	0.642	0.585	0.629	0.586	0.501
206	0.766	0.654	0.778	0.673	0.775	0.675	0.417
207	0.559	0.432	0.568	0.446	0.596	0.510	0.171
209	0.415	0.281	0.533	0.423	0.587	0.494	0.280
213	0.539	0.425	0.551	0.440	0.561	0.448	0.478
217	0.601	0.492	0.433	0.299	0.510	0.386	0.135
223	0.669	0.559	0.689	0.587	0.729	0.655	0.409
225	0.617	0.485	0.620	0.518	0.676	0.611	0.184
227	0.628	0.541	0.627	0.501	0.658	0.529	0.462
231	0.655	0.542	0.656	0.543	0.660	0.577	0.692
234	0.648	0.532	0.631	0.551	0.682	0.617	0.342
237	0.702	0.587	0.658	0.545	0.700	0.615	0.200
240	0.562	0.427	0.722	0.617	0.726	0.649	0.214
243	0.538	0.455	0.553	0.471	0.595	0.532	0.144
246	0.530	0.373	0.577	0.448	0.580	0.496	0.006
249	0.636	0.498	0.545	0.396	0.656	0.557	0.739
251	0.542	0.436	0.659	0.531	0.718	0.639	0.215
252	0.434	0.306	0.563	0.463	0.583	0.489	0.642
255	0.423	0.316	0.439	0.334	0.558	0.473	0.486
260	0.546	0.408	0.447	0.319	0.444	0.320	0.351
263	0.456	0.338	0.573	0.438	0.666	0.571	0.275
265	0.531	0.413	0.485	0.373	0.553	0.489	0.270
270	0.555	0.397	0.523	0.395	0.613	0.515	0.598
272	0.602	0.475	0.614	0.493	0.612	0.503	0.203
273	0.491	0.338	0.574	0.421	0.652	0.528	0.547
276	0.684	0.550	0.494	0.340	0.476	0.313	0.170
279	0.684	0.550	0.699	0.577	0.781	0.706	0.258
286	0.498	0.393	0.521	0.422	0.531	0.448	0.304
294	0.441	0.302	0.430	0.280	0.463	0.323	0.600
308	0.736	0.647	0.752	0.675	0.796	0.747	0.342
322	0.657	0.599	0.662	0.606	0.687	0.642	0.593
522	0.625	0.551	0.640	0.575	0.665	0.635	0.807
527	0.821	0.774	0.793	0.722	0.804	0.741	0.531
609	0.772	0.669	0.770	0.675	0.838	0.782	0.225
613	0.795	0.651	0.783	0.633	0.886	0.792	0.175
614	0.481	0.327	0.496	0.344	0.579	0.452	0.029
621	0.802	0.731	0.814	0.747	0.872	0.845	0.305
623	0.540	0.452	0.486	0.404	0.587	0.551	0.439
700	0.609	0.470	0.607	0.463	0.661	0.535	0.581

Table 1: Comparison of semantic segmentation performance on 80 scenes in SceneNN dataset. Our proposed CRF model consistently outperforms the naive approach that directly fuses neural network predictions to 3D (Direct) [3], and Semantic-Fusion (SF) [7]. The fifth column further reports the average precision of the instance-based semantic segmentation task.

ID	SF		Ours		ID	SF		Ours	
	A	wIoU	A	wIoU		A	wIoU	A	wIoU
0000	0.495	0.384	0.527	0.443	0060	0.782	0.698	0.828	0.755
0001	0.604	0.485	0.566	0.459	0061	0.515	0.346	0.472	0.381
0003	0.573	0.505	0.649	0.589	0062	0.711	0.597	0.724	0.624
0004	0.359	0.270	0.432	0.387	0063	0.690	0.580	0.734	0.643
0005	0.767	0.623	0.903	0.849	0065	0.610	0.428	0.632	0.456
0006	0.648	0.487	0.731	0.599	0067	0.742	0.628	0.788	0.738
0007	0.568	0.415	0.574	0.457	0068	0.624	0.515	0.752	0.722
0008	0.726	0.590	0.801	0.722	0069	0.347	0.200	0.547	0.461
0009	0.795	0.667	0.845	0.750	0070	0.433	0.286	0.457	0.356
0010	0.684	0.567	0.719	0.614	0071	0.708	0.540	0.809	0.698
0012	0.551	0.386	0.612	0.511	0072	0.392	0.320	0.394	0.348
0013	0.530	0.308	0.641	0.436	0073	0.586	0.467	0.684	0.604
0015	0.780	0.634	0.832	0.730	0074	0.469	0.247	0.577	0.366
0016	0.603	0.370	0.683	0.495	0075	0.574	0.421	0.653	0.523
0017	0.576	0.396	0.702	0.555	0076	0.685	0.513	0.727	0.587
0018	0.787	0.712	0.825	0.813	0077	0.521	0.391	0.569	0.461
0019	0.666	0.517	0.760	0.652	0078	0.415	0.317	0.474	0.378
0020	0.728	0.651	0.797	0.752	0079	0.527	0.427	0.542	0.481
0021	0.567	0.366	0.885	0.820	0080	0.364	0.317	0.367	0.342
0022	0.848	0.748	0.746	0.651	0081	0.522	0.356	0.512	0.362
0023	0.822	0.687	0.933	0.879	0082	0.691	0.497	0.751	0.577
0024	0.577	0.375	0.692	0.521	0083	0.548	0.349	0.579	0.409
0026	0.576	0.486	0.622	0.573	0084	0.782	0.703	0.808	0.761
0027	0.753	0.600	0.787	0.670	0085	0.597	0.503	0.659	0.590
0030	0.670	0.511	0.671	0.544	0086	0.777	0.691	0.802	0.736
0032	0.550	0.460	0.537	0.495	0087	0.644	0.486	0.596	0.471
0033	0.838	0.721	0.925	0.861	0088	0.783	0.637	0.741	0.653
0035	0.531	0.394	0.571	0.471	0089	0.473	0.292	0.532	0.352
0036	0.554	0.414	0.668	0.599	0090	0.695	0.532	0.768	0.636
0037	0.759	0.609	0.878	0.797	0091	0.316	0.175	0.404	0.248
0038	0.813	0.685	0.868	0.794	0092	0.546	0.367	0.593	0.437
0039	0.496	0.337	0.552	0.399	0093	0.538	0.345	0.660	0.482
0041	0.747	0.574	0.690	0.596	0094	0.490	0.306	0.577	0.456
0042	0.876	0.844	0.780	0.692	0095	0.619	0.437	0.571	0.455
0043	0.709	0.592	0.597	0.480	0096	0.671	0.581	0.692	0.634
0044	0.694	0.537	0.885	0.860	0097	0.697	0.559	0.761	0.642
0045	0.837	0.761	0.900	0.875	0098	0.411	0.315	0.423	0.347
0047	0.563	0.418	0.680	0.597	0099	0.475	0.254	0.628	0.420
0048	0.727	0.616	0.750	0.678	0100	0.574	0.499	0.606	0.551
0052	0.631	0.398	0.540	0.291	0101	0.408	0.325	0.516	0.482
0053	0.770	0.638	0.812	0.709	0102	0.501	0.313	0.577	0.416
0055	0.548	0.393	0.650	0.544	0103	0.764	0.636	0.745	0.630
0056	0.666	0.480	0.799	0.669	0105	0.906	0.847	0.948	0.925
0057	0.416	0.260	0.434	0.303	0106	0.615	0.451	0.713	0.587
0058	0.477	0.305	0.584	0.446	0107	0.532	0.378	0.555	0.424

ID	SF		Ours	
	A	wIoU	A	wIoU
0108	0.307	0.225	0.291	0.232
0109	0.541	0.397	0.576	0.461
0110	0.472	0.369	0.415	0.327
0111	0.526	0.418	0.505	0.426
0112	0.798	0.652	0.950	0.910
0113	0.673	0.564	0.688	0.603
0114	0.463	0.308	0.502	0.377
0115	0.735	0.625	0.714	0.639
0116	0.133	0.045	0.205	0.087
0117	0.612	0.474	0.676	0.571
0118	0.654	0.465	0.682	0.501
0119	0.227	0.206	0.229	0.215
0120	0.639	0.504	0.715	0.604
0121	0.435	0.266	0.421	0.270
0122	0.747	0.616	0.755	0.640
0123	0.659	0.592	0.621	0.576
0124	0.636	0.505	0.695	0.599
0125	0.669	0.519	0.718	0.611
0126	0.393	0.314	0.449	0.401
0127	0.453	0.411	0.526	0.461
0128	0.370	0.200	0.425	0.282
0129	0.612	0.450	0.732	0.622
0131	0.463	0.383	0.512	0.474
0132	0.549	0.450	0.631	0.568
0133	0.723	0.566	0.758	0.619
0134	0.348	0.224	0.358	0.268
0135	0.464	0.393	0.500	0.437
0136	0.330	0.252	0.327	0.277
0137	0.479	0.356	0.504	0.402
0138	0.490	0.321	0.640	0.492
0139	0.419	0.324	0.495	0.412
0140	0.561	0.471	0.569	0.490
0141	0.631	0.524	0.568	0.514
0142	0.476	0.420	0.482	0.442
0143	0.399	0.305	0.471	0.383
0145	0.549	0.414	0.592	0.521
0146	0.659	0.488	0.668	0.492
0147	0.453	0.348	0.447	0.359
0148	0.595	0.477	0.595	0.531
0149	0.497	0.346	0.469	0.350
0150	0.574	0.369	0.686	0.505
0151	0.514	0.388	0.602	0.524
0152	0.501	0.391	0.510	0.426
0153	0.437	0.385	0.463	0.440
0154	0.492	0.276	0.508	0.291

ID	SF		Ours	
	A	wIoU	A	wIoU
0156	0.666	0.526	0.674	0.560
0157	0.424	0.336	0.392	0.302
0158	0.329	0.240	0.314	0.240
0159	0.544	0.435	0.571	0.496
0160	0.353	0.230	0.405	0.296
0161	0.680	0.498	0.772	0.620
0162	0.644	0.512	0.591	0.460
0163	0.654	0.457	0.634	0.479
0164	0.528	0.423	0.533	0.458
0165	0.672	0.607	0.668	0.638
0166	0.453	0.380	0.475	0.409
0167	0.574	0.488	0.621	0.546
0168	0.396	0.290	0.452	0.378
0169	0.529	0.347	0.598	0.420
0170	0.451	0.386	0.488	0.420
0171	0.628	0.476	0.575	0.482
0172	0.348	0.265	0.394	0.326
0173	0.644	0.496	0.666	0.536
0174	0.398	0.319	0.466	0.418
0176	0.717	0.526	0.910	0.834
0177	0.355	0.292	0.319	0.283
0178	0.440	0.335	0.629	0.593
0179	0.504	0.287	0.627	0.418
0181	0.578	0.416	0.627	0.500
0182	0.593	0.405	0.709	0.582
0183	0.385	0.226	0.436	0.286
0184	0.165	0.140	0.167	0.162
0185	0.609	0.385	0.584	0.381
0186	0.309	0.227	0.328	0.250
0187	0.716	0.597	0.759	0.682
0188	0.615	0.442	0.603	0.454
0189	0.625	0.404	0.789	0.647
0191	0.439	0.285	0.454	0.308
0192	0.378	0.234	0.273	0.271
0193	0.547	0.431	0.625	0.545
0194	0.492	0.302	0.504	0.312
0195	0.497	0.410	0.579	0.543
0196	0.709	0.593	0.758	0.658
0197	0.518	0.387	0.581	0.461
0198	0.487	0.408	0.529	0.473
0199	0.749	0.602	0.688	0.556
0200	0.529	0.340	0.642	0.469
0201	0.653	0.530	0.673	0.575
0202	0.612	0.483	0.663	0.545
0203	0.385	0.225	0.424	0.269

