Instance Segmentation & Object Tracking in Point Clouds

Jiantao Gao

2020.6.26

Paper List

- PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation (CVPR 2020 oral)
- OccuSeg: Occupancy-aware 3D Instance Segmentation (CVPR 2020)
- Leveraging Shape Completion for 3D Siamese Tracking (CVPR 2019)

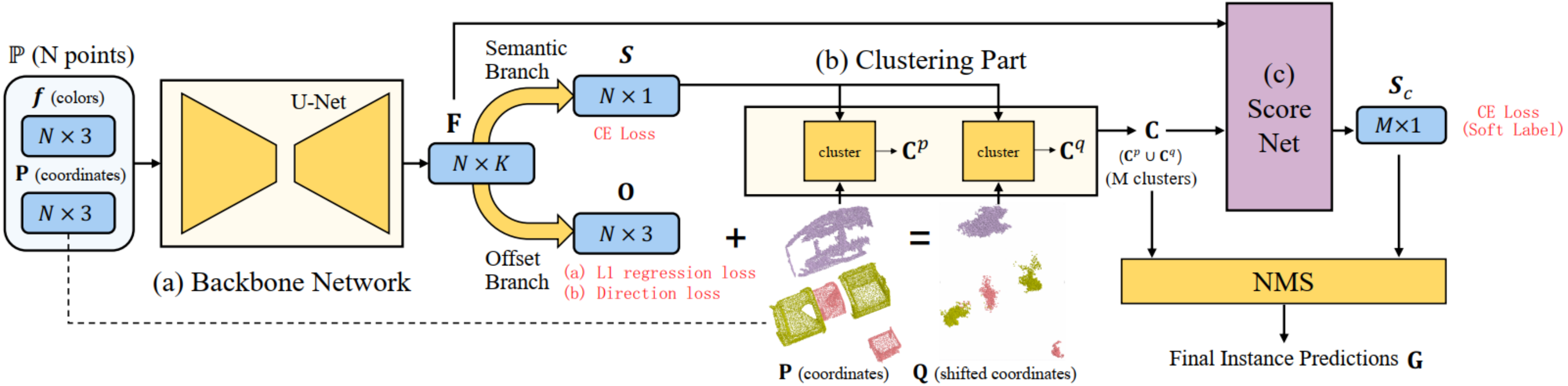
3D Semantic Instance Segmentation

Input: 3D Point Cloud *Object Center Votes & Aggregated Proposals* *Output: 3D Semantic Instances*

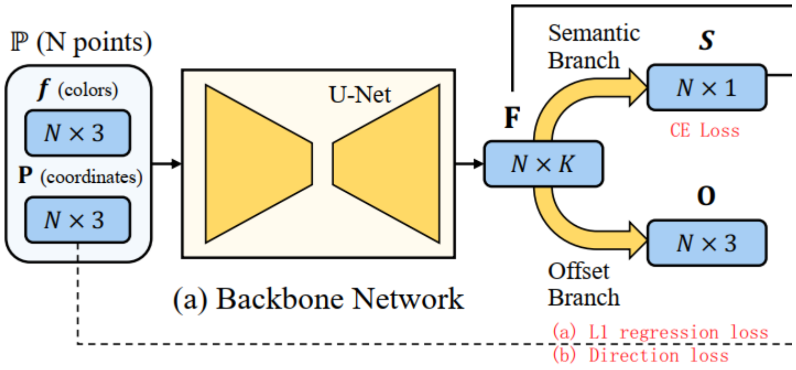
PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation

Li Jiang[1*] Hengshuang Zhao[1*] Shaoshuai Shi[1] Shu Liu[2] Chi-Wing Fu[1] Jiaya Jia[1,2]
[1]The Chinese University of Hong Kong [2]SmartMore
{lijiang, hszhao, cwfu, leojia}@cse.cuhk.edu.hk ssshi@ee.cuhk.edu.hk sliu@smartmore.com

Pipeline:


ℙ (N points)
f (colors)
N × 3
P (coordinates)
N × 3
U-Net
(a) Backbone Network
F
N × K
Semantic Branch
S
N × 1
CE Loss
Offset Branch
O
N × 3
(a) L1 regression loss
(b) Direction loss
+
=
P (coordinates)
Q (shifted coordinates)
(b) Clustering Part
cluster
C^p
cluster
C^q
C
(C^p ∪ C^q)
(M clusters)
(c) Score Net
S_c
M×1
CE Loss
(Soft Label)
NMS
Final Instance Predictions G

Backbone Network:



1. Semantic Branch: CE Loss
2. Offset Branch:

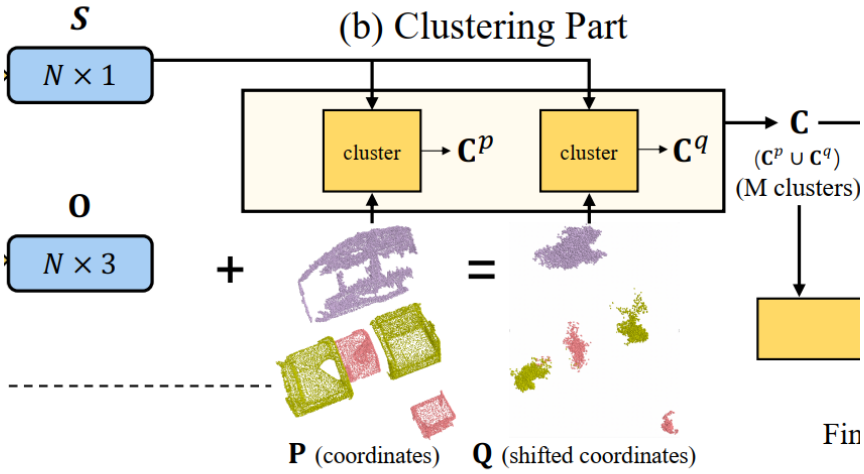
   a. L1 regression loss:

$$L_{o\_reg} = \frac{1}{\sum_i m_i} \sum_i ||o_i - (\hat{c}_i - p_i)|| \cdot m_i$$

   b. Direction Loss:

$$L_{o\_dir} = -\frac{1}{\sum_i m_i} \sum_i \frac{o_i}{||o_i||_2} \cdot \frac{\hat{c}_i - p_i}{||\hat{c}_i - p_i||_2} \cdot m_i.$$

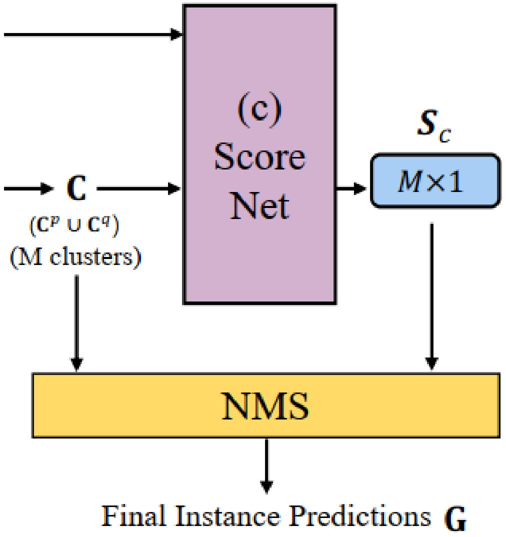
Clustering Part:



