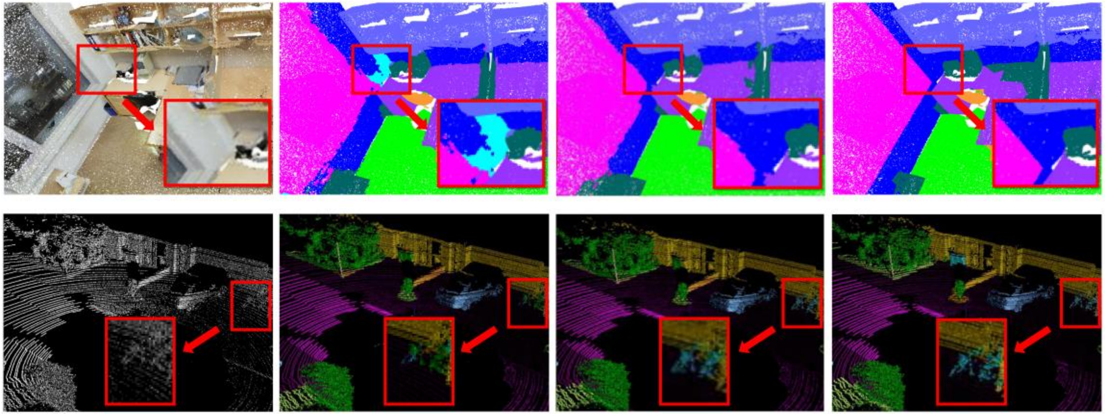
Semantic Segmentation for Real Point Cloud Scenes via Bilateral Augmentation and Adaptive Fusion

Shi Qiu, Saeed Anwar, Nick Barnes

{shi.qiu, saeed.anwar, nick.barnes} @anu.edu.au

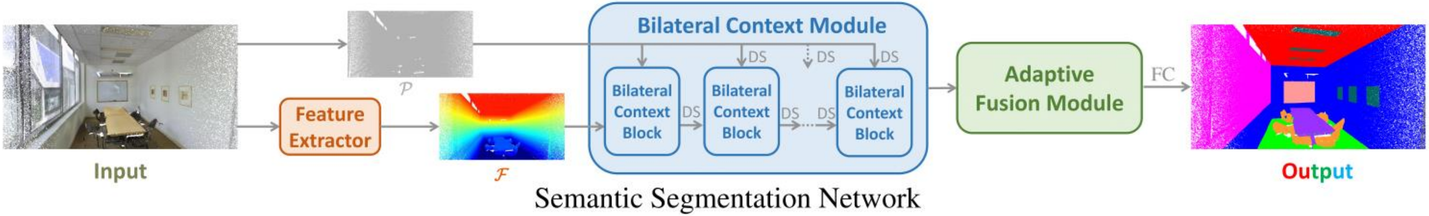
Motivation


Input RandLA-Net Ours Ground-truth

Issues in SOTA methods:

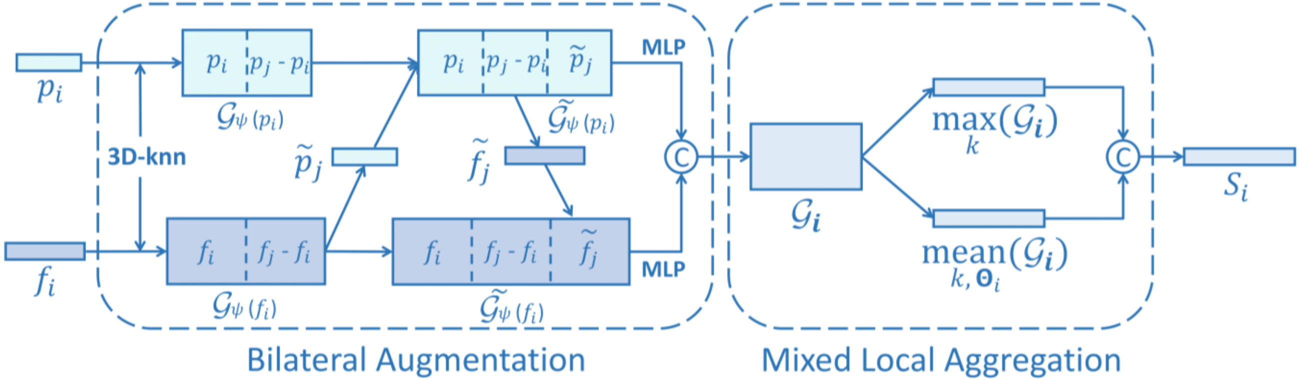
- ambiguity in close points
- redundant features
- inadequate global representations