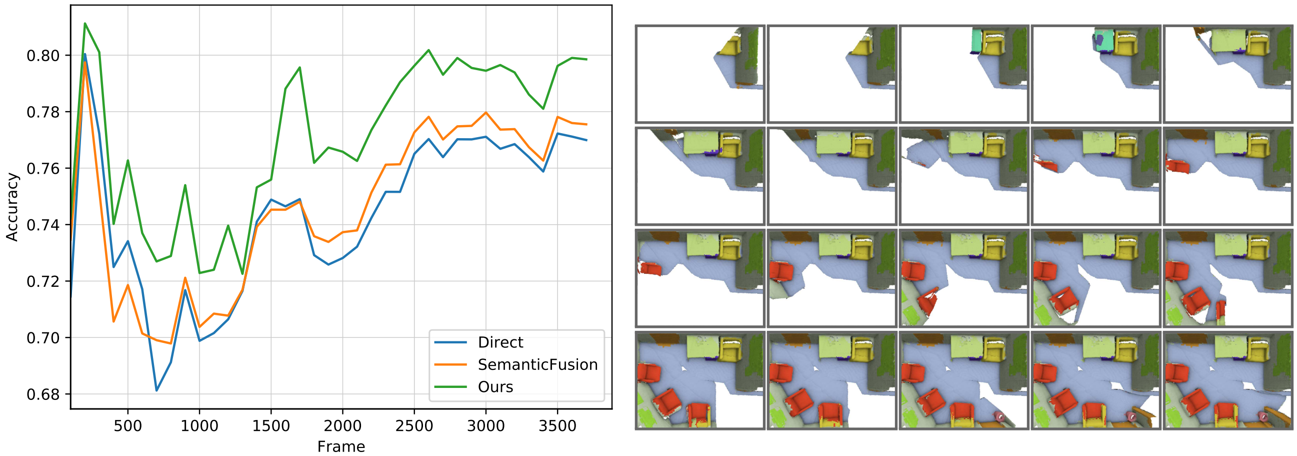
ID	SF		Ours	
	A	wIoU	A	wIoU
0204	0.785	0.656	0.689	0.561
0205	0.633	0.481	0.708	0.590
0206	0.320	0.243	0.350	0.291
0207	0.323	0.280	0.355	0.335
0209	0.477	0.395	0.465	0.426
0210	0.535	0.361	0.553	0.393
0211	0.381	0.206	0.408	0.232
0212	0.550	0.437	0.428	0.375
0213	0.488	0.339	0.524	0.403
0214	0.602	0.469	0.619	0.498
0215	0.732	0.614	0.750	0.659
0216	0.648	0.503	0.793	0.697
0217	0.494	0.444	0.491	0.442
0218	0.570	0.502	0.566	0.540
0219	0.608	0.398	0.727	0.564
0220	0.491	0.363	0.505	0.434
0221	0.541	0.389	0.600	0.489
0222	0.384	0.314	0.397	0.335
0223	0.501	0.399	0.544	0.473
0224	0.229	0.162	0.248	0.181
0225	0.516	0.434	0.549	0.488
0226	0.516	0.413	0.530	0.442
0227	0.716	0.534	0.780	0.634
0228	0.535	0.319	0.579	0.369
0229	0.442	0.311	0.715	0.564
0230	0.554	0.429	0.607	0.505
0232	0.848	0.740	0.907	0.832
0233	0.459	0.328	0.459	0.341
0234	0.532	0.344	0.549	0.368
0235	0.572	0.392	0.628	0.476
0236	0.393	0.314	0.383	0.342
0237	0.458	0.409	0.437	0.389
0238	0.443	0.368	0.527	0.494
0239	0.442	0.356	0.561	0.460
0241	0.730	0.600	0.818	0.759
0242	0.742	0.582	0.783	0.641
0244	0.556	0.481	0.623	0.578
0245	0.593	0.456	0.630	0.531
0246	0.395	0.255	0.481	0.344
0247	0.556	0.411	0.563	0.427
0248	0.366	0.246	0.305	0.250
0249	0.651	0.466	0.696	0.531
0250	0.635	0.571	0.629	0.580
0251	0.677	0.552	0.690	0.573
0252	0.642	0.417	0.740	0.554

ID	SF		Ours	
	A	wIoU	A	wIoU
0253	0.236	0.185	0.243	0.214
0254	0.542	0.389	0.599	0.494
0255	0.493	0.330	0.599	0.465
0256	0.441	0.271	0.491	0.328
0257	0.805	0.726	0.646	0.546
0258	0.548	0.492	0.675	0.623
0259	0.623	0.499	0.615	0.539
0260	0.427	0.353	0.482	0.435
0261	0.596	0.418	0.647	0.497
0262	0.428	0.297	0.488	0.402
0263	0.462	0.336	0.514	0.422
0264	0.441	0.325	0.460	0.361
0265	0.643	0.527	0.684	0.608
0266	0.642	0.546	0.602	0.507
0267	0.696	0.582	0.665	0.559
0268	0.245	0.178	0.255	0.189
0270	0.314	0.240	0.372	0.309
0271	0.474	0.380	0.490	0.438
0272	0.475	0.330	0.575	0.459
0273	0.399	0.286	0.460	0.347
0274	0.667	0.479	0.718	0.546
0275	0.513	0.360	0.591	0.459
0276	0.593	0.420	0.699	0.561
0277	0.454	0.300	0.521	0.401
0278	0.558	0.497	0.624	0.581
0279	0.526	0.421	0.621	0.540
0280	0.359	0.299	0.370	0.320
0281	0.750	0.635	0.682	0.570
0282	0.421	0.229	0.448	0.262
0283	0.724	0.562	0.853	0.752
0284	0.504	0.410	0.512	0.442
0285	0.698	0.531	0.755	0.605
0286	0.430	0.370	0.437	0.388
0287	0.502	0.396	0.548	0.483
0288	0.552	0.460	0.584	0.516
0289	0.677	0.551	0.700	0.590
0290	0.267	0.188	0.146	0.083
0291	0.764	0.623	0.781	0.654
0293	0.401	0.362	0.384	0.356
0294	0.585	0.381	0.595	0.403
0295	0.507	0.423	0.563	0.519
0296	0.550	0.421	0.583	0.475
0297	0.730	0.581	0.772	0.647
0298	0.505	0.376	0.534	0.413
0299	0.662	0.531	0.706	0.597

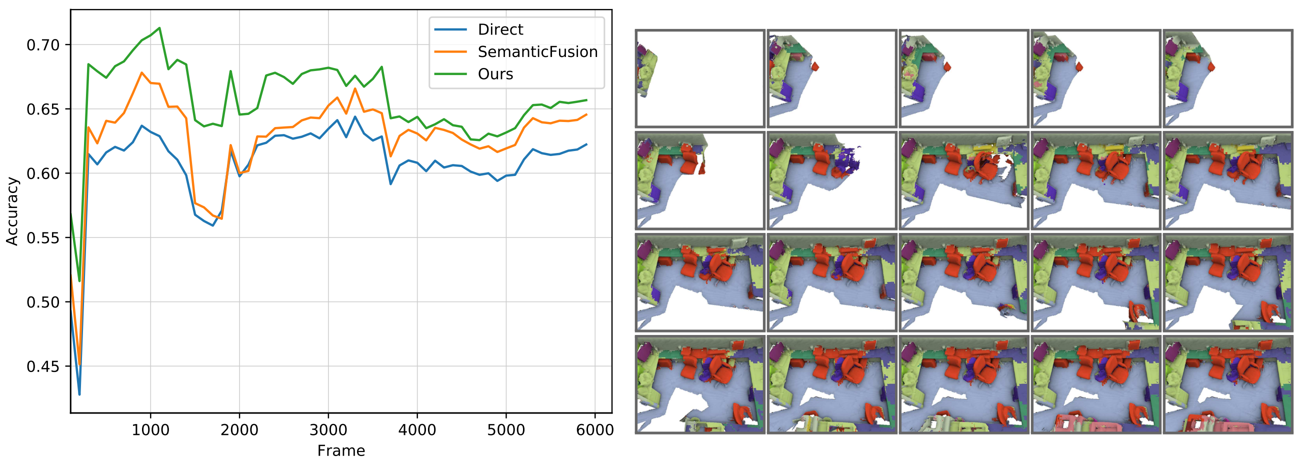
ID	SF		Ours		ID	SF		Ours	
	A	wIoU	A	wIoU		A	wIoU	A	wIoU
0300	0.829	0.723	0.850	0.755	0345	0.498	0.363	0.574	0.474
0301	0.334	0.258	0.407	0.360	0346	0.429	0.406	0.429	0.398
0302	0.569	0.380	0.661	0.483	0348	0.412	0.312	0.403	0.338
0303	0.453	0.376	0.507	0.451	0349	0.439	0.351	0.437	0.369
0304	0.683	0.572	0.701	0.624	0350	0.602	0.553	0.651	0.630
0305	0.420	0.357	0.510	0.497	0351	0.530	0.386	0.564	0.437
0306	0.372	0.276	0.315	0.277	0352	0.410	0.269	0.513	0.395
0307	0.485	0.292	0.598	0.419	0353	0.379	0.315	0.422	0.377
0308	0.258	0.190	0.317	0.274	0355	0.434	0.352	0.413	0.331
0309	0.510	0.354	0.514	0.371	0356	0.381	0.313	0.414	0.368
0310	0.428	0.251	0.513	0.350	0358	0.712	0.592	0.656	0.534
0311	0.511	0.338	0.528	0.371	0359	0.585	0.359	0.707	0.515
0312	0.504	0.291	0.638	0.451	0360	0.669	0.557	0.727	0.685
0313	0.306	0.169	0.336	0.182	0361	0.175	0.066	0.198	0.081
0314	0.521	0.400	0.601	0.573	0363	0.431	0.372	0.485	0.420
0315	0.389	0.277	0.494	0.414	0364	0.466	0.373	0.488	0.411
0316	0.599	0.413	0.721	0.593	0365	0.666	0.479	0.712	0.552
0317	0.585	0.535	0.613	0.591	0367	0.646	0.547	0.686	0.608
0318	0.136	0.089	0.168	0.118	0368	0.474	0.351	0.497	0.399
0319	0.438	0.281	0.515	0.433	0369	0.598	0.463	0.632	0.553
0320	0.435	0.363	0.454	0.410	0370	0.642	0.496	0.649	0.517
0321	0.640	0.520	0.750	0.637	0371	0.746	0.579	0.795	0.665
0322	0.719	0.572	0.763	0.638	0372	0.635	0.457	0.705	0.566
0323	0.444	0.301	0.501	0.355	0374	0.679	0.563	0.694	0.603
0324	0.700	0.540	0.837	0.754	0375	0.569	0.403	0.699	0.588
0325	0.344	0.241	0.388	0.320	0376	0.461	0.331	0.625	0.555
0326	0.550	0.413	0.608	0.542	0378	0.386	0.269	0.437	0.338
0327	0.421	0.269	0.431	0.318	0379	0.447	0.291	0.557	0.418
0328	0.189	0.089	0.388	0.287	0380	0.574	0.388	0.689	0.555
0329	0.611	0.434	0.638	0.484	0381	0.772	0.636	0.784	0.680
0330	0.352	0.141	0.492	0.263	0382	0.358	0.270	0.451	0.399
0331	0.409	0.335	0.442	0.394	0383	0.391	0.230	0.479	0.323
0332	0.564	0.445	0.605	0.491	0384	0.573	0.448	0.616	0.554
0333	0.409	0.221	0.416	0.237	0385	0.504	0.306	0.584	0.442
0334	0.586	0.432	0.706	0.679	0387	0.776	0.652	0.893	0.836
0335	0.535	0.409	0.535	0.446	0388	0.778	0.636	0.804	0.681
0336	0.601	0.459	0.720	0.638	0390	0.256	0.095	0.275	0.093
0337	0.548	0.479	0.700	0.618	0391	0.543	0.337	0.552	0.375
0338	0.608	0.474	0.677	0.559	0392	0.544	0.382	0.566	0.430
0339	0.594	0.419	0.529	0.382	0393	0.482	0.357	0.497	0.396
0340	0.443	0.339	0.484	0.403	0395	0.458	0.354	0.452	0.361
0341	0.458	0.277	0.525	0.355	0396	0.745	0.660	0.794	0.744
0342	0.744	0.578	0.799	0.653	0397	0.538	0.433	0.570	0.503
0343	0.387	0.198	0.471	0.274	0398	0.532	0.451	0.559	0.502
0344	0.686	0.534	0.751	0.638	0399	0.733	0.588	0.681	0.535

