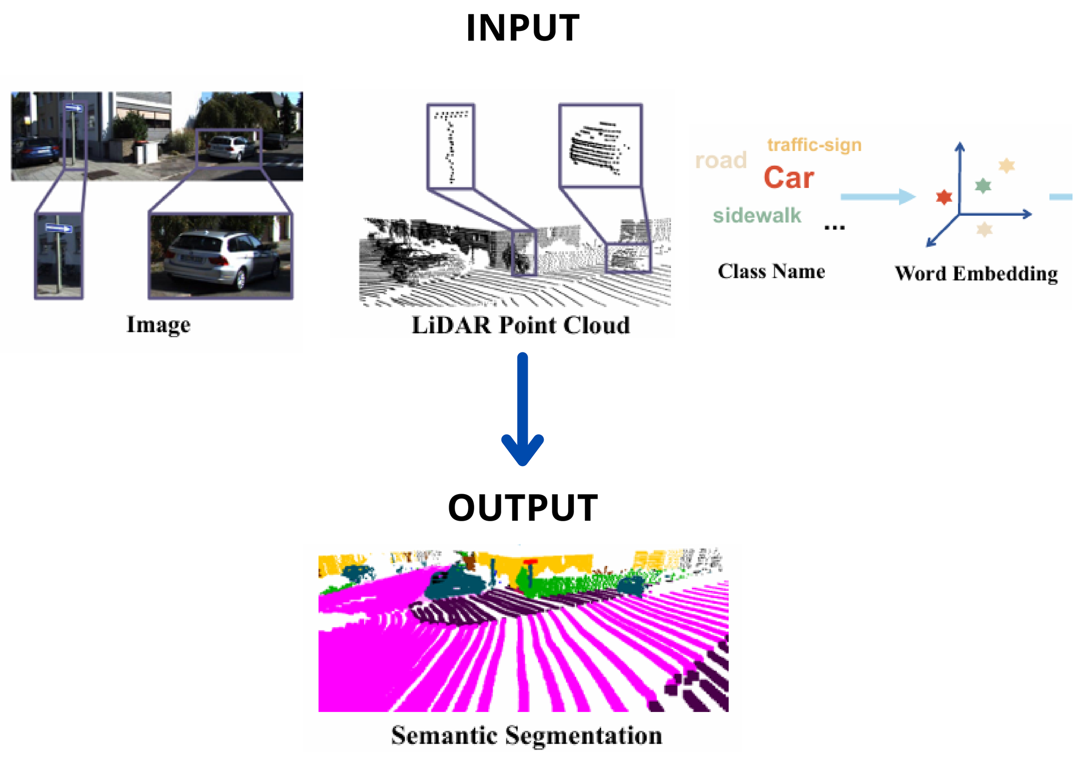
See More and Know More:
Zero-shot Point Cloud Segmentation via Multi-modal Visual Data