Network Overview



- Geometric Context: $\mathcal{P} \in \mathbb{R}^{N\times3}$
- Semantic Context: $\mathcal{F} \in \mathbb{R}^{N\times C}$
- DS: Down Sampling
- FC: Fully Connected Layers

Bilateral Context Block



- Bilateral Augmentation:

learning 3-DoF offsets from semantic features to enhance local geometric context:

$$\tilde{p_j} = \mathcal{M}(\mathcal{G}_\psi(f_i)) + p_j, \quad \tilde{p_j} \in \mathbb{R}^3.$$

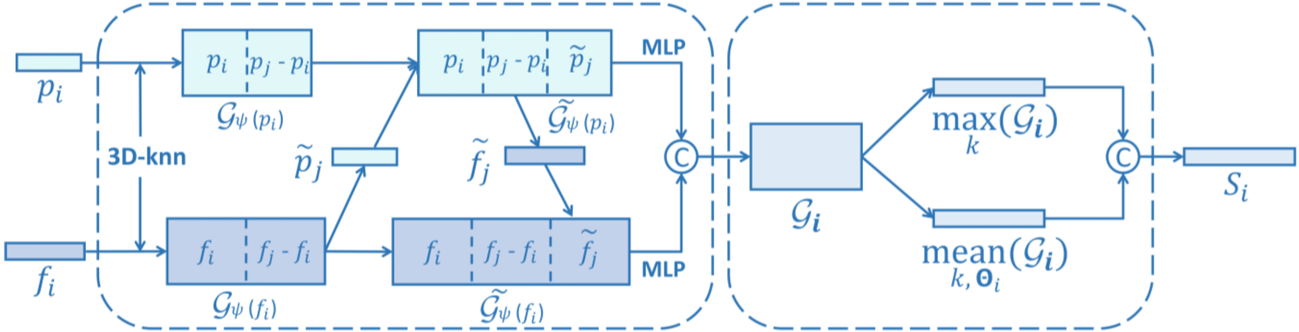
$$\tilde{\mathcal{G}_\psi}(p_i) = [p_i; p_j - p_i; \tilde{p_j}];$$

learning *d*-DoF offsets from geometric features to enhance local semantic context:

$$\tilde{f_j} = \mathcal{M}(\tilde{\mathcal{G}_\psi}(p_i)) + f_j, \quad \tilde{f_j} \in \mathbb{R}^d;$$

$$\tilde{\mathcal{G}_\psi}(f_i) = [f_i; f_j - f_i; \tilde{f_j}]$$

Bilateral Context Block



- Augmentation Loss:

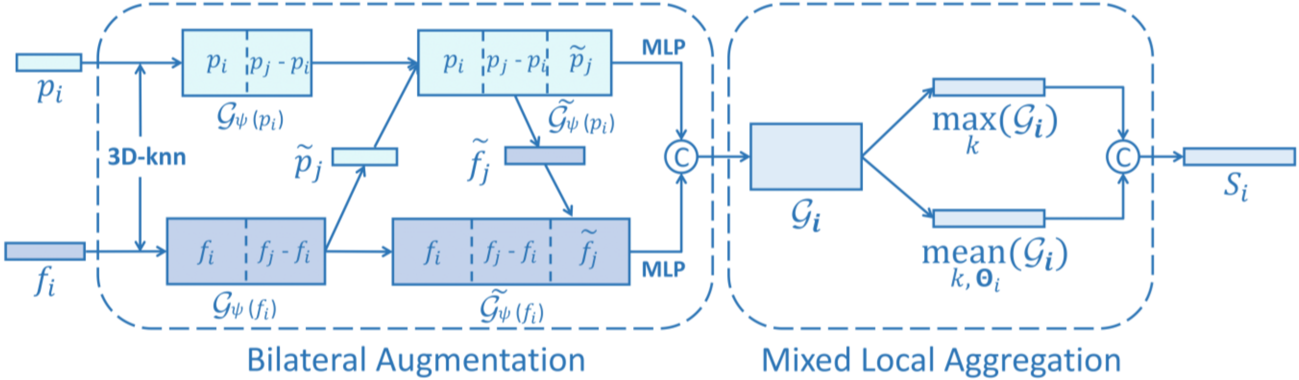
For a point p_i and its k shifted neighbors $\tilde{p_j}$:

$$\mathcal{L}(p_i) = \left\| \frac{1}{k} \sum_{j=1}^{k} \tilde{p_j} - p_i \right\|_2. \qquad \mathcal{L}_m = \sum_{i=1}^{N_m} \mathcal{L}(p_i).$$