Table 2: Comparison of semantic segmentation performance on ScanNet dataset with SemanticFusion (SF) [7]. Some scenes are missing due to crashes during processing.

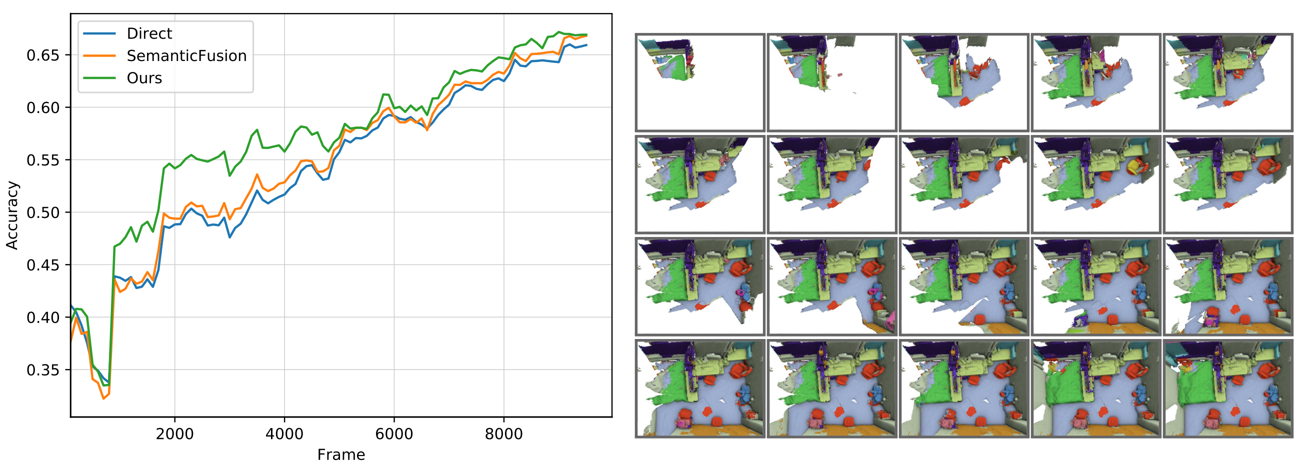
SceneNN/011



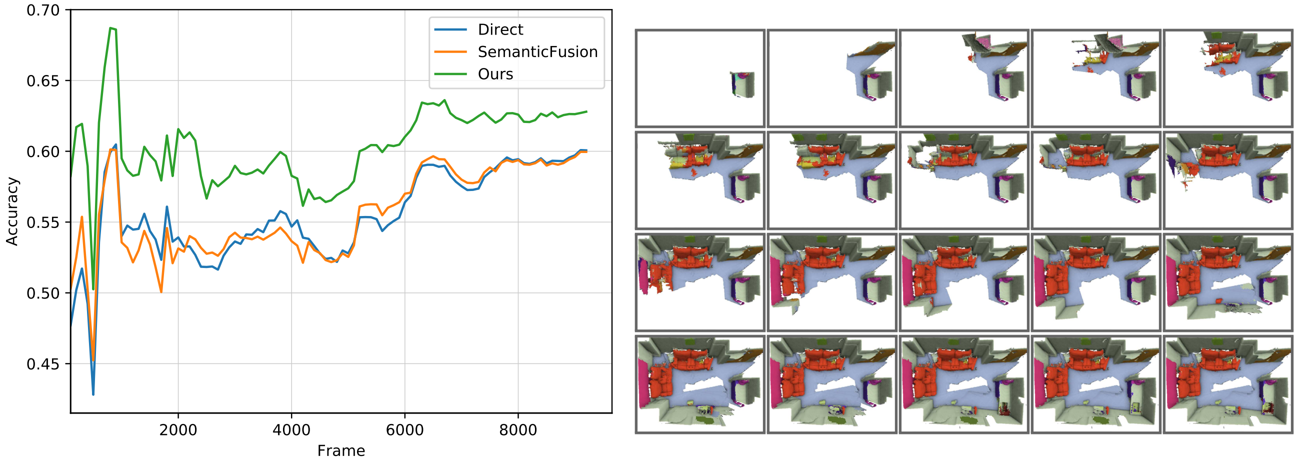
SceneNN/086



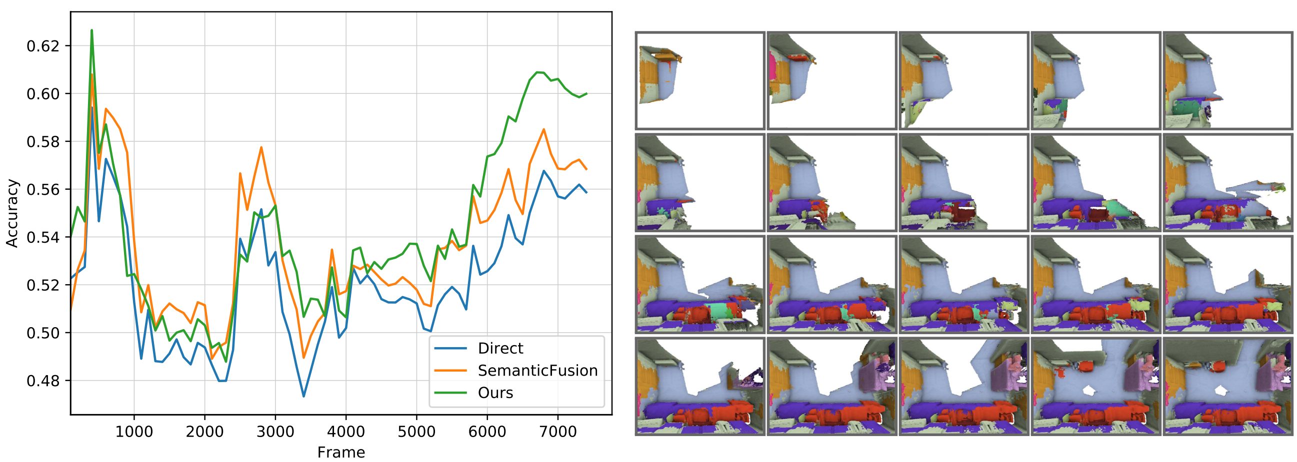
SceneNN/096



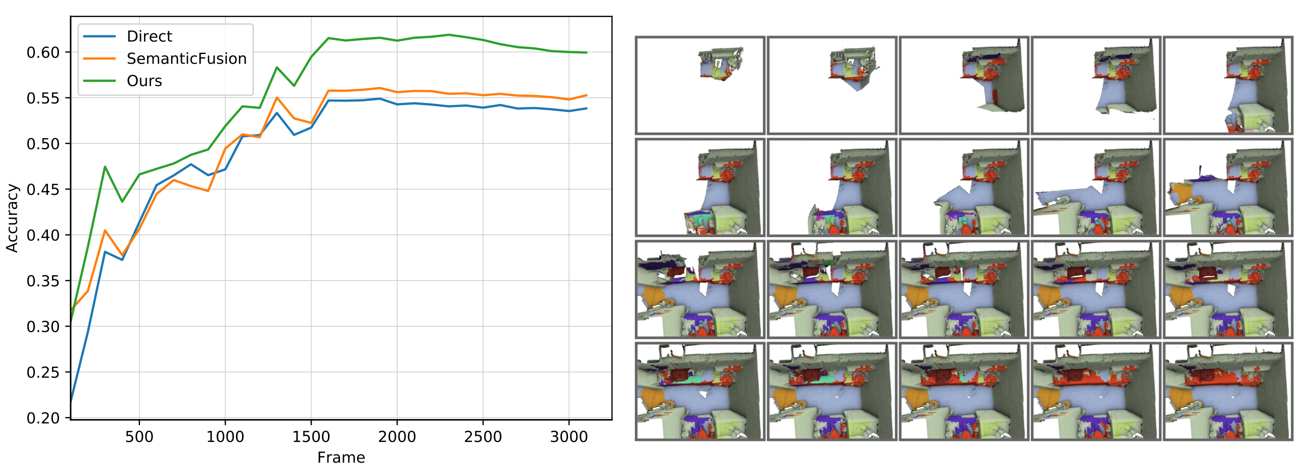
SceneNN/098



SceneNN/207



SceneNN/243



3. Additional offline results

3.1. SceneNN

We show the segmentation results on 20 SceneNN scenes. Each scene is presented with five figures consisting of ground truth segmentation, results from direct fusion using SegNet [2] and FCN-8s [6], and results from our proposed CRF using SegNet and FCN-8s. The weighted IoU scores of scenes from SceneNN are provided in Table 3.