Yuhang Lu, Qi Jiang, Runnan Chen, Yuenan Hou, Xinge Zhu, Yuexin Ma

GOAL

Point cloud semantic segmentation over:
seen & unseen objects

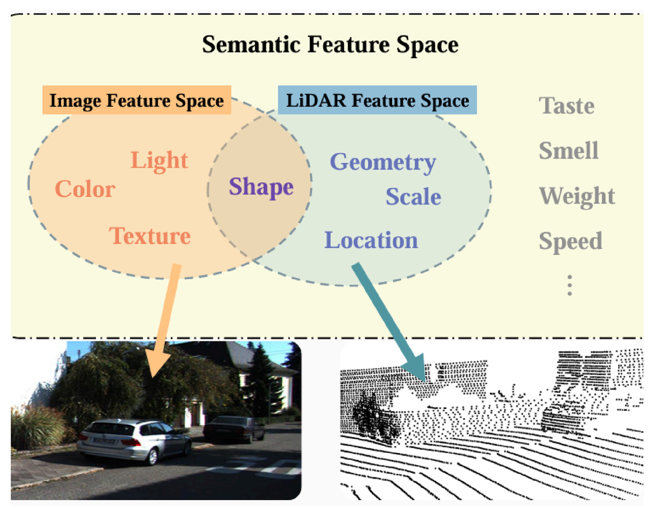


INTRODUCTION

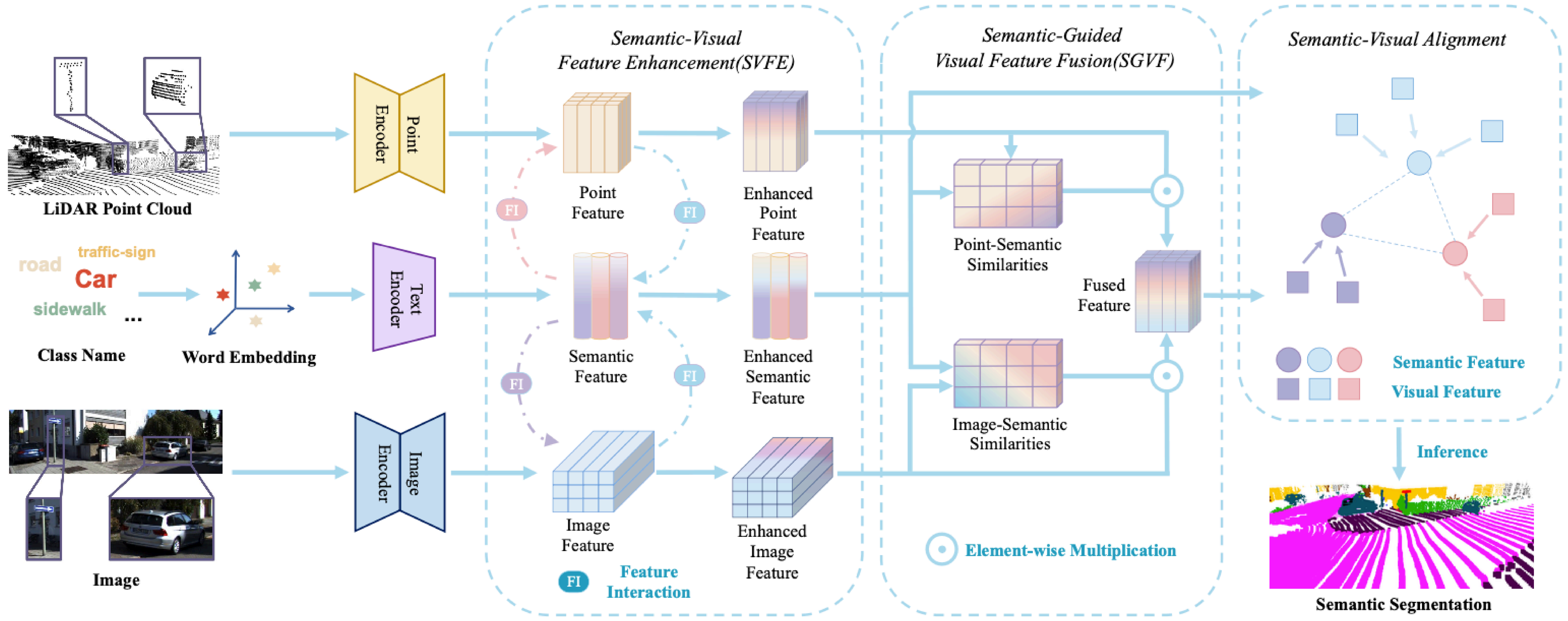
- Point cloud **semantic segmentation** over: seen & unseen objects
- **Train a model** to recognize unseen objects
- Useful for **autonomous driving**

CHALLENGES

- **Hard to generalize** over unlabelled training set of 3D data
- **Manual labelling** is infeasible
- **Few** 3D semantic segmentation models
- Generative methods
  - fake features generator (3DGenZ)
- Projection-based methods alignment (TGF)