For the network:

$$\mathcal{L}_{all} = \mathcal{L}_{CE} + \sum_{m=1}^{M} \omega_m \cdot \mathcal{L}_m$$

Bilateral Context Block



- Mixed Local Aggregation:

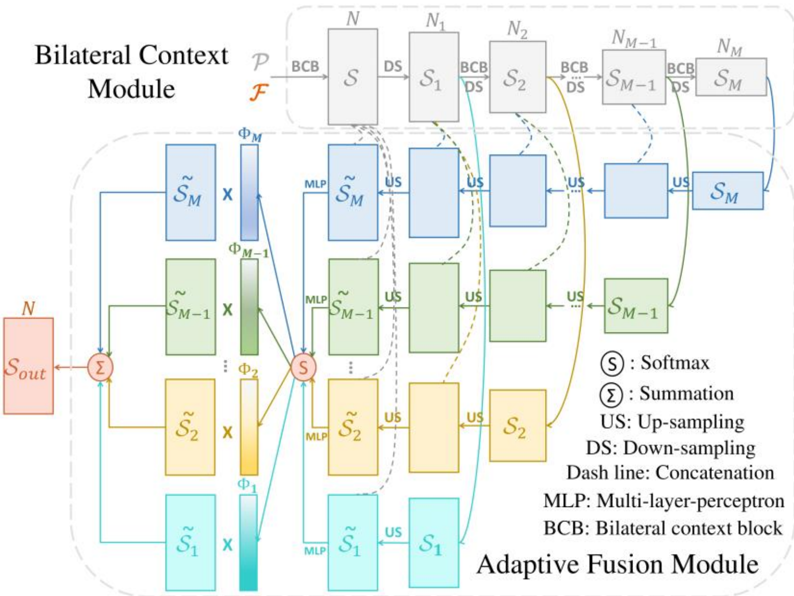
Augmented Local Context:

$$\mathcal{G}_i = \operatorname{concat}\Big(\mathcal{M}(\tilde{\mathcal{G}}_\psi(p_i)), \mathcal{M}(\tilde{\mathcal{G}}_\psi(f_i))\Big) \in \mathbb{R}^{k \times d'}.$$

Output point feature:

$$s_i = \operatorname{concat}\Big(\max_k(\mathcal{G}_i), \operatorname*{mean}_{k,\Theta_i}(\mathcal{G}_i)\Big), \quad s_i \in \mathbb{R}^{2d'}$$