3.2. ScanNet

We also conducted additional experiments on ScanNet [4], a new richly-annotated 3D indoor scene dataset. This dataset consists of 1513 annotated scenes of various indoor settings. However, we didn't chose this data for our main experiments because of these two issues: (1) the annotations are performed on a sub-sampled mesh; and (2) their annotations are not completed, with a lots of un-annotated regions.

Providing dense annotations in 3D is a non-trivial task. For ScanNet, they had to out-source the segmentation task to workers on Amazon Mechanical Tusk. However, quality control will be a big concern, as there might be subtle disagreements between manual segmentation results. With this experiment, we want to show that our method can speed up this process by providing an initial high-quality semantic labels automatically. Later, user intervention can help refining these labels without much effort.

We chose to show results of 10 scenes from ScanNet. We present the results with the same format as in SceneNN. The segmentation results are obtained on high-quality meshes, while the ground truth meshes are in lower quality. results are shown in Table 4.

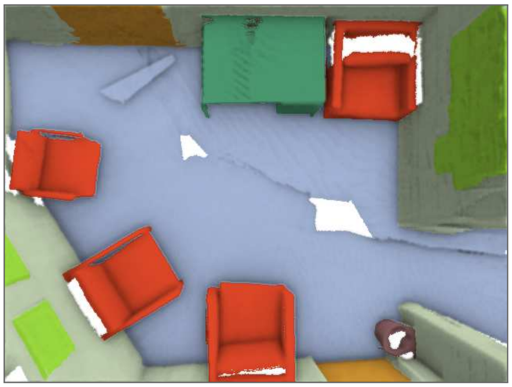
wIoU	**SegNet**		**FCN-8s**		**SSCNet**	
Scene	Base	Ours	Base	Ours	Base	Ours
011	0.622	**0.783**	0.550	0.682	0.299	0.378
016	0.425	0.636	0.463	0.519	0.514	**0.679**
030	0.413	0.570	0.421	**0.624**	0.418	0.383
061	0.351	**0.728**	0.162	0.276	0.640	0.641
078	0.405	**0.594**	0.393	0.562	0.374	0.428
086	0.459	**0.613**	0.356	0.551	0.464	0.448
096	0.495	**0.605**	0.460	0.545	0.545	0.567
206	0.502	0.718	0.463	0.729	**0.800**	0.771
223	0.524	**0.685**	0.567	0.669	0.528	0.526
255	0.410	**0.559**	0.429	0.580	0.450	0.526

Table 3: Weighted IoU offline results on SceneNN [5].

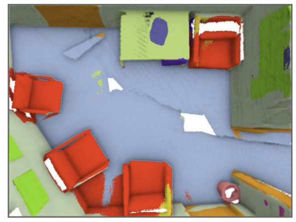
Acc.	**SegNet**			**FCN-8s**		
Scene	Base	SF	Ours	Base	SF	Ours
0000	0.442	0.469	0.480	0.512	0.553	**0.598**
0002	0.481	0.512	**0.549**	0.451	0.483	0.501
0006	0.610	0.613	0.600	0.627	**0.652**	0.644
0014	0.513	0.525	0.449	0.534	**0.549**	0.503
0028	0.554	0.607	0.664	0.550	0.617	**0.693**
0029	0.614	0.660	0.625	0.615	0.686	**0.736**
0031	0.510	0.543	0.577	0.649	0.713	**0.729**
0034	0.631	0.691	0.661	0.660	**0.698**	0.679
0050	0.548	0.585	**0.620**	0.526	0.561	0.574
0054	0.481	0.513	**0.534**	0.451	0.470	0.482

Table 4: Accuracy offline results of direct method [3], SemanticFusion (SF) [7] and ours on ScanNet [4].

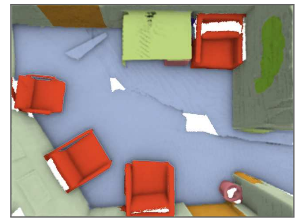
SceneNN/011



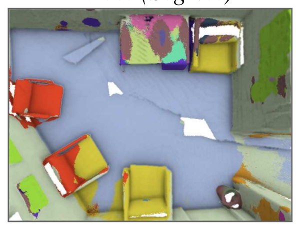
Ground truth



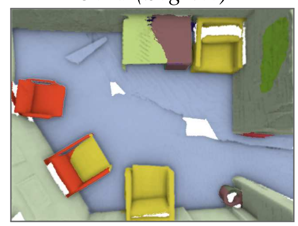
Direct (SegNet)



Ours (SegNet)



Direct (FCN)



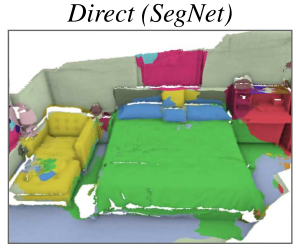
Ours (FCN)

SceneNN/016

Ground truth

Direct (SegNet)

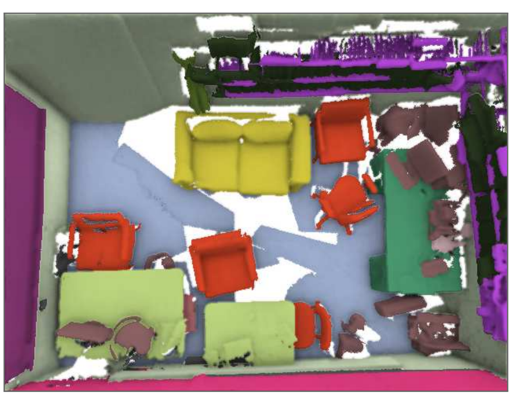
Ours (SegNet)



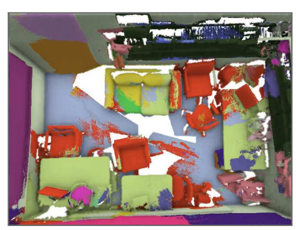
Direct (FCN)

Ours (FCN)

SceneNN/030

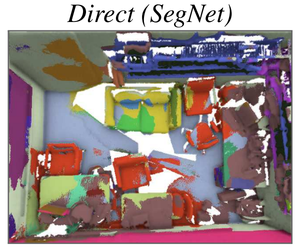


Ground truth

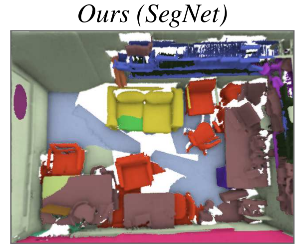


Direct (SegNet)

Ours (SegNet)

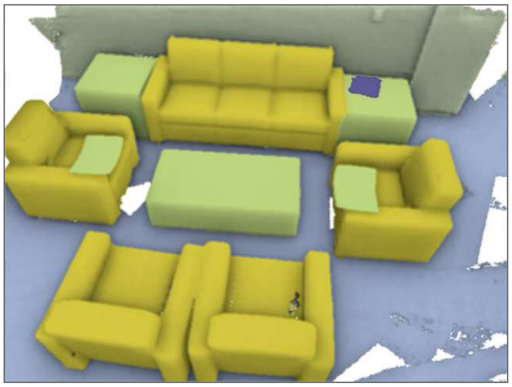


Direct (FCN)

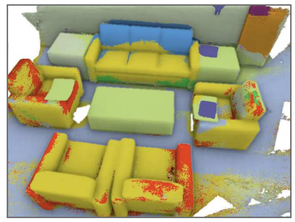


Ours (FCN)

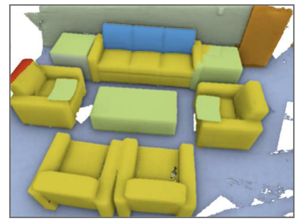
SceneNN/061



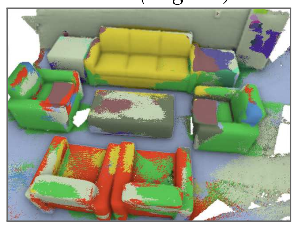
Ground truth



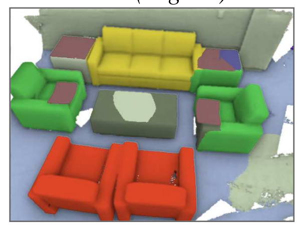
Direct (SegNet)



Ours (SegNet)

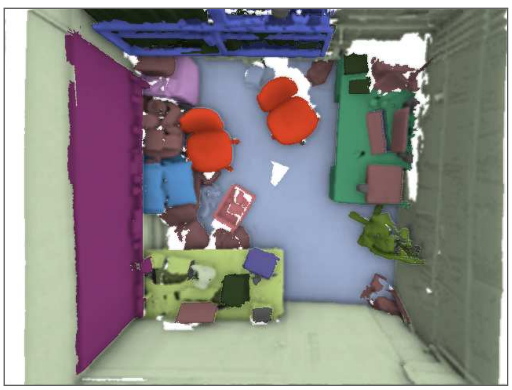


Direct (FCN)



Ours (FCN)

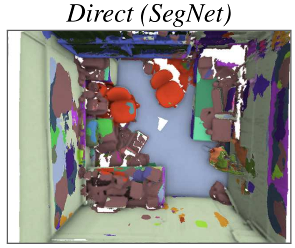
SceneNN/078



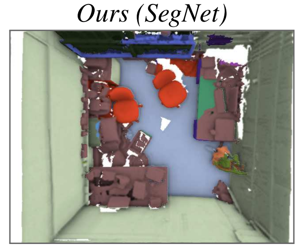
Ground truth

Direct (SegNet)

Ours (SegNet)



Direct (FCN)

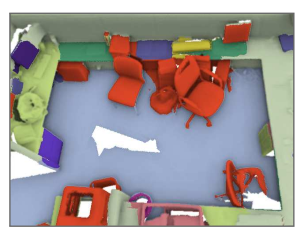


Ours (FCN)

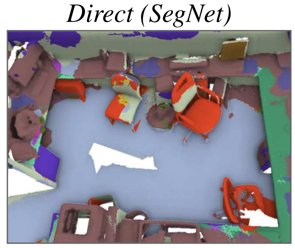
SceneNN/086

Ground truth

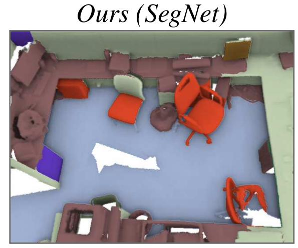
Direct (SegNet)



Ours (SegNet)