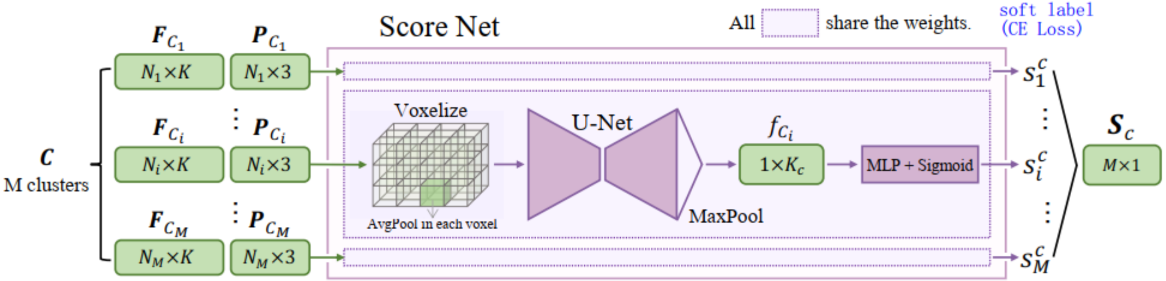
Fail to separate same category objects that are close to each other (two pictures that hang side-by-side on the wall)

For points near object boundary, the predicted offsets may not be accurate.

ScoreNet:



CE Loss
(Soft Label)



(a) ScoreNet

$$\hat{s}_i^c = \begin{cases} 0 & iou_i < \theta_l \\ 1 & iou_i > \theta_h \\ \frac{1}{\theta_h-\theta_l}\cdot(iou_i-\theta_l) & otherwise \end{cases},$$

$$L_{c\_score} = -\frac{1}{M}\sum_{i=1}^{M}(\hat{s}_i^c log(s_i^c)+(1-\hat{s}_i^c)log(1-s_i^c)). \quad (7)$$

Experiments:

ScanNet v2 :

Method	Avg AP_{50}	bathtub	bed	bookshe.	cabinet	chair	counter	curtain	desk	door	otherfu.	picture	refrige.	s. curtain	sink	sofa	table	toilet	window
SGPN [49]	0.143	0.208	0.390	0.169	0.065	0.275	0.029	0.069	0.000	0.087	0.043	0.014	0.027	0.000	0.112	0.351	0.168	0.438	0.138
3D-BEVIS [11]	0.248	0.667	0.566	0.076	0.035	0.394	0.027	0.035	0.098	0.099	0.030	0.025	0.098	0.375	0.126	0.604	0.181	0.854	0.171
R-PointNet [54]	0.306	0.500	0.405	0.311	0.348	0.589	0.054	0.068	0.126	0.283	0.290	0.028	0.219	0.214	0.331	0.396	0.275	0.821	0.245
DPC [12]	0.355	0.500	0.517	0.467	0.228	0.422	**0.133**	0.405	0.111	0.205	0.241	0.075	0.233	0.306	0.445	0.439	0.457	0.974	0.23
3D-SIS [19]	0.382	1.000	0.432	0.245	0.190	0.577	0.013	0.263	0.033	0.320	0.240	0.075	0.422	0.857	0.117	0.699	0.271	0.883	0.235
MASC [27]	0.447	0.528	0.555	0.381	0.382	0.633	0.002	0.509	0.260	0.361	0.432	0.327	0.451	0.571	0.367	0.639	0.386	0.980	0.276
PanopticFusion [32]	0.478	0.667	0.712	0.595	0.259	0.550	0.000	0.613	0.175	0.250	0.434	0.437	0.411	0.857	0.485	0.591	0.267	0.944	0.35
3D-BoNet [53]	0.488	1.000	0.672	0.590	0.301	0.484	0.098	0.620	0.306	0.341	0.259	0.125	0.434	0.796	0.402	0.499	0.513	0.909	0.439
MTML [23]	0.549	1.000	**0.807**	0.588	0.327	0.647	0.004	**0.815**	0.180	0.418	0.364	0.182	0.445	1.000	0.442	0.688	**0.571**	**1.000**	0.396
PointGroup (Ours)	**0.636**	**1.000**	0.765	**0.624**	**0.505**	**0.797**	0.116	0.696	**0.384**	**0.441**	**0.559**	**0.476**	**0.596**	**1.000**	**0.666**	**0.756**	0.556	0.997	**0.513**

Table 1: 3D instance segmentation results on ScanNet v2 testing set with AP_{50} scores. Our proposed PointGroup approach yields the highest average AP_{50}, outperforming all state-of-the-art methods by a large margin. All numbers are from the ScanNet benchmark on 15/11/2019.

Experiments:

S3DIS:

Method	AP_{50}	$mPrec_{50}$	$mRec_{50}$
SGPN† [49]	-	0.360	0.287
ASIS† [50]	-	0.553	0.424
PointGroup†	**0.578**	**0.619**	**0.621**
SGPN‡ [49]	0.544	0.382	0.312
PartNet‡ [31]	-	0.564	0.434
ASIS‡ [50]	-	0.636	0.475
3D-BoNet‡ [53]	-	0.656	0.476
PointGroup‡	**0.640**	**0.696**	**0.692**

Table 5: Instance segmentation results on the S3DIS validation set. Methods marked with † are evaluated on Area 5; those marked with ‡ are on the 6-fold cross validation.

Experiments:

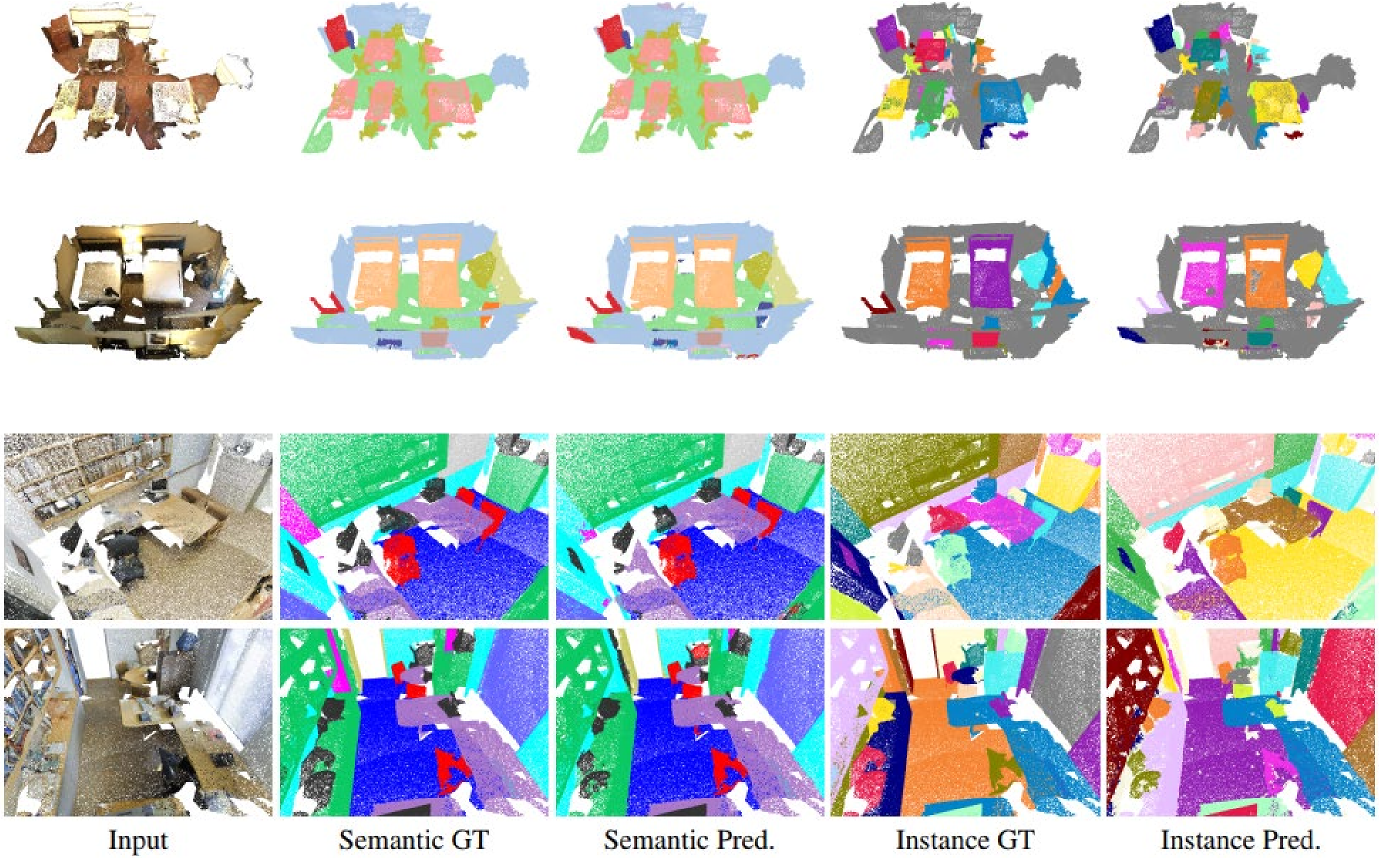


Figure 5: Visualization of the semantic and instance segmentation results on ScanNet v2 (top) and S3DIS (bottom). For instance predictions, different colors represent separate instances, and the semantic results indicate the categories of instances.

Ablation Study on ScanNet v2:

Method	Metric	mean	bathtub	bed	bookshe.	cabinet	chair	counter	curtain	desk	door	otherfu.	picture	refrige.	s. curtain	sink	sofa	table	toilet	window
Original **P**	AP	0.283	0.414	0.327	0.244	0.167	0.493	0.083	0.269	0.089	0.193	0.286	0.205	0.207	0.373	0.226	0.361	0.251	0.684	0.231
	AP_{50}	0.507	0.692	0.647	0.481	0.347	0.685	0.231	0.508	0.308	0.384	0.453	0.359	0.301	0.632	0.537	0.660	0.531	0.961	0.413
	AP_{25}	0.659	0.840	0.764	0.597	0.496	0.791	0.588	0.614	0.686	0.529	0.600	0.432	0.401	0.660	0.775	0.777	0.721	0.995	0.601
Shifted **Q**	AP	0.328	0.499	0.383	0.248	0.217	0.713	0.008	0.241	0.165	0.216	0.318	0.211	0.238	0.422	0.292	0.383	0.362	0.799	0.194
	AP_{50}	0.529	0.738	0.694	0.550	0.435	0.884	0.035	0.389	0.410	0.413	0.501	0.363	0.366	0.617	0.590	0.648	0.571	0.948	0.375
	AP_{25}	0.677	0.863	0.795	0.699	0.617	0.931	0.426	0.541	0.697	0.538	0.623	0.446	0.366	0.765	0.826	0.848	0.669	0.999	0.533
Both **P** & **Q**	AP	0.348	0.597	0.376	0.267	0.253	0.712	0.069	0.266	0.140	0.229	0.339	0.208	0.246	0.416	0.298	0.434	0.385	0.758	0.275
	AP_{50}	0.569	0.805	0.696	0.549	0.481	0.877	0.224	0.449	0.416	0.420	0.530	0.377	0.372	0.644	0.611	0.715	0.629	0.983	0.462
	AP_{25}	0.713	0.865	0.795	0.744	0.673	0.925	0.648	0.616	0.741	0.548	0.654	0.482	0.383	0.711	0.828	0.851	0.742	1.000	0.636

Table 2: Ablation results using different coordinate sets on the ScanNet v2 validation set. Adopting both the original and shifted coordinates for clustering yields the best 3D instance segmentation performance.

Input | (i) **P** Only | (ii) **Q** Only | (iii) **P** and **Q** | Shifted Coord.

Figure 4: Instance predictions produced by models trained with clustering on (i) **P** only, (ii) shifted coordinates **Q** only, and (iii) both. The last column shows the predicted instances of (iii) represented with **Q**, where stuff points are ignored.

Ablation Study on ScanNet v2:

Method	avg AP	avg AP_{50}	avg AP_{25}
r = 2cm	0.285	0.501	0.651
r = 3cm	**0.348**	**0.569**	**0.713**
r = 4cm	0.337	0.552	0.700
r = 5cm	0.342	0.552	0.699

Table 3: Ablation results for clustering with different radii r on the ScanNet v2 validation set.

	#Points	Total Time	BB	Clustering on P and Q				SCN	NMS
				BQ_p	CL_p	BQ_q	CL_q		
1	239,261	865	332	95	16	95	70	176	82
2	45,557	261	177	5	2	5	5	52	14
3	186,857	567	281	44	9	45	31	95	62
4	60,071	271	180	6	3	7	15	55	6
avg	132,937	**491**	243	38	8	38	30	95	41

Table 4: Inference time (ms). BB denotes backbone + two branches; BQ denotes ballquery; subscripts p and q denote clustering on P and Q respectively; CL denotes our clustering algorithm; and SCN denotes ScoreNet.

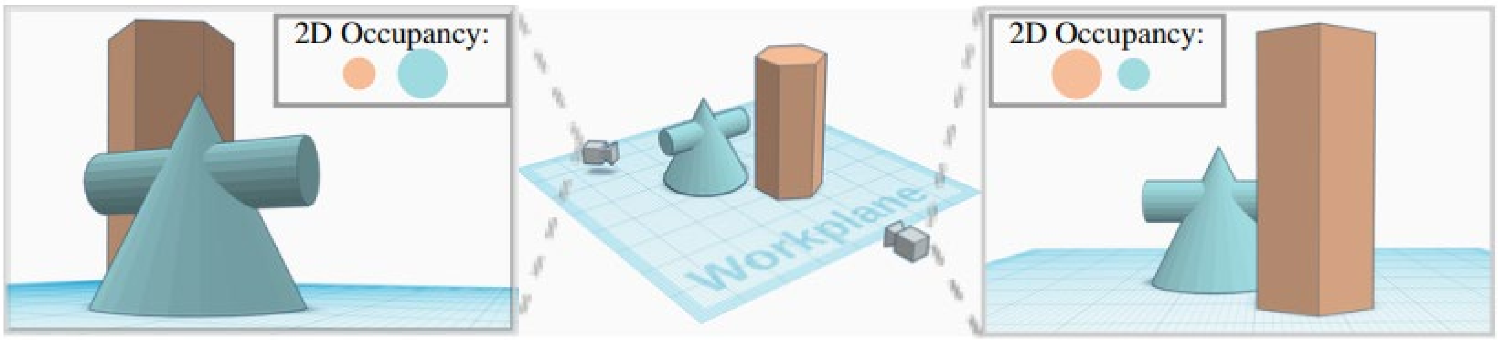
OccuSeg: Occupancy-aware 3D Instance Segmentation