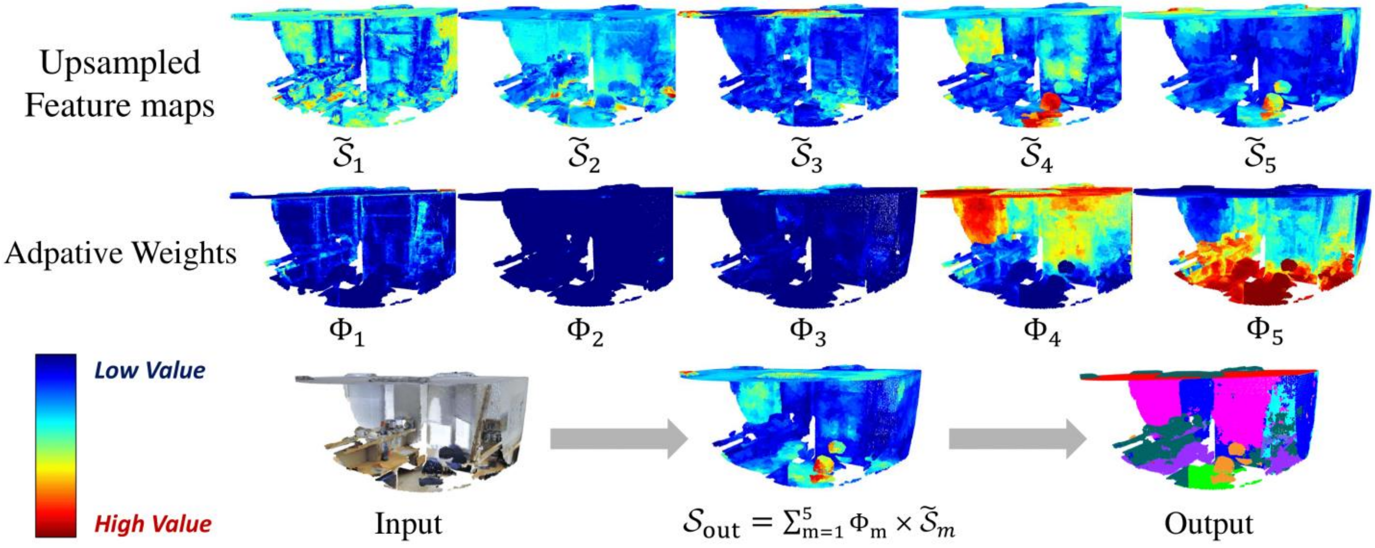
Adaptive Fusion



Algorithm 1: Adaptive Fusion Module Pipeline

input: M multi-resolution feature maps $\{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_M\}$.
output: $\mathcal{S}_{out}$ for semantic segmentation.
for $\mathcal{S}_m \in \{\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_M\}$ **do**
  upsample: $\tilde{\mathcal{S}_m} \leftarrow \mathcal{S}_m$;
  summarize: $\phi_m \leftarrow \tilde{\mathcal{S}_m}$;
end for
obtain: $\forall \tilde{\mathcal{S}_m} \in \{\tilde{\mathcal{S}_1}, \tilde{\mathcal{S}_2}, ..., \tilde{\mathcal{S}_M}\}$, $\tilde{\mathcal{S}_m} \in \mathbb{R}^{N \times c}$; and $\forall \phi_m \in \{\phi_1, \phi_2, ..., \phi_M\}$, $\phi_m \in \mathbb{R}^N$.
regress: $\{\Phi_1, \Phi_2, ..., \Phi_M\} \leftarrow \{\phi_1, \phi_2, ..., \phi_M\}$, where $\Phi_m \in \mathbb{R}^N$.
return:
$\mathcal{S}_{out} = \sum_{m=1}^{M} \Phi_m \times \tilde{\mathcal{S}_m}$.

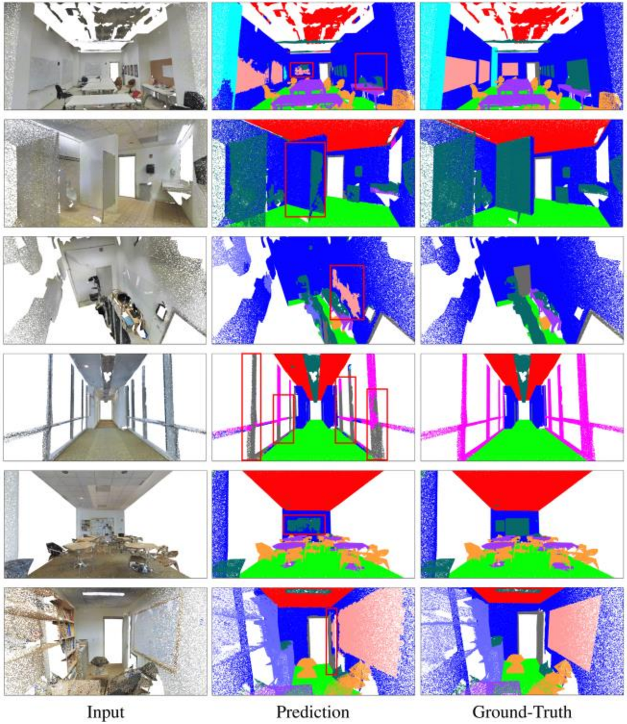
Adaptive Fusion



S3DIS

Table 1: Semantic segmentation (6-fold cross-validation) results (%) on the *S3DIS* dataset [2]. (**mAcc**: average class accuracy, **OA**: overall accuracy, **mIoU**: mean Intersection-over-Union. "-" indicates unknown result.)