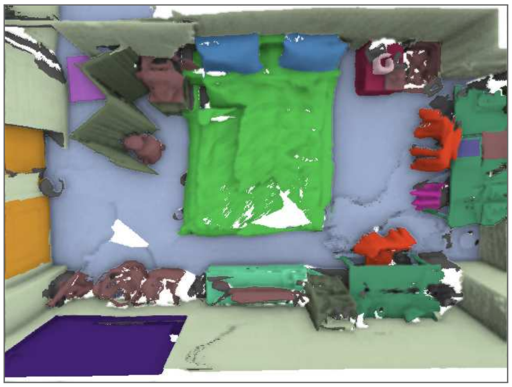


Direct (FCN)



Ours (FCN)

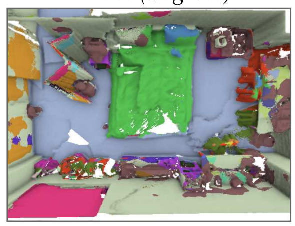
SceneNN/089



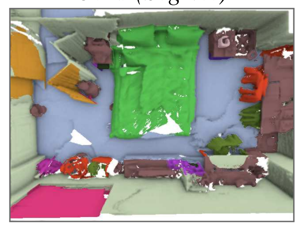
Ground truth

Direct (SegNet)

Ours (SegNet)

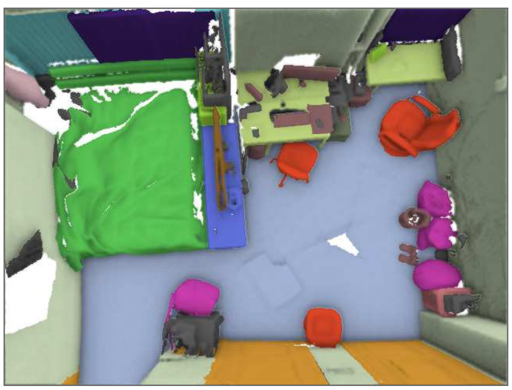


Direct (FCN)



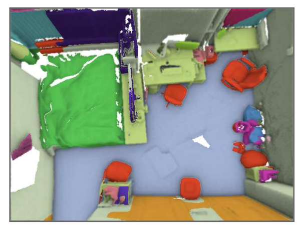
Ours (FCN)

SceneNN/096

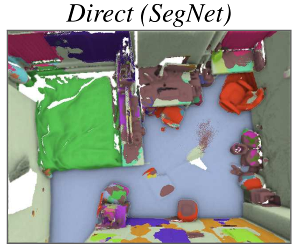


Ground truth

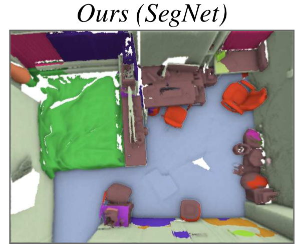
Direct (SegNet)



Ours (SegNet)

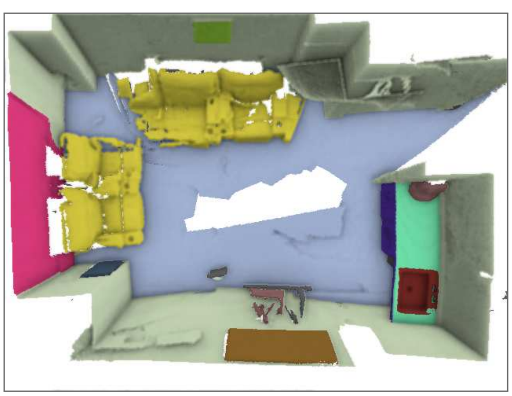


Direct (FCN)

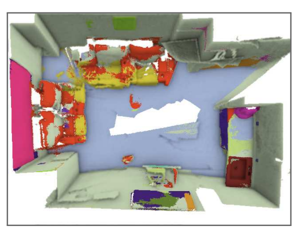


Ours (FCN)

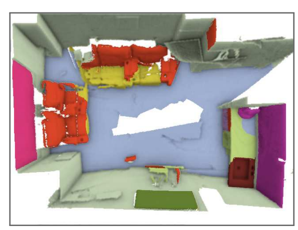
SceneNN/098



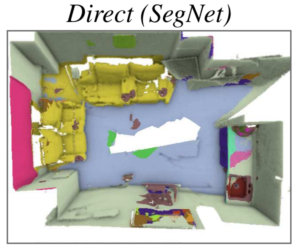
Ground truth



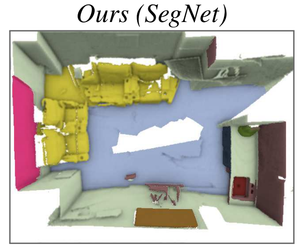
Direct (SegNet)



Ours (SegNet)

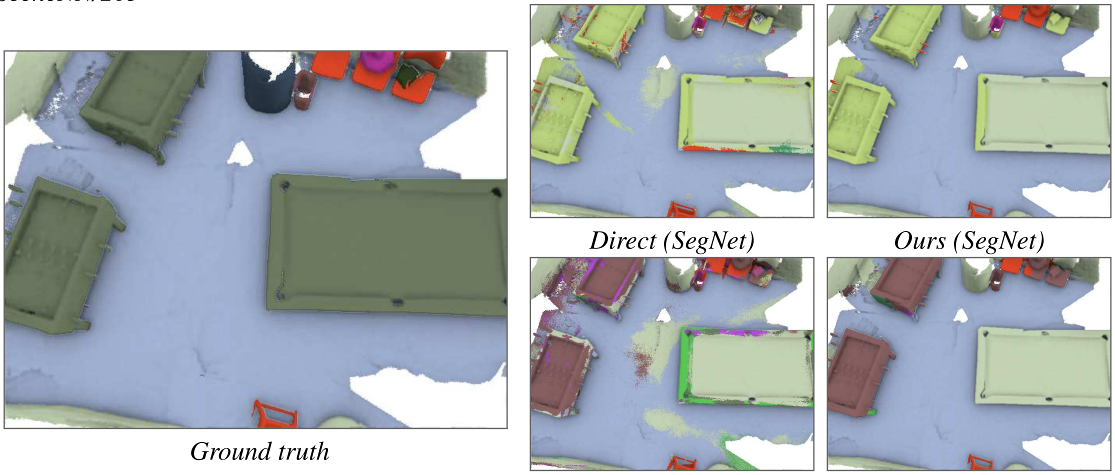


Direct (FCN)



Ours (FCN)

