AAAI 2024

VLM2Scene: Self-Supervised Image-Text-LiDAR Learning with Foundation Models for Autonomous Driving Scene Understanding

Guibiao Liao, Jiankun Li, Xiaoqing Ye

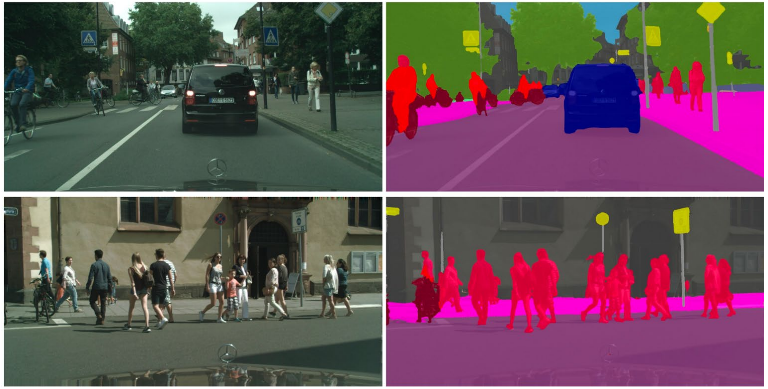
- Problem/Objective
  - Self-supervised 3D representation for Scene understanding

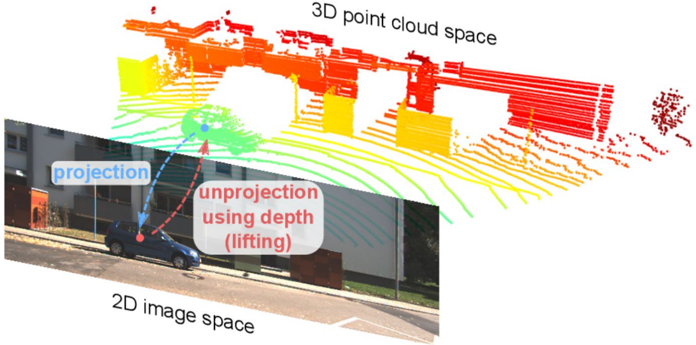
- Contribution/Key Idea
  - 3D에서의 Vision / Language foundation models(VLMs)을 동시 사용
  - Challenge of LiDAR(sparse&noise)의 극복 가능성을 보여줌 by Image-text

김범준

- **Scene Understanding**

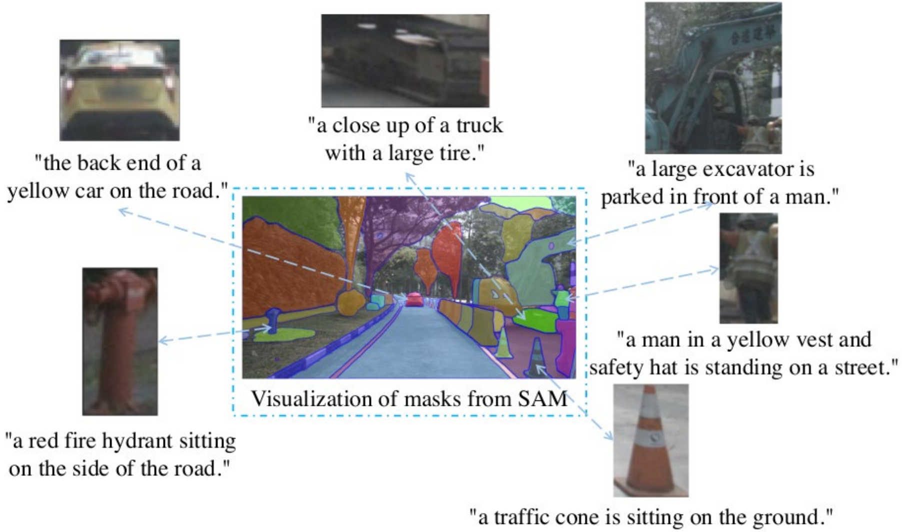
- Image understanding(ex. CLIP)의 3D 버전
  - Semantic 정보를 이용한
    - 3D Object detection
    - **3D semantic segmentation**
    - etc
  - 주되게는 LiDAR only / LiDAR + Camera
    - LiDAR의 noise, sparse → Challenge