year	Method	mAcc	OA	mIoU
2017	PointNet [37]	66.2	78.6	47.6
	PointNet++ [38]	67.1	81.0	54.5
2018	A-SCN [50]	-	81.6	52.7
	PointCNN [30]	75.6	88.1	65.4
	SPG [26]	73.0	85.5	62.1
2019	DGCNN [46]	-	84.1	56.1
	KP-Conv [44]	79.1	-	70.6
	ShellNet [54]	-	87.1	66.8
	PointWeb [55]	76.2	87.3	66.7
	SSP+SPG [25]	78.3	87.9	68.4
2020	Seg-GCN [28]	77.1	87.8	68.5
	PointASNL [51]	79.0	88.8	68.7
	RandLA-Net [19]	82.0	88.0	70.0
	MPNet [17]	-	86.8	61.3
	InsSem-SP [32]	74.3	88.5	64.1
	Ours	**83.1**	**88.9**	**72.2**



Semantic3D and SemanticKITTI

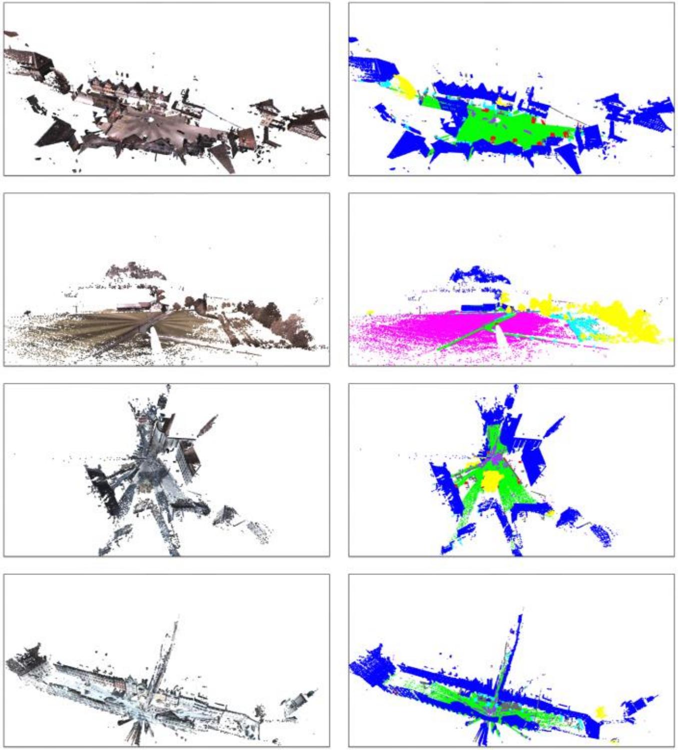
Table 2: Semantic segmentation (semantic-8) results (%) on the *Semantic3D* dataset [14].

Method	OA	mIoU	man-made terrain	natural terrain	high vegetation	low vegetation	buildings	hard scape	scanning artefacts	cars
TMLC-MS [15]	85.0	49.4	91.1	69.5	32.8	21.6	87.6	25.9	11.3	55.3
EdgeConv-PN [9]	89.4	61.0	91.2	69.8	51.4	58.5	90.6	33.0	24.9	68.6
PointNet++ [38]	85.7	63.1	81.9	78.1	64.3	51.7	75.9	36.4	43.7	72.6
SnapNet [6]	91.0	67.4	89.6	79.5	74.8	56.1	90.9	36.5	34.3	77.2
PointConv [48]	91.8	69.2	92.2	79.2	73.1	62.7	92.0	28.7	43.1	82.3
PointGCR [36]	92.1	69.5	93.8	80.0	64.4	66.4	93.2	39.2	34.3	85.3
PointConv-CE [31]	92.3	71.0	92.4	79.6	72.7	62.0	93.7	40.6	44.6	82.5
RandLA-Net [19]	94.2	71.8	96.0	88.6	65.3	62.0	**95.9**	49.8	27.8	89.3
SPG [26]	92.9	**76.2**	91.5	75.6	**78.3**	**71.7**	94.4	**56.8**	**52.9**	88.4
Ours	**94.9**	75.4	**97.9**	**95.0**	70.6	63.1	94.2	41.6	50.2	**90.3**