Lei Han[1,2], Tian Zheng[1], Lan Xu[1,2], and Lu Fang[1✉]

[1]Tsinghua University [2]Hong Kong University of Science and Technology

Motivation:

2D/3D Occupancy: the number of pixels/voxels occupied by each instance

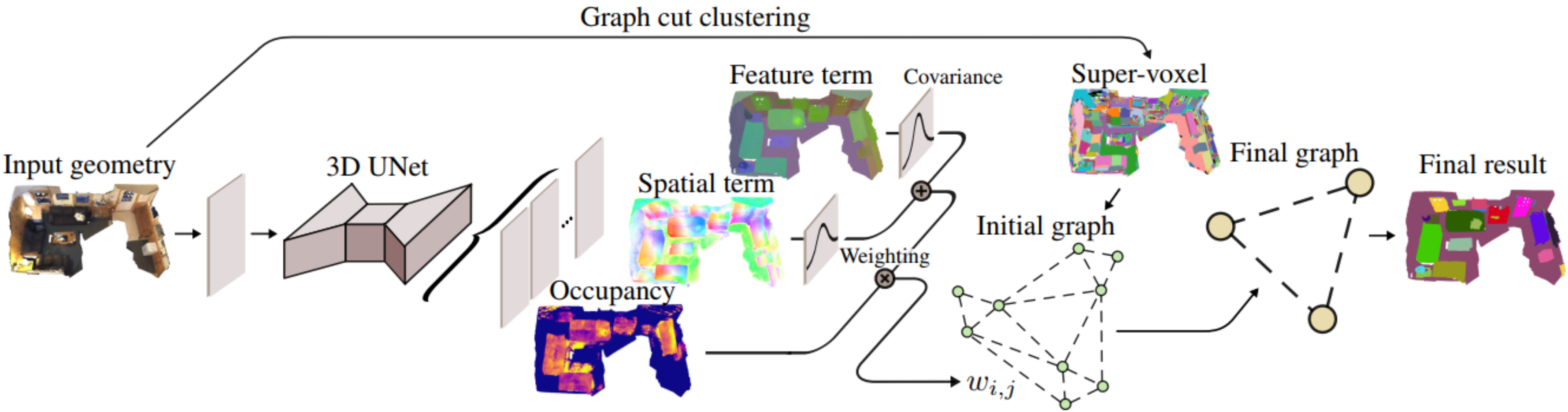


2D Occupancy is uncertain on 2D image,

3D Occupancy is certain and can be predicted robustly for the reconstructed 3D model

Introduce such occupancy signal to guide the clustering stage of 3D instance segmentation ?

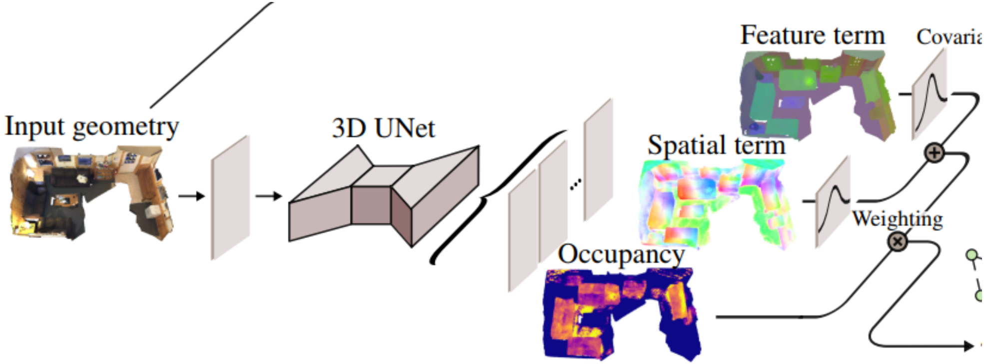
Pipeline:



Multi-task Learning **Instance Clustering**

Multi-task Learning:

$$\mathcal{L}_{\text{sp}} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i=1}^{N_c}||\mathbf{d}_i + \mu_i - \frac{1}{N_c}\sum_{i=1}^{N_c}\mu_i||$$

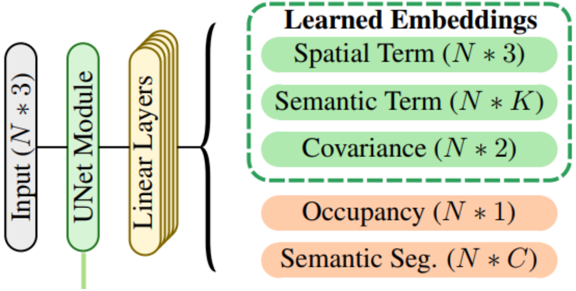


$$\mathcal{L}_{\text{se}} = \mathcal{L}_{\text{var}} + \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{reg}},$$

$$\mathcal{L}_{\text{var}} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i=1}^{N_C}[||\mathbf{u}_c - \mathbf{s}_i|| - \delta_v]_+^2, \tag{5}$$

$$\mathcal{L}_{\text{dist}} = \frac{1}{C(C-1)}\sum_{c_A=1}^{C}\sum_{c_B=c_A+1}^{C}[2\delta_d - ||\mathbf{u}_{c_A} - \mathbf{u}_{c_B}||]_+^2, \tag{6}$$

$$\mathcal{L}_{\text{reg}} = \frac{1}{C}\sum_{c=1}^{C}||\mathbf{u}_c||. \tag{7}$$



$$\mathrm{L_e}\begin{cases}\mathrm{L_{sp}}\\ \mathrm{L_{se}} = \mathrm{L_{var}} + \mathrm{L_{dist}} + \mathrm{L_{reg}}\\ \mathrm{L_{cov}}\end{cases}$$