SceneNN/205

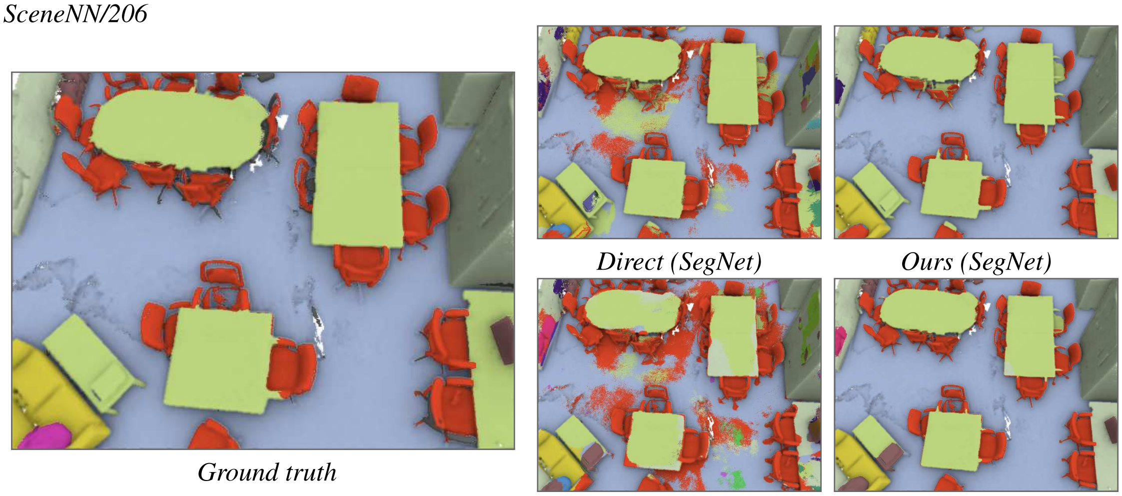


Ground truth

Direct (SegNet) *Ours (SegNet)*

Direct (FCN) *Ours (FCN)*

SceneNN/206

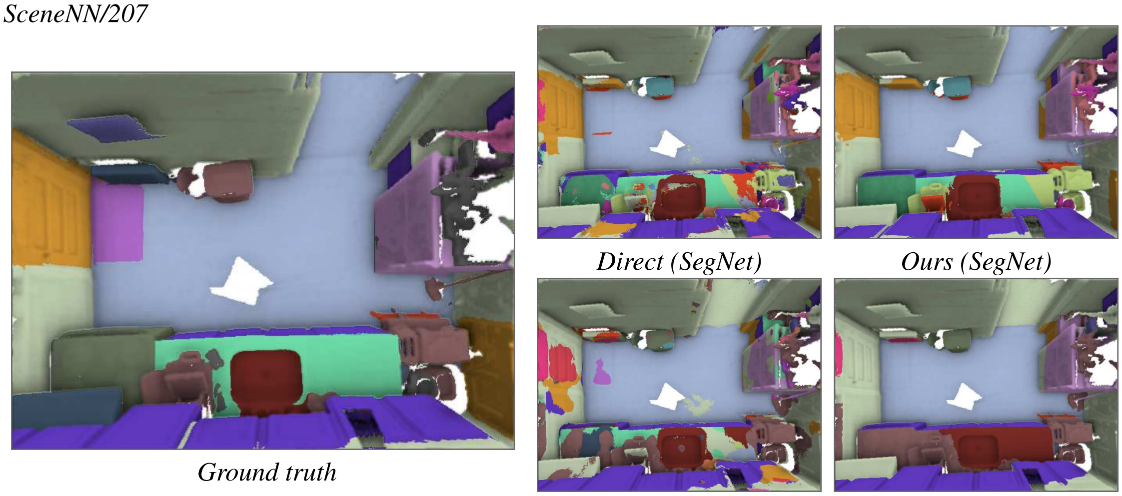


Ground truth

Direct (SegNet) *Ours (SegNet)*

Direct (FCN) *Ours (FCN)*

SceneNN/207

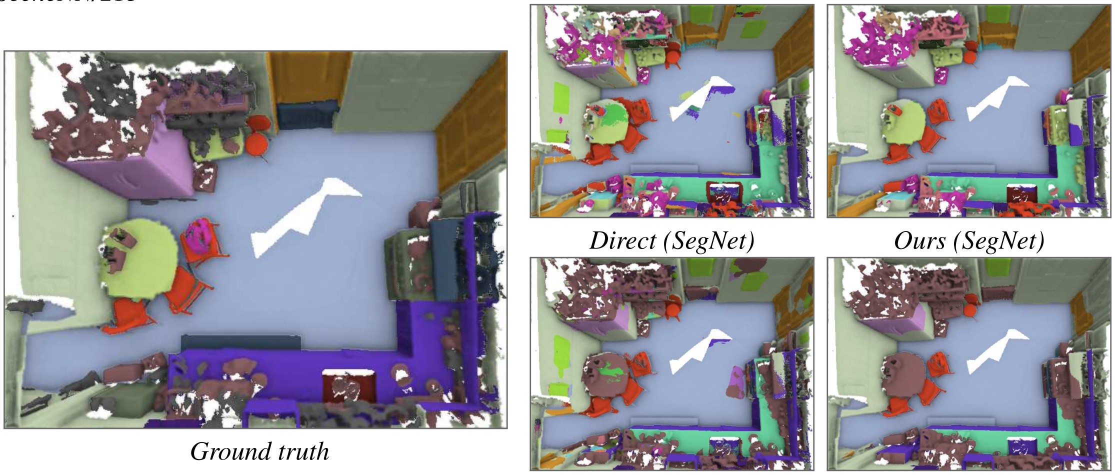


Ground truth

Direct (SegNet) *Ours (SegNet)*

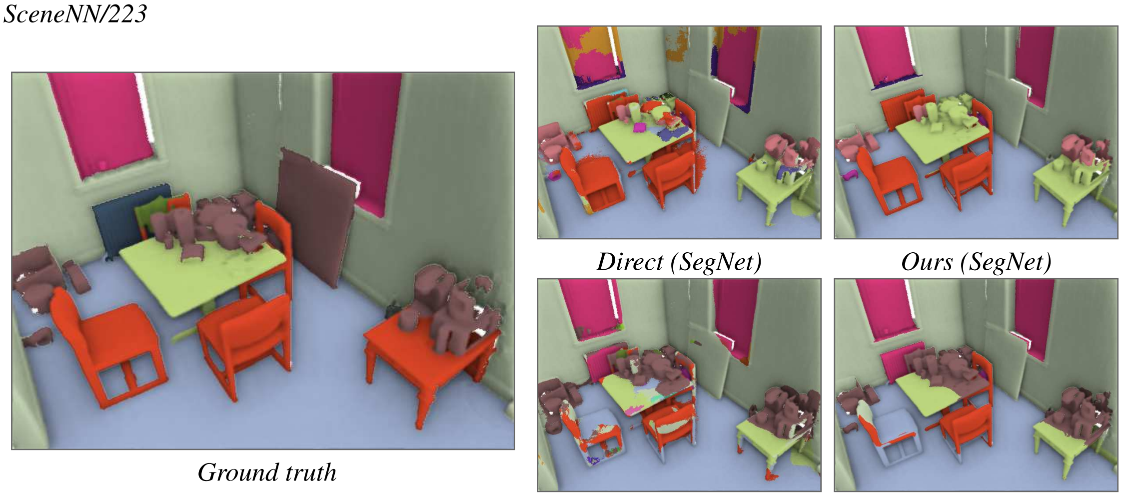
Direct (FCN) *Ours (FCN)*

SceneNN/213



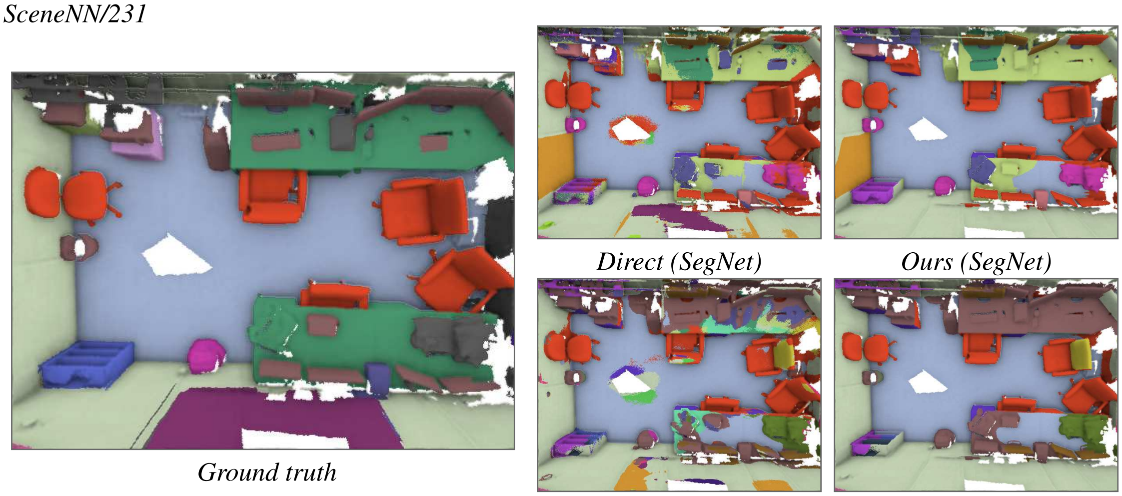
Ground truth — *Direct (SegNet)* — *Ours (SegNet)* — *Direct (FCN)* — *Ours (FCN)*

SceneNN/223



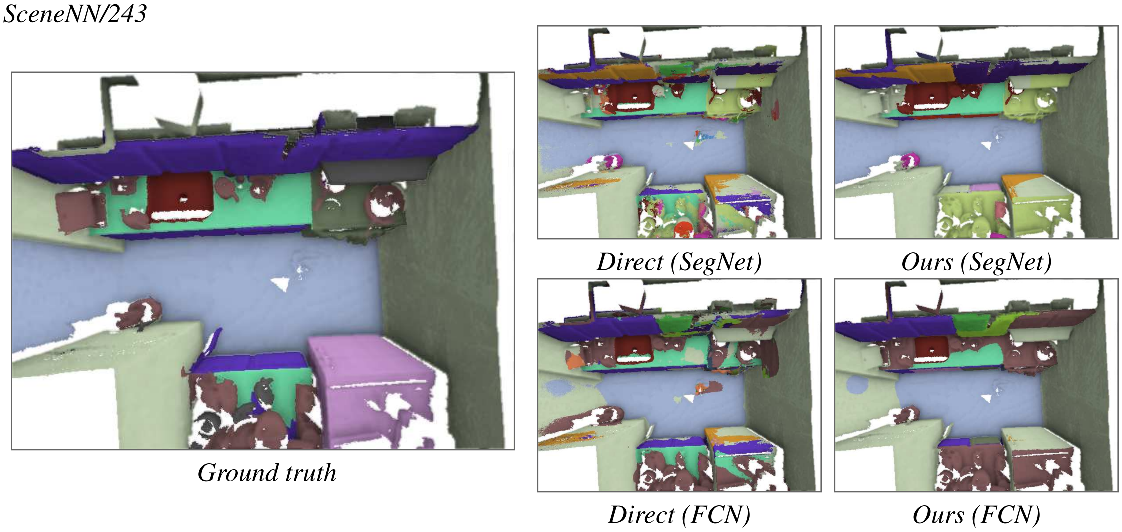
Ground truth — *Direct (SegNet)* — *Ours (SegNet)* — *Direct (FCN)* — *Ours (FCN)*

SceneNN/231

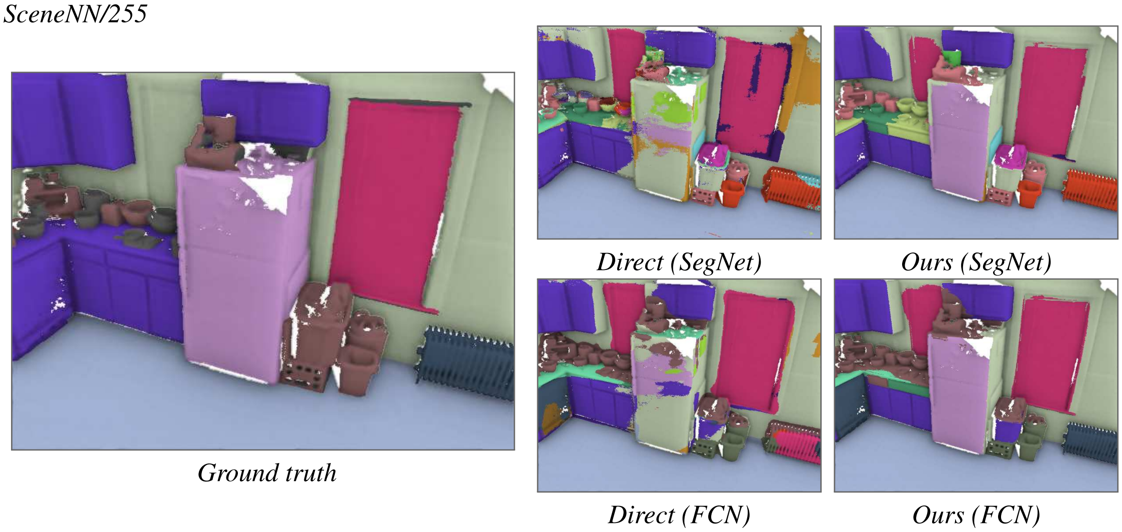


Ground truth — *Direct (SegNet)* — *Ours (SegNet)* — *Direct (FCN)* — *Ours (FCN)*