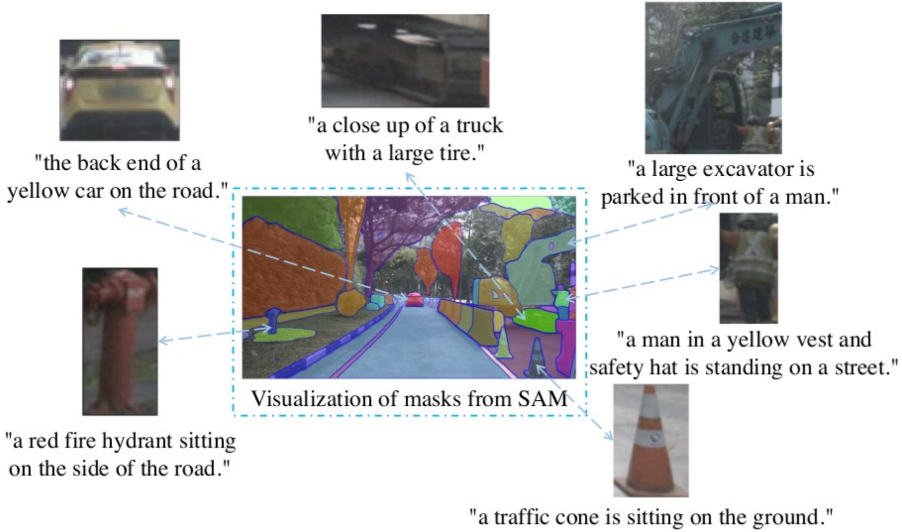
- # **Introduction**

- This Paper overcome the limitation of
    - Cost of 3D annotation
    - Difficulty of transferring 2D to 3D
    - VLMs that are only used for 2D
    - Potential noise and sparse of points
    - Lack of realistic and detailed description