$\mathrm{L_o}$

$\mathrm{L_c}$: CE Loss

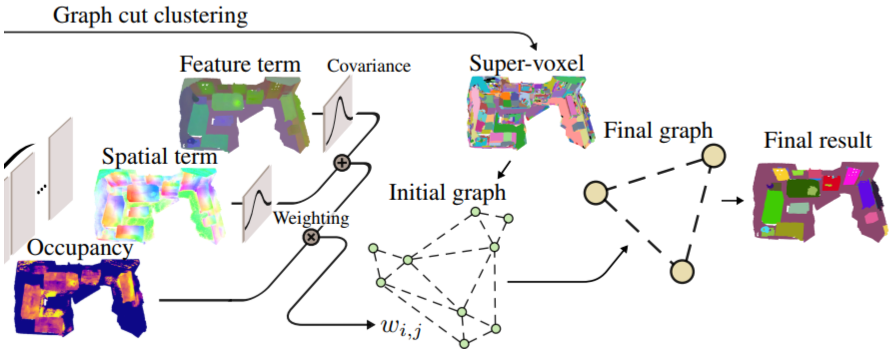
$$\mathcal{L}_{\text{cov}} = -\frac{1}{C}\sum_{c=1}^{C}\frac{1}{N}\sum_{i=1}^{N}[y_i log(p_i) + (1-y_i)log(1-p_i)]$$

$$p_i = \exp(-(\frac{||\mathbf{s}_i - \mathbf{u}_c||}{\sigma_s^c})^2 - (\frac{||\mu_i + \mathbf{d}_i - \mathbf{e}_c||}{\sigma_d^c})^2)$$

$$\mathcal{L}_{\text{o}} = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{i=1}^{N_c}||o_i - \log(N_c)||$$

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{c}} + \mathcal{L}_{\text{e}} + \mathcal{L}_{\text{o}}.$$

Instance Clustering:



1. Bottom-up Strategy:

   Input voxels - - *grouping* - - > super-voxels v_i

$$\mathbf{D}_i = \frac{1}{|\Omega_i|} \sum_{k \in \Omega_i} (\mathbf{d}_i + \mu_i), \qquad S_i \ldots O_i \ldots \sigma_i$$

Ω_i denotes the collection of all the voxels belonging to the super-voxel v_i

2. Establish an undirected graph $G = (V, E, W)$:

   $v_i \in V$: the generated super-voxels

   $(v_i, v_j) \in E$: the pairs of vertices with a weight $w_{i,j}$

   $w_{i,j} \in W$: the weights for the pairs of vertices

$$w_{i,j} = \frac{\exp(-(\frac{||\mathbf{S_i}-\mathbf{S_j}||}{\sigma_s})^2 - (\frac{||\mathbf{D_i}-\mathbf{D_j}||}{\sigma_d})^2)}{max(r, 0.5)}, \qquad r_i = \frac{O_i}{|\Omega_i|}.$$

3. Select the edge $e_i = (v_i, v_j) \in E$ with the highest weight $w_{i,j}$, if $w_{i,j} > T_0 = 0.5$, merge the super-voxels v_i, v_j as a new vertex

4. Iterate until none of the weight is larger than $T_0 = 0.5$.

5. Finally, the remaining vertices are labeled as instances if their occupancy ratio $0.3 < r < 2$

Experiments:

ScanNet v2 :

Method	mAP	bath	bed	bkshf	cab	chair	cntr	curt	desk	door	ofurn	pic	fridg	showr	sink	sofa	tabl	toil	wind
3D-SIS [18]	16.1	40.7	15.5	6.8	4.3	34.6	0.1	13.4	0.5	8.8	10.6	3.7	13.5	32.1	2.8	33.9	11.6	46.6	9.3
PanopticFusion [31]	21.4	25.0	33.0	27.5	10.3	22.8	0.0	34.5	2.4	8.8	20.3	18.6	16.7	36.7	12.5	22.1	11.2	66.6	16.2
3D-BoNet [48]	25.3	51.9	32.4	25.1	13.7	34.5	3.1	41.9	6.9	16.2	13.1	5.2	20.2	33.8	14.7	30.1	30.3	65.1	17.8
MTML [22]	28.2	57.7	38.0	18.2	10.7	43.0	0.1	**42.2**	5.7	17.9	16.2	7.0	22.9	51.1	16.1	49.1	31.3	65.0	16.2
Occipital-SCS	32.0	67.9	35.2	33.4	22.9	43.6	2.5	41.2	5.8	16.1	24.0	8.5	26.2	49.6	18.7	46.7	32.8	77.5	23.1
OccuSeg	**44.3**	**85.2**	**56.0**	**38.0**	**24.9**	**67.9**	**9.7**	34.5	**18.6**	**29.8**	**33.9**	**23.1**	**41.3**	**80.7**	**34.5**	**50.6**	**42.4**	**97.2**	**29.1**

Table 1. Quantitative comparison on the ScanNetV2 [4] benchmark in terms of mAP score on 18 classes. Our approach achieves **the best performance in 17 out of 18 classes**. Note that the ScanNetV2 benchmark data is accessed on 11/14/2019.

	mAP	mAP@0.5	mAP@0.25
3D-SIS [18]	16.1	38.2	55.8
3D-BoNet [48]	25.3	48.8	68.7
MASC [28]	25.4	44.7	61.5
MTML [22]	28.2	54.9	73.1
Occipital-SCS	32.0	51.2	68.8
OccuSeg	**44.3**	**63.4**	**73.9**

Table 2. Quantitative results on the ScanNetV2 [4] benchmark in terms of mAP, mAP@0.5 and mAP@0.25, respectively. Our approach outperforms previous methods by a significant margin. ScanNetV2 benchmark data is accessed on 11/14/2019.

Experiments:

S3DIS:

	mPrec	mRec
PartNet [30]	56.4	43.4
ASIS [44]	63.6	47.5
3D-BoNet [48]	65.6	47.6
OccuSeg	**72.8**	**60.3**

Table 3. Comparison on the S3DIS [1] dataset. Our method outperforms previous methods in terms of mean Precision (mPrec) and mean recall (mRec) with an IoU threshold of 0.5.

Experiments:

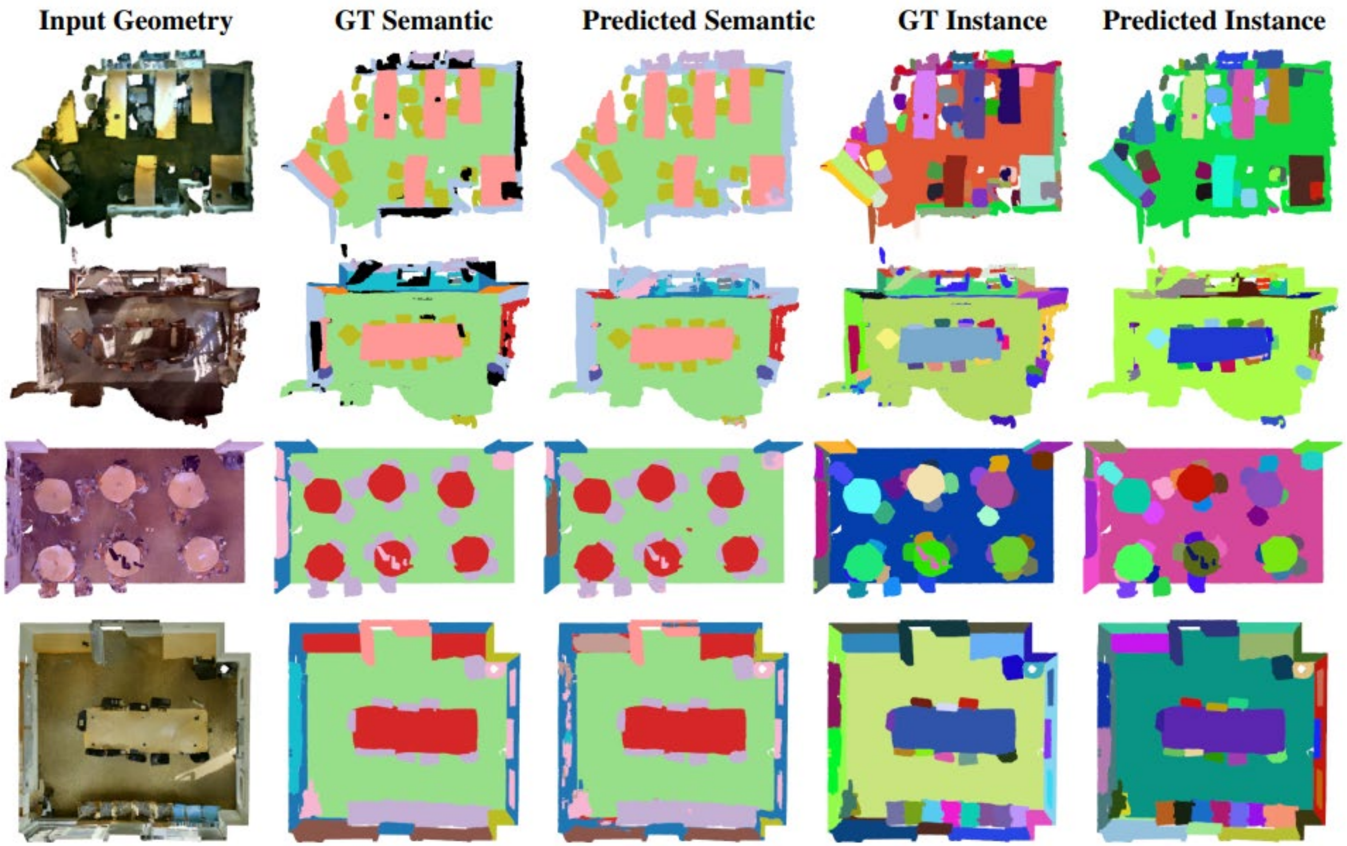


Figure 6. Representative 3D instance segmentation results on the validation set of public datasets, including ScanNetV2 [4] and S3DIS [1].

Experiments:

SceneNN:

Method	**mAP@0.5**	wall	floor	cabinet	bed	chair	sofa	table	desk	tv	prop
MT-PNet [36]	8.5	13.1	27.3	0.0	15.0	21.2	0.0	0.7	0.0	6.0	2.0
MLS-CRF [36]	12.1	13.9	44.5	0.0	32.9	12.9	0.0	5.7	10.8	0.0	0.8
OccuSeg	**47.1**	**39.0**	**93.8**	**5.7**	**66.7**	**91.3**	**8.7**	**50.0**	**31.6**	**76.9**	**7.14**

Table 4. Quantitative results on the SceneNN [19] dataset in terms of mAP@0.5 score of each class. Our approach achieves the best performance for all the 10 classes.

Ablation Study on ScanNet v2:

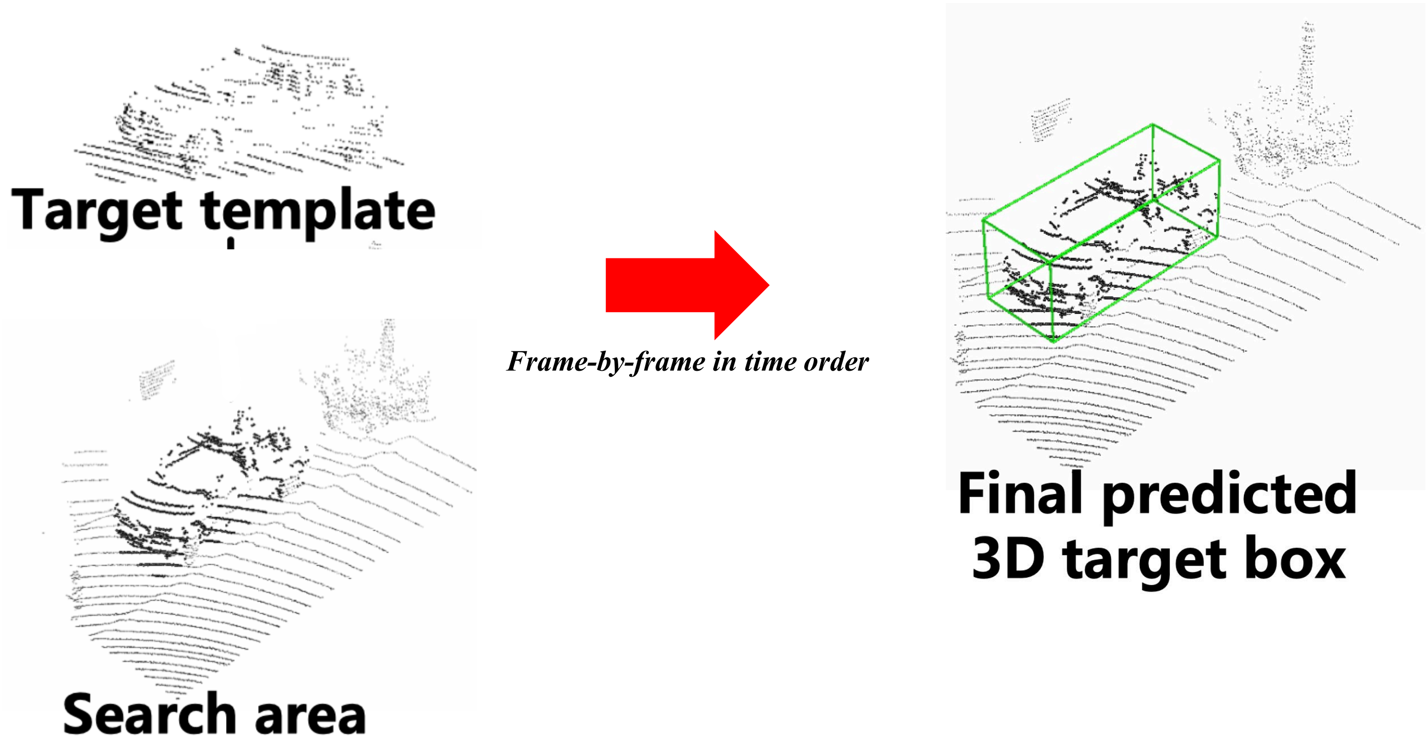
	mAP	mAP@0.5	mAP@0.25
w/o_feature	36.7	51.8	62.6
w/o_spatial	42.8	58.5	69.7
w/o_occupancy	40.9	55.7	67.4
OccuSeg	44.2	60.7	71.9

Table 6. Ablation study of each component of our method on the ScanNetV2 validation split, in terms of mAP, mAP@0.5 and mAP@0.25.

	Details	Total
SGPN [43]	network(GPU): 650 group merging(CPU): 46562 block merging(CPU): 2221	49433
ASIS [44]	network(GPU): 650 mean shift(CPU): 53886 block merging(CPU): 2221	56757
GSPN [50]	network(GPU): 500 point sampling(GPU): 2995 neighbour search(CPU): 468	3963
3D-SIS [18]	network (GPU+CPU): 38841	38841
3D-BoNet [48]	network(GPU): 650 SCN (GPU parallel): 208 block merging(CPU): 2221	2871
OccuSeg	network(GPU): 59 supervoxel(CPU): 375 clustering(GPU+CPU): 160	**594**

Table 5. The processing time (seconds) on the validation set of ScanNetV2 [4]. Note that all the other methods are evaluated based on their released codes according to [48].