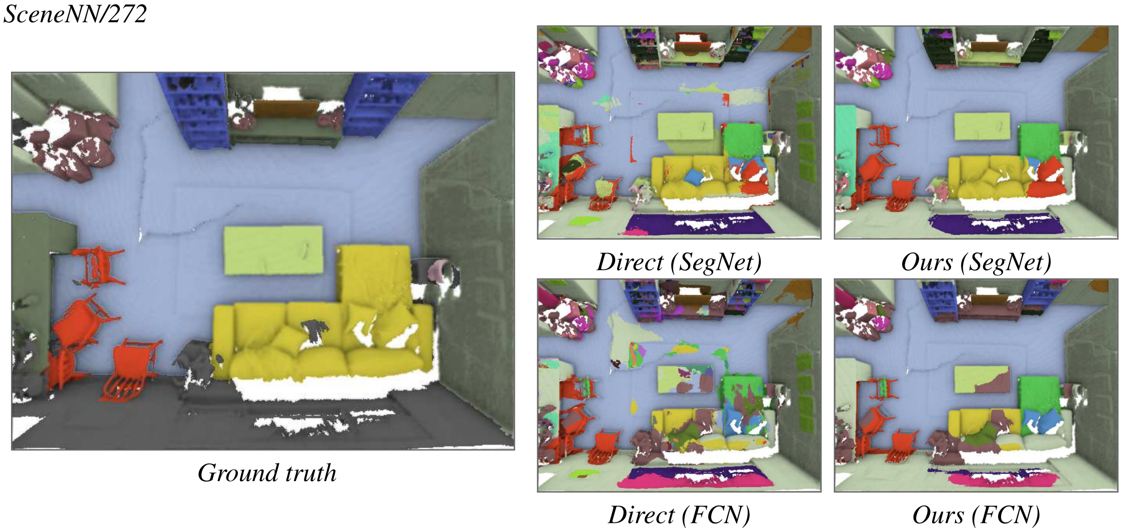
SceneNN/243



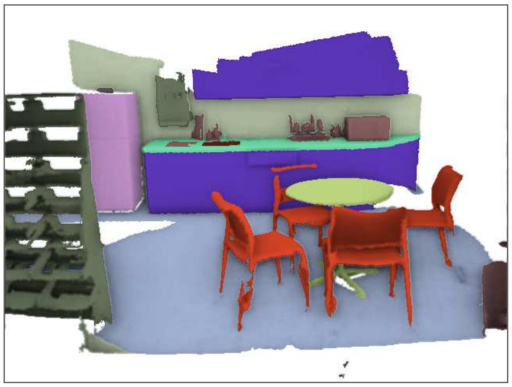
SceneNN/255



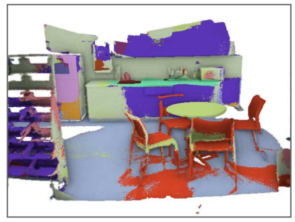
SceneNN/272



SceneNN/322

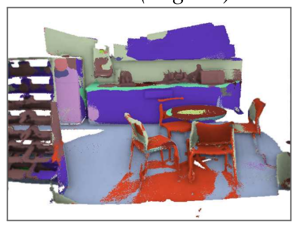


Ground truth

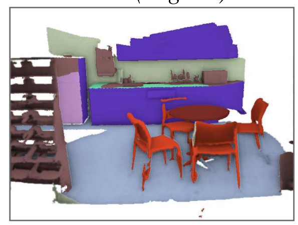


Direct (SegNet)

Ours (SegNet)



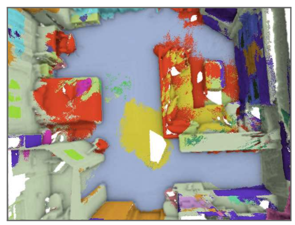
Direct (FCN)



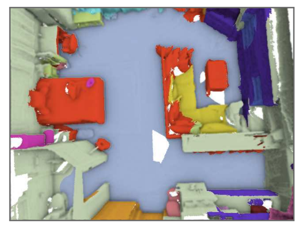
Ours (FCN)

ScanNet/scene0000

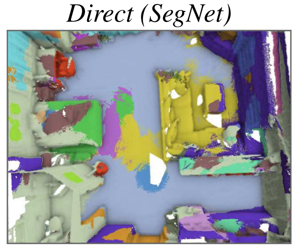
Ground truth



Direct (SegNet)



Ours (SegNet)



Direct (FCN)

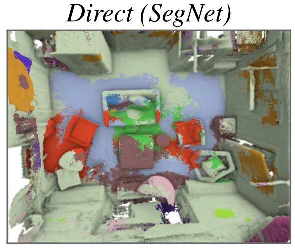
Ours (FCN)

ScanNet/scene0002

Ground truth

Direct (SegNet)

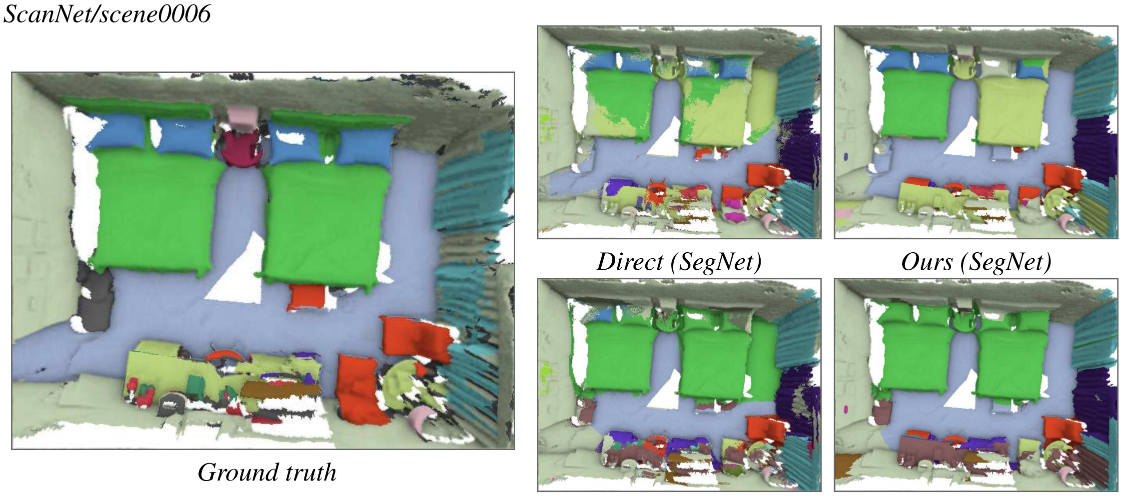
Ours (SegNet)



Direct (FCN)

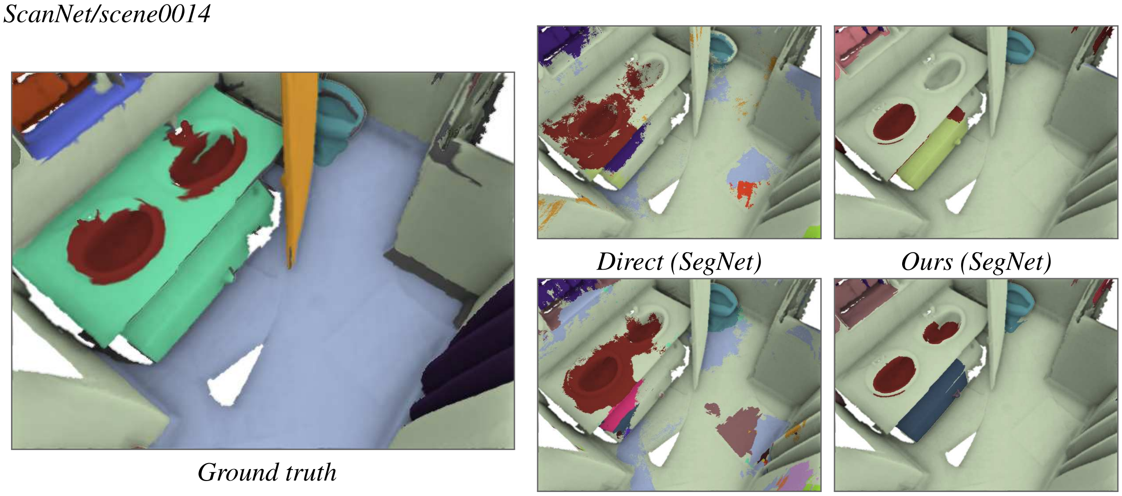
Ours (FCN)

ScanNet/scene0006



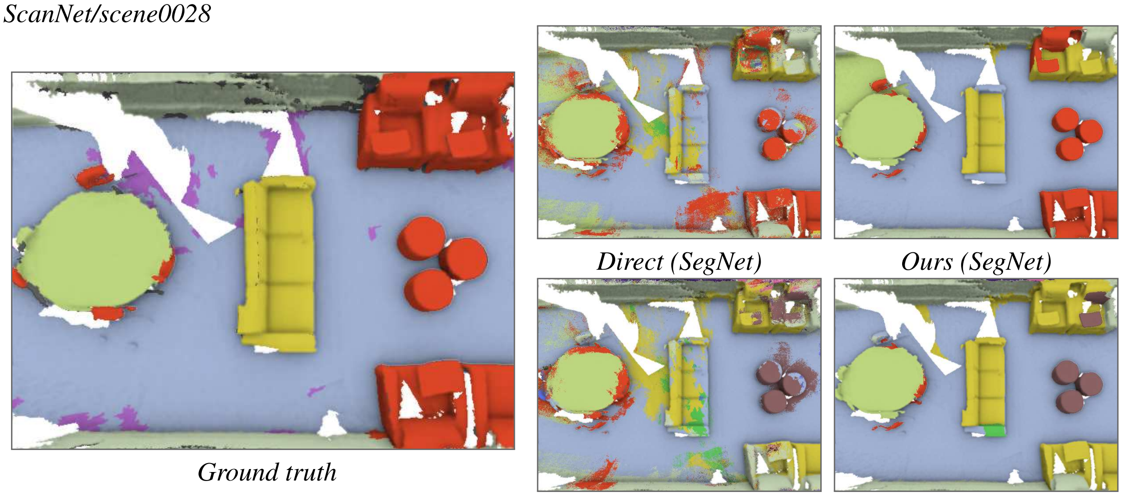
Ground truth | *Direct (SegNet)* | *Ours (SegNet)* | *Direct (FCN)* | *Ours (FCN)*

ScanNet/scene0014



Ground truth | *Direct (SegNet)* | *Ours (SegNet)* | *Direct (FCN)* | *Ours (FCN)*

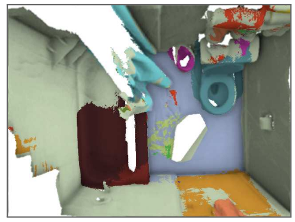
ScanNet/scene0028



Ground truth | *Direct (SegNet)* | *Ours (SegNet)* | *Direct (FCN)* | *Ours (FCN)*

ScanNet/scene0029

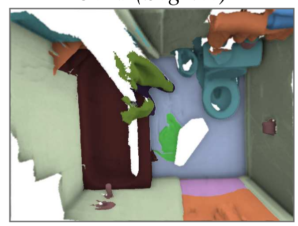
Ground truth



Direct (SegNet)

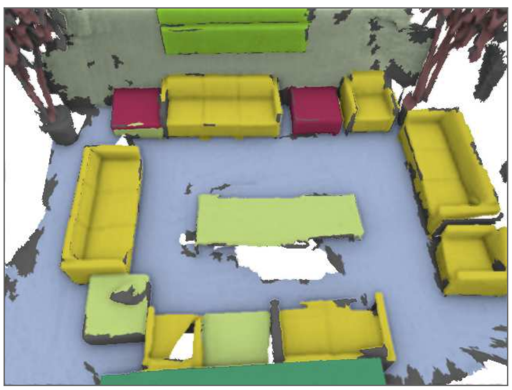
Ours (SegNet)

Direct (FCN)

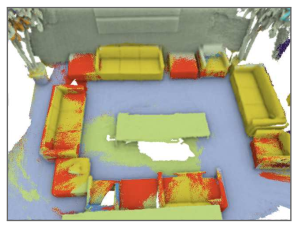


Ours (FCN)

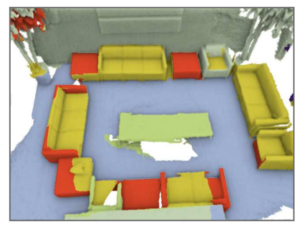
ScanNet/scene0031



Ground truth



Direct (SegNet)