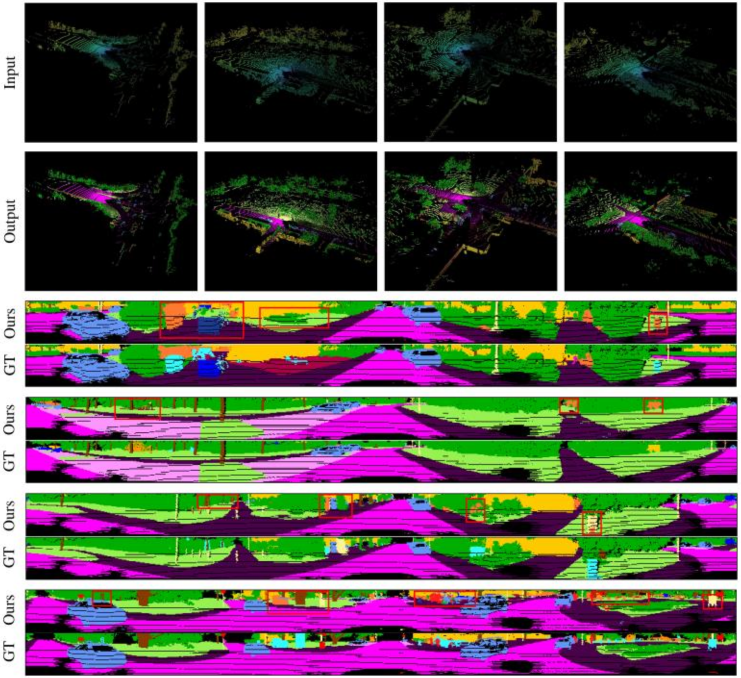
Table 3: Semantic segmentation (single-scan) results (%) on the *SemanticKITTI* dataset [4].

Method	mIoU	road	sidewalk	parking	other-ground	building	car	truck	bicycle	motorcycle	other-vehicle	vegetation	trunk	terrain	person	bicyclist	motorcyclist	fence	pole	traffic-sign
PointNet [37]	14.6	61.6	35.7	15.8	1.4	41.4	46.3	0.1	1.3	0.3	0.8	31.0	4.6	17.6	0.2	0.2	0.0	12.9	2.4	3.7
PointNet++ [38]	20.1	72.0	41.8	18.7	5.6	62.3	53.7	0.9	1.9	0.2	0.2	46.5	13.8	30.0	0.9	1.0	0.0	16.9	6.0	8.9
SquSegV2 [47]	39.7	88.6	67.6	45.8	17.7	73.7	81.8	13.4	18.5	17.9	14.0	71.8	35.8	60.2	20.1	25.1	3.9	41.1	20.2	36.3
TangentConv [43]	40.9	83.9	63.9	33.4	15.4	83.4	90.8	15.2	2.7	16.5	12.1	79.5	49.3	58.1	23.0	28.4	8.1	49.0	35.8	28.5
PointASNL [51]	46.8	87.4	74.3	24.3	1.8	83.1	87.9	39.0	0.0	25.1	29.2	84.1	52.2	**70.6**	34.2	**57.6**	0.0	43.9	57.8	36.9
RandLA-Net [19]	53.9	90.7	73.7	60.3	20.4	86.9	94.2	40.1	26.0	25.8	38.9	81.4	61.3	66.8	49.2	48.2	7.2	56.3	49.2	47.7
PolarNet [53]	54.3	90.8	74.4	61.7	21.7	90.0	93.8	22.9	40.3	30.1	28.5	84.0	65.5	67.8	43.2	40.2	5.6	67.8	51.8	57.5
MinkNet42 [8]	54.3	91.1	69.7	63.8	29.3	**92.7**	94.3	26.1	23.1	26.2	36.7	83.7	68.4	64.7	43.1	36.4	7.9	57.1	57.3	60.1
FusionNet [52]	**61.3**	**91.8**	**77.1**	**68.8**	**30.8**	92.5	95.3	41.8	**47.5**	**37.7**	34.5	**84.5**	**69.8**	68.5	**59.5**	56.8	11.9	**69.4**	**60.4**	**66.5**
Ours	59.9	90.9	74.4	62.2	23.6	89.8	**95.4**	**48.7**	31.8	35.5	**46.7**	82.7	63.4	67.9	49.5	55.7	**53.0**	60.8	53.7	52.0

Semantic3D and SemanticKITTI



Input Prediction



Conclusions

- Augmenting the local context bilaterally.
- Fusing multi-resolution features for each point adaptively.
- Evaluating the network on various point cloud benchmarks.
- Expecting different frameworks and downstream tasks in the future.

Contact the authors

Visit the project page