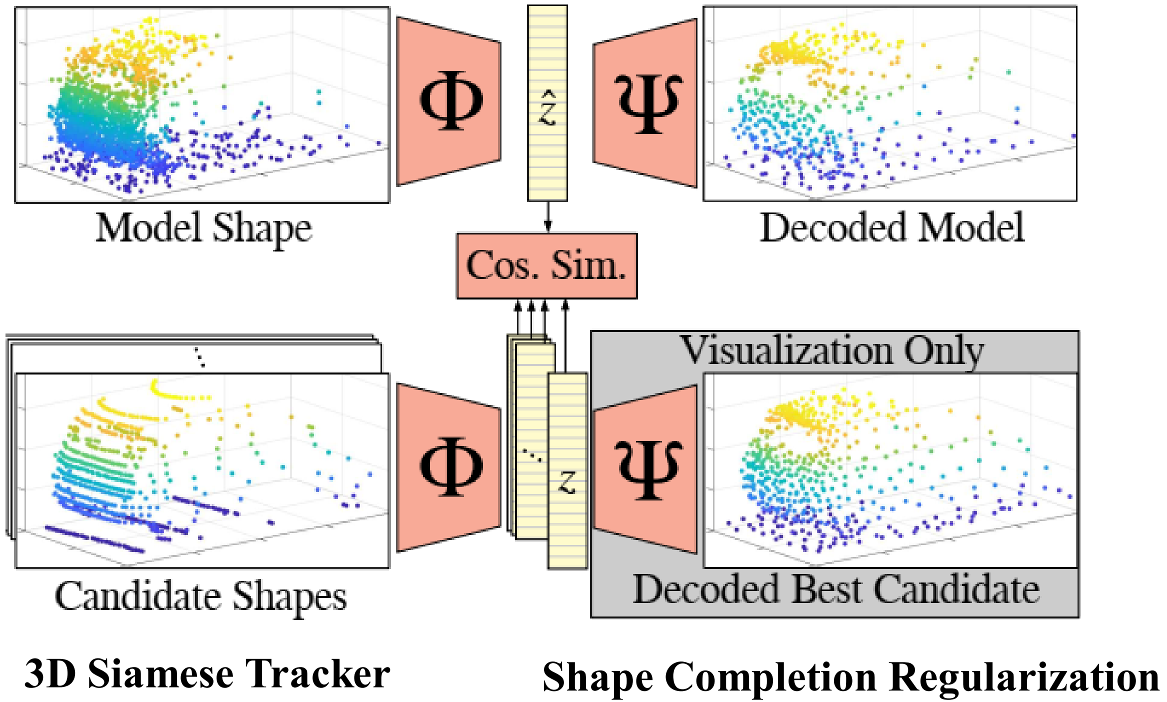
3D Single Object tracking



Leveraging Shape Completion for 3D Siamese Tracking

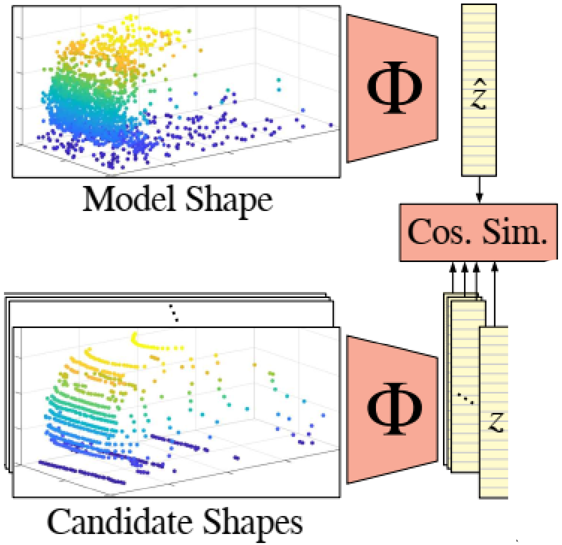
Silvio Giancola*, Jesus Zarzar*, and Bernard Ghanem
King Abdullah University of Science and Technology (KAUST), Saudi Arabia
{silvio.giancola,jesusalejandro.zarzartorano,bernard.ghanem}@kaust.edu.sa

Pipeline:



Model Shape
Φ
$\hat{z}$
Ψ
Decoded Model
Cos. Sim.
Visualization Only
Candidate Shapes
Φ
z
Ψ
Decoded Best Candidate
3D Siamese Tracker
Shape Completion Regularization

3D Siamese Tracker:



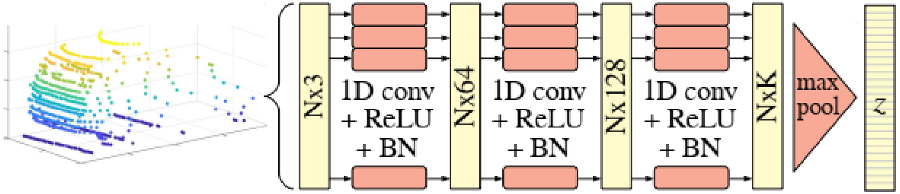


Figure 2. Our encoder takes as input a point cloud with $N = 2048$ points. Point clouds are encoded into a K-dimensional ($K = 128$) latent vector $\mathbf{z}$ using 3 layers of 1D CNN with ReLU and BN.

Similarity Metric:

$$CosSim(\mathbf{z}, \hat{\mathbf{z}}) = \frac{\mathbf{z}^\top \hat{\mathbf{z}}}{\|\mathbf{z}\|_2 \|\hat{\mathbf{z}}\|_2}$$

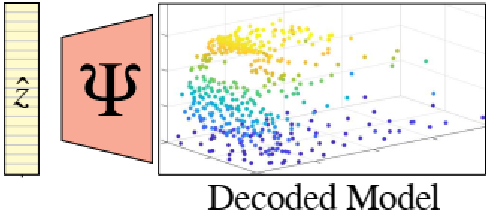
Tracking Loss:

$$\mathcal{L}_{tr} = \frac{1}{n} \sum_{\mathbf{x}} \Big(CosSim(\phi(\mathbf{x}), \phi(\hat{\mathbf{x}})) - \rho\big(d(\mathbf{x}, \hat{\mathbf{x}})\big) \Big)^2$$

$d(\cdot,\cdot)$: the L2-norm $\|\cdot\|^2$ of the difference

$\rho(\cdot)$: Gaussian function with $\mu = 0,\ \delta = 1$

Shape Completion Regularization:



Completion Loss:

$$\mathcal{L}_{comp} = \sum_{\hat{\mathbf{x}}_i \in \hat{\mathbf{x}}} \min_{\tilde{\mathbf{x}}_j \in \tilde{\mathbf{x}}} \|\hat{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|_2^2 + \sum_{\tilde{\mathbf{x}}_j \in \tilde{\mathbf{x}}} \min_{\hat{\mathbf{x}}_i \in \hat{\mathbf{x}}} \|\hat{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|_2^2$$

$\psi(\cdot)$: Two fully connected layers

Input: N x K (2048 x 128)

Output: M x 3 (2048 x 3)

The tracking loss enforces encoded partial shapes to be similar to their respective encoded model

The completion loss enforces the encoded model to hold semantic information to enable its decoding. Thus, this regularization is used to enforce the latent space learned by the Siamese network to hold meaningful shape semantic information.