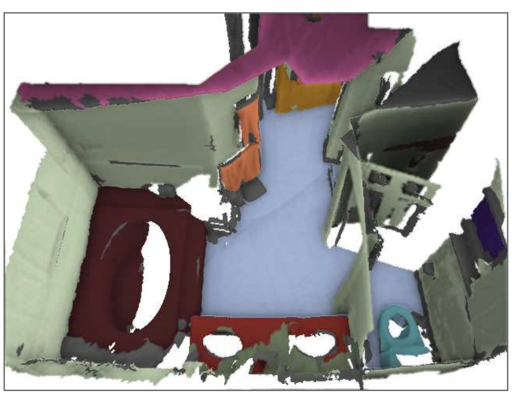


Ours (SegNet)

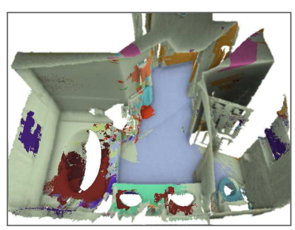
Direct (FCN)

Ours (FCN)

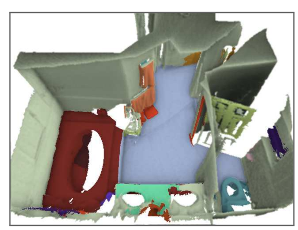
ScanNet/scene0034



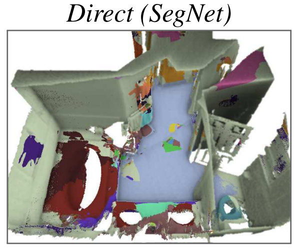
Ground truth



Direct (SegNet)



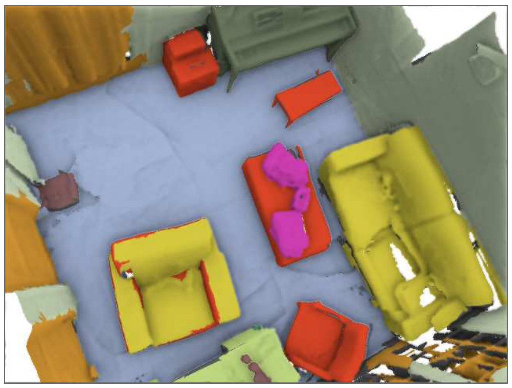
Ours (SegNet)



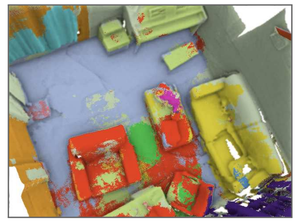
Direct (FCN)

Ours (FCN)

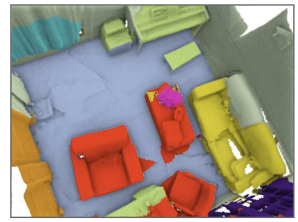
ScanNet/scene0050



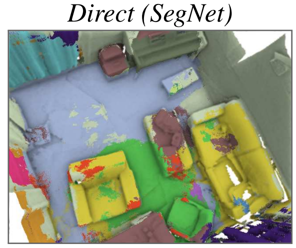
Ground truth



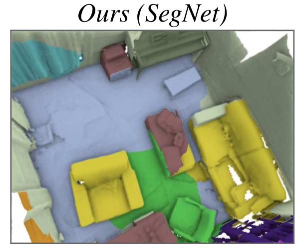
Direct (SegNet)



Ours (SegNet)

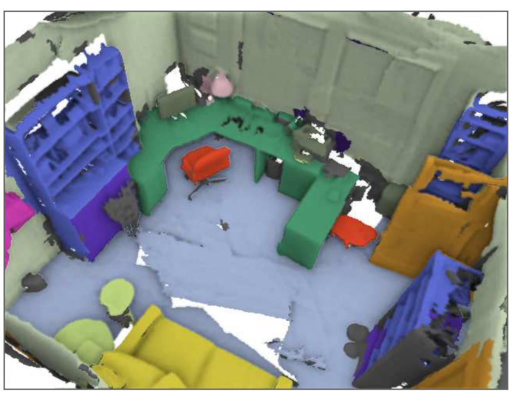


Direct (FCN)



Ours (FCN)

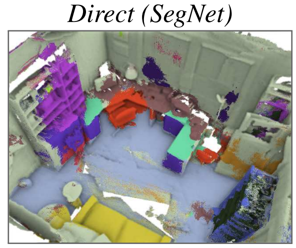
ScanNet/scene0054



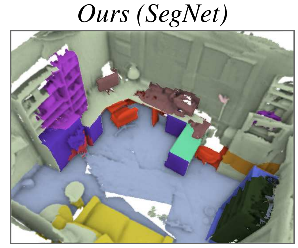
Ground truth

Direct (SegNet)

Ours (SegNet)



Direct (FCN)



Ours (FCN)

4. Mean Field Updates

This section discusses the mean field updates for the RaCRF model introduced in the main paper. For consistency, we use the same notation as in the main paper. We use mean field inference to approximate the joint distribution with the product of marginals $\prod_i Q(x_i = l)$. We denote $Q(\mathbf{x_r})$ as the marginal probability of a clique. In general, mean field updates of such CRF take the following form

$$Q^{t+1}(x_i = l) = \frac{1}{Z_i} \exp\left(- \sum_{\{\mathbf{x}_r | x_j = l\}} \sum Q^t(\mathbf{x}_{r-i}) \psi(\mathbf{x}_r) \right) \tag{3}$$

where Q^{t+1} is the marginal after the t^{th} iteration, $\mathbf{x}_r$ is an label assignment to all variables in the r^{th} region, $\mathbf{x}_{r-i}$ is an assignment to all variables in the r^{th} region except for x_i, $\psi(\mathbf{x}_r)$ is the cost of assigning $\mathbf{x}_r$, and Z_i is the normalization factor that converts $Q(x_i = l)$ into a proper probability distribution after updating.

Updates from objectness potentials The contribution from ψ^A to the update of $Q(\mathbf{x}_r^{(i)} = l)$ takes the form

$$\sum_{\mathbf{x}_r | x_r^{(i)} = l\}} Q(\mathbf{x}_{r-i}) \psi^O(\mathbf{x}_r) = \begin{cases} \frac{1}{|\mathbf{x}_r|} Q(y_r = 0), & \text{if } l \in \mathcal{O}, \\ \frac{1}{|\mathbf{x}_r|} Q(y_r = 1), & \text{if } l \notin \mathcal{O} \end{cases} \tag{4}$$

where $\mathbf{x}_{r-i}$ is a label assignment to the r^{th} region, ignoring the i^{th} vertex. A similar derivation can be found in [1], but the authors used object detection instead of an object proposal method in their formulation.

Updates from consistency potentials The contribution from $\psi^C(\mathbf{x}_r)$ to the mean field update of $Q(\mathbf{x}_r^{(i)} = l)$ is derived as follows

$$\begin{aligned} &\sum_{\{\mathbf{x}_r | x_r^{(i)} = l\}} Q(\mathbf{x}_{r-i}) \psi^C(\mathbf{x}_r) = \\ &\quad - \sum_{l_k \in \mathcal{L}} f_r(l_k) \log f_r(l_k) \left(\prod_{x \in \mathbf{x}_{r-i}} Q(x = l) \right) \\ &\quad - \sum_{l_k \in \mathcal{L}} f_r(l_k) \log f_r(l_k) \left(1 - \prod_{x \in \mathbf{x}_{r-i}} Q(x = l) \right) \end{aligned} \tag{5}$$

Note that we only need to calculate the entropy of a region once after one iteration, and it can be re-used for every vertices inside that region. The frequency calculation can also be used for updates of co-occurrence potentials.

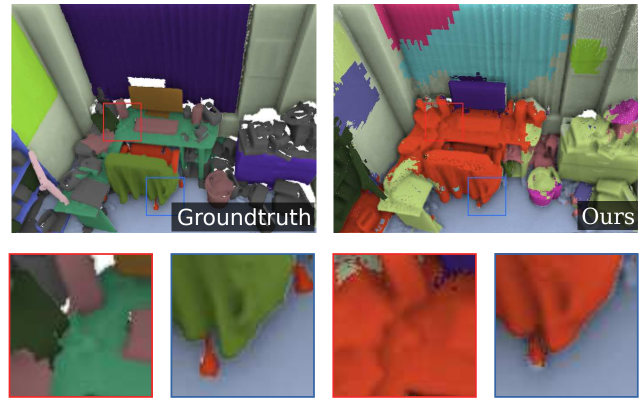


Figure 3: Failure case of the proposed CRF model. The model fails to produce fine segmentation such as cluttered objects on the table, or clothing on a chair. The final segmentation result loses some of the fine details, since the model favors consistency over complexity.

Updates from relationship potentials The contribution from $\psi^R(\mathbf{x}_r, \mathbf{x}_q)$ to the mean field update of $Q(\mathbf{x}_r^{(i)} = l)$ is derived as follows

$$\begin{aligned} &\sum_{\{\mathbf{x}_r | x_r^{(i)} = l\}} \sum_{\{\mathbf{x}_q | x_q^{(i)} = m\}} Q(\mathbf{x}_{r-i}) Q(\mathbf{x}_q) \psi^R(\mathbf{x}_r, \mathbf{x}_q) = \\ &\quad - \sum_{l_i \in \mathcal{L}} \sum_{l_j \in \mathcal{L}} \log f_r(l_i) \log f_q(l_j) \log \gamma_{l_i, l_j} \\ &\quad \times \prod_{x_j \in \mathbf{x}_{r-i}} \prod_{x_k \in \mathbf{x}_q} Q(x_j = l) Q(x_k = m) \end{aligned} \tag{6}$$

Again, we observe that the frequency can be pre-calculated for every regions. However, this update needs to compute the marginal product between pairs of vertices, so it only feasible when the number of connections is small. In our case, we only consider neighboring regions based on mesh connectivity, the computation is still manageable.

5. Failure cases

To highlight a typical failure case, we show an example of over-smoothing in Figure 3.

References

[1] A. Arnab, S. Jayasumana, S. Zheng, and P. H. Torr. Higher order conditional random fields in deep neural networks. In *European Conference on Computer Vision*, pages 524–540. Springer, 2016.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *arXiv preprint arXiv:1511.00561*, 2015.

[3] T. Cavallari and L. Di Stefano. Semanticfusion: Joint labeling, tracking and mapping. In *Computer Vision–ECCV 2016 Workshops*, pages 648–664. Springer, 2016.

[4] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[5] B.-S. Hua, Q.-H. Pham, D. T. Nguyen, M.-K. Tran, L.-F. Yu, and S.-K. Yeung. Scenenn: A scene meshes dataset with annotations. In *International Conference on 3D Vision (3DV)*, 2016.

[6] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[7] J. McCormac, A. Handa, A. Davison, and S. Leutenegger. Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *ICRA*, 2016.