MODEL



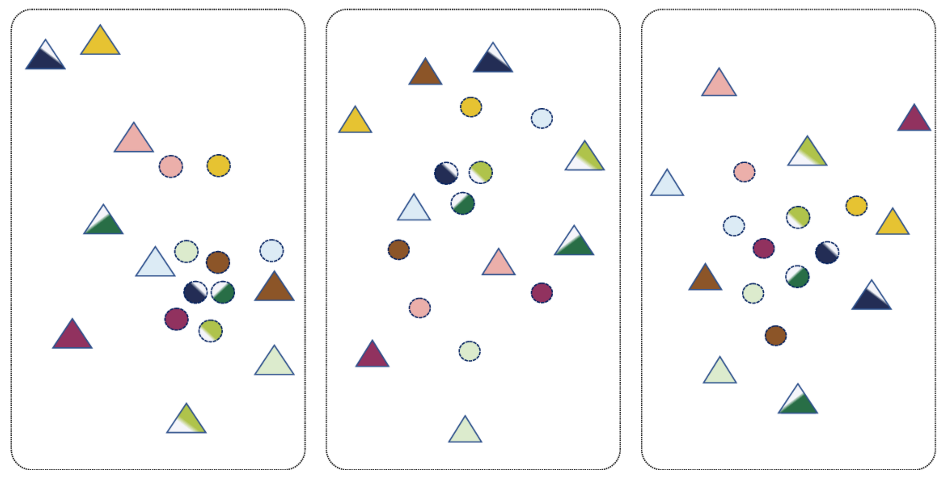
Visual-Semantic Feature Extraction

- **Cylinder3D**, to extract 3D features from point cloud
- **ResUnet**, to extract 2D features from image
- **W2V**, **Glove**, followed by **MLP** to map the word embedding to its semantic feature

Semantic-Guided Visual Feature Fusion

Combine effectively visual features under the guide of the semantic features

1. **Weight matrices**, multihead-attention of enhanced visual features and enhanced semantic features
2. **Element-wise multiplication** between weight matrix and enhanced visual feature
3. **MLP with Softmax**



(a) Before SVFE (b) After SVFE (c) After SGVF

Semantic-Visual Feature Enhancement

Transformer Decoder with cross-attention layer to decrease the gap between visual and semantic features

1. **Semantic Feature Enhancement**, semantic features as query and visual features as key and value
2. **Visual Feature Enhancement**, visual features as query and semantic features as key and value

Semantic-Visual Alignment

Align semantic and visual feature spaces under seen classes supervision

- Loss function for seen classes, to have compact distribution within classes and distinguishable one between classes
- Loss function for unseen classes, to push unseen classes' features apart to seen ones (seen bias)

Inference

The class of each point is determined by the similarity between its fused visual feature and semantic features of all classes (seen and unseen)

EXPERIMENTS

Datasets:

1. SemanticKITTI
2. nuScenes

Evaluation metrics:

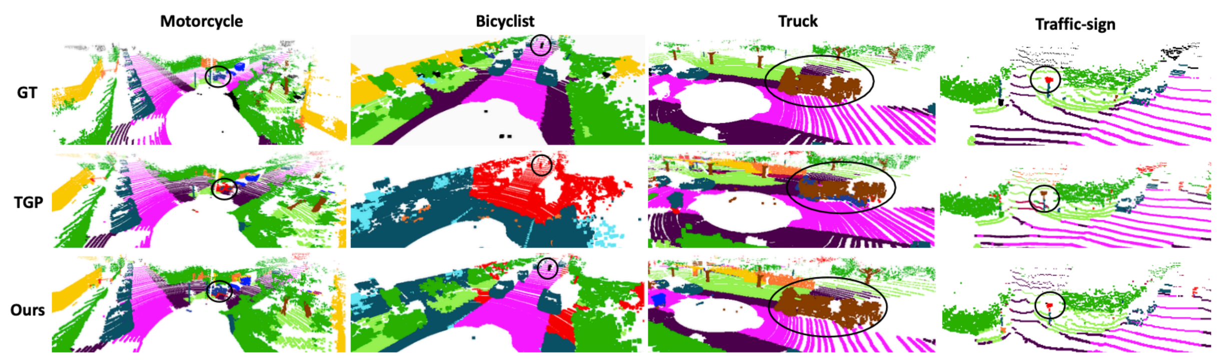
- harmonic mean IoU

$$\mathrm{hIoU} = \frac{2 \times \mathrm{mIoU}_{seen} \times \mathrm{mIoU}_{unseen}}{\mathrm{mIoU}_{seen} + \mathrm{mIoU}_{unseen}}$$

RESULTS

Our model outperforms SOTA with **more than 50% improvement** of unseen category mIOU.
Furthermore, it yields real-time performance (0.097 s/f)

Setting	Model	SemanticKITTI Seen mIoU	SemanticKITTI Uneen mIoU	SemanticKITTI Improvement	SemanticKITTI Overall mIoU	SemanticKITTI Overall hIoU	nuScenes Seen mIoU	nuScenes Uneen mIoU	nuScenes Improvement	nuScenes Overall mIoU	nuScenes Overall hIoU
0	TGP[15]	-	-	-	59.1	-	-	-	-	67.9	-
	Ours	-	-	-	**62.6**	-	-	-	-	**69.1**	-
2	3DGenZ[45]	40.9	12.4	-	37.9	19.0	**67.8**	4.2	-	**59.9**	7.9
	TGP[15]	58.3	28.8	+3.5%	55.2	38.6	58.9	26.9	+25.3%	54.9	36.9
	Ours	**59.5**	**29.8**	-	**56.4**	**39.7**	59.4	**33.7**	-	56.2	**43.0**
	Supervised	61.5	71.8	-	62.6	66.3	70.1	61.9	-	69.1	65.7
4	3DGenZ[45]	41.4	10.8	-	35.0	17.1	**67.2**	3.1	-	51.2	5.9
	TGP[15]	54.6	17.3	+54.9%	46.7	26.3	65.7	14.8	+56.1%	53.0	24.2
	Ours	**58.8**	**26.8**	-	**52.1**	**36.8**	66.4	**23.1**	-	**55.6**	**34.3**
	Supervised	60.3	71.2	-	62.6	65.3	71.9	60.6	-	69.1	65.8
6	3DGenZ[45]	40.3	6.5	-	29.6	11.2	53.8	3.2	-	34.8	6.0
	TGP[15]	53.6	13.3	+79.7%	40.9	21.3	**68.8**	14.1	+56.7%	48.3	23.4
	Ours	**56.6**	**23.9**	-	**46.3**	**33.6**	66.8	**22.1**	-	**50.0**	**33.2**
	Supervised	56.8	75.3	-	62.6	64.8	74.5	60.1	-	69.1	66.5
8	3DGenZ[45]	38.3	1.3	-	22.7	2.5	36.5	2.1	-	19.3	4.0
	TGP[15]	**53.2**	8.6	+70.9%	**34.4**	14.8	**68.4**	13.7	+56.9%	41.1	22.8
	Ours	46.0	**14.7**	-	32.8	**22.3**	68.2	**21.5**	-	**44.9**	**32.7**
	Supervised	52.1	77.1	-	62.6	62.2	73.5	64.7	-	69.1	68.8



ABLATION

Model	Seen mIoU	Unseen mIoU	Overall mIoU	Overall hIoU
Ours	58.8	**26.8**	**52.1**	**36.8**
Ours w/o SGVF	58.8	23.4	51.3	33.5
Ours w/o SVFE	**59.0**	19.9	50.8	29.8
Ours w/o Image	58.3	20.0	50.2	29.8