Experiments:

Kitti:

Table 1. Ablation study for different losses we are training with. We report the OPE Success/Precision metrics for different losses averaged over 5 runs. Best results shown in bold.

Ablation	Success	Precision
(i) Before Training (Random)	39.06	41.79
(ii) Pre-trained on ShapeNet	44.54	49.38
(iii) Ours – Completion only	65.36	70.62
(iv) Ours – Tracking only	73.96	78.68
(v) Ours – λ_{comp}@$1e^{-6}$	**76.94**	**81.38**

	Method	Previous result	Previous GT	Current GT
Success	SC3D [11]	41.3	64.6	76.9
	P2B (ours)	**56.2**	**82.4**	**84.0**
Precision	SC3D [11]	57.9	74.5	81.3
	P2B (ours)	**72.8**	**90.1**	**90.3**

Table 2. **Comprehensive comparison with SC3D.** The right three columns differ in their ways to generate search area.

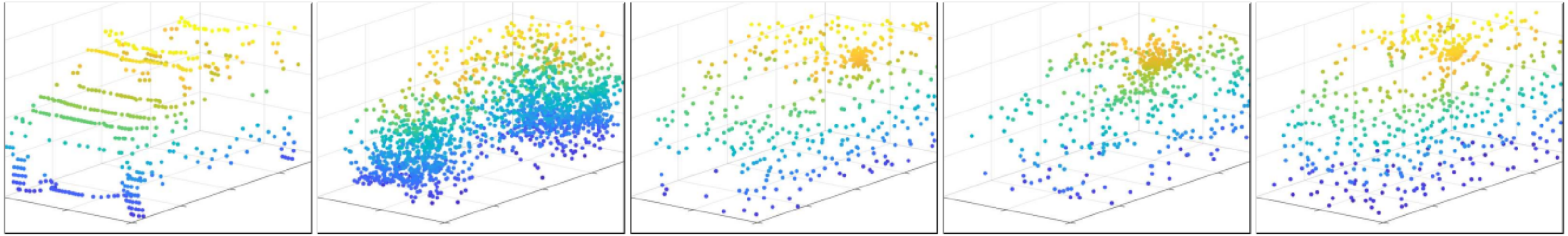


Figure 4. Example of model completion (from left to right): **(i)** Candidate point cloud, **(ii)** Decoded candidate point cloud when it is pre-trained with ShapeNet, **(iii)** Decoded candidate point cloud when it is trained with completion loss only ($\lambda_{comp} = \infty$), **(iv)** Decoded candidate point cloud when it is trained with tracking loss only ($\lambda_{comp} = 0$) (the decoder trained for completion is used for fair comparison), **(v)** Decoded candidate point cloud when it is trained with both tracking and completion losses ($\lambda_{comp} = 1e^{-6}$).

Ablation Study:

$$\mathcal{L} = \mathcal{L}_{tr} + \lambda_{comp}\mathcal{L}_{comp}$$

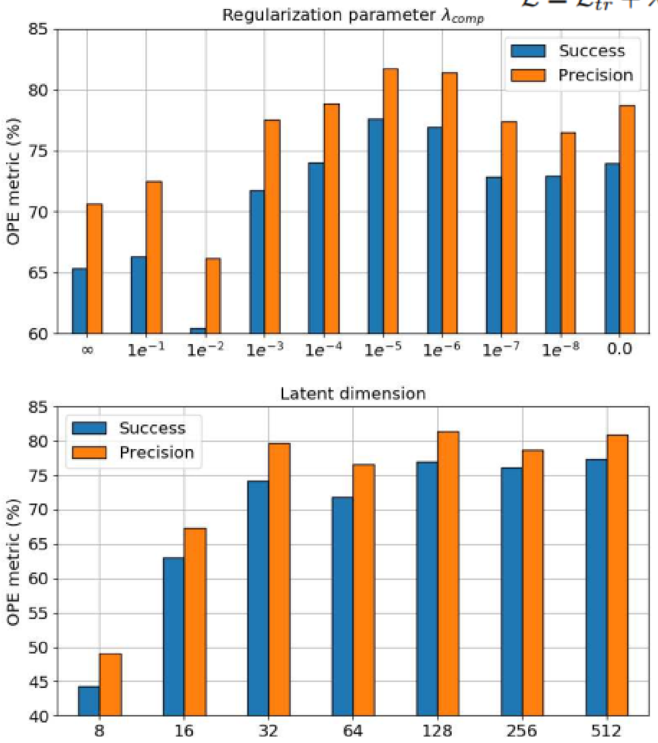


Figure 3. Ablation study for different regularization λ_{comp} of the shape completion (*top*) and for the latent representation size K (*bottom*). We report the OPE Success/Precision metrics for different values of λ_{comp} and K averaged over 5 runs.

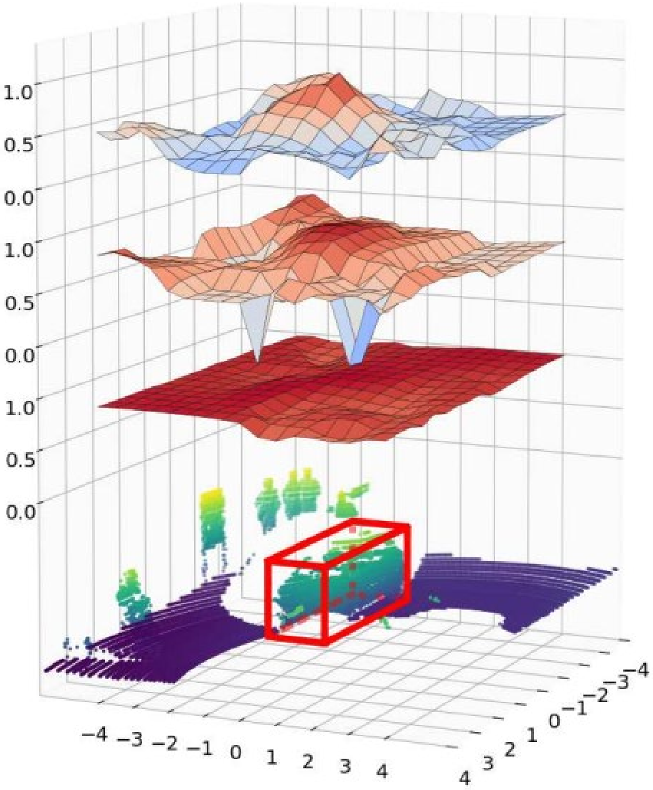


Figure 5. Heatmap of model cosine similarity scores on an exhaustive search space grid: From bottom to top: **(i)** activation using random weights model, **(ii)** activation on pre-trained model (ShapeNet), **(iii)** our model.

Thanks