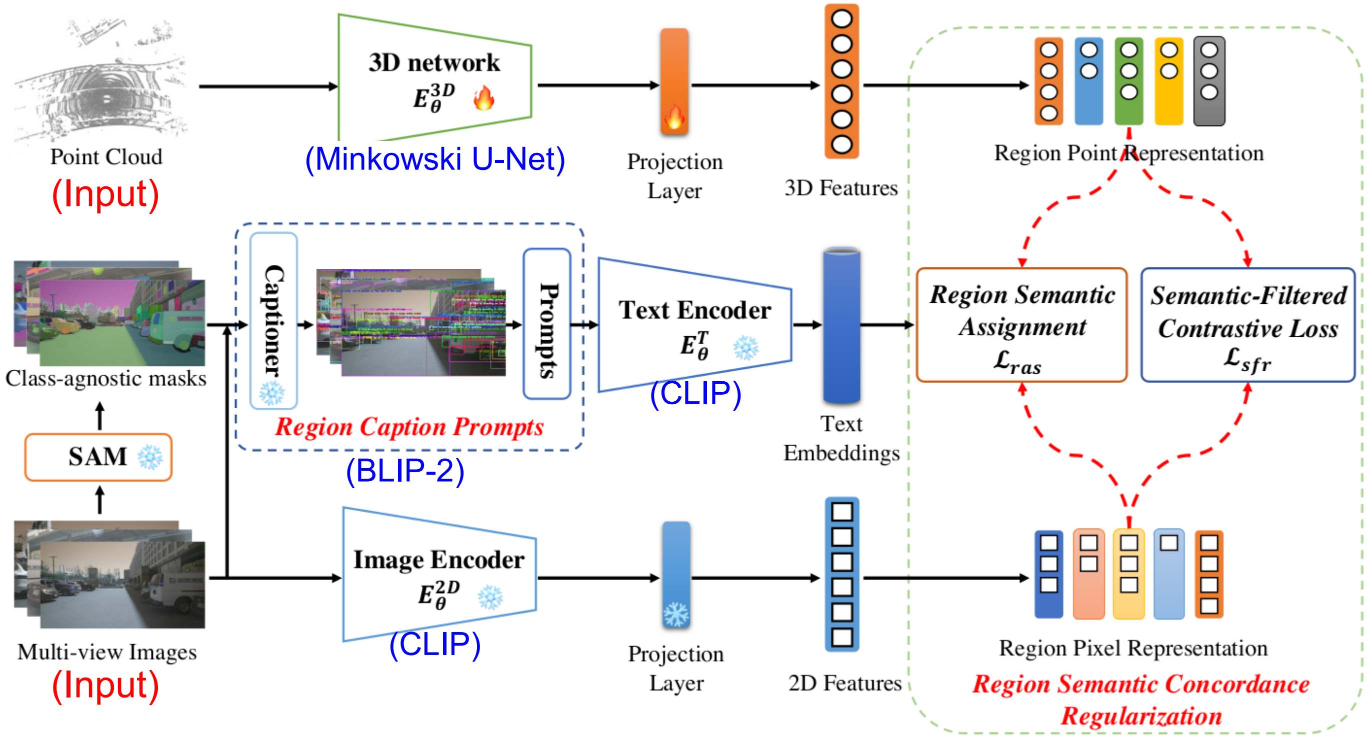


- # Introduction

- This Paper overcome the limitation of
  - Cost of 3D annotation
    → **Self-supervised method**
  - Difficulty of transferring 2D to 3D
  - VLMs that are only used for 2D
    → **VLMs(CLIP, BLIP-2, SAM) 정보를 통합하여 3D에 적용**
  - Potential noise and sparse of points
    → **Semantic-Filter Region(SFR) 도입**
  - Lack of realistic and detailed description
    → **Region Caption Prompts(RCP) 도입**



- # Method - VLM2Scene

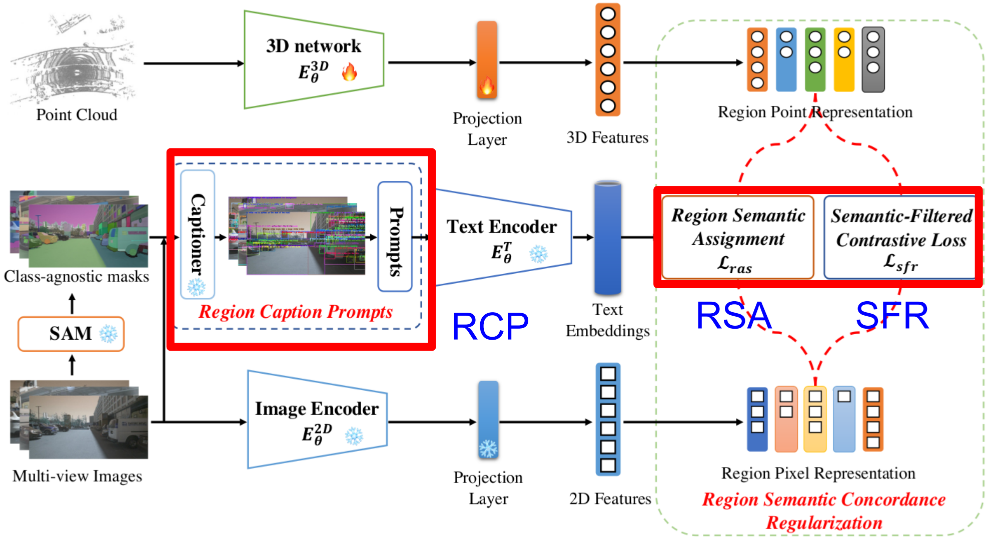


- **Method - Details**

- Region Caption Prompts(RCP)
  - 이미지 내의 각 영역의 구체적인 텍스트 설명을 생성(위치, 관계, 색상 속성 등)

- Region Semantic Assignment(RSA)
  - 이미지, 포인트 내의 영역에 가장 적합한 카테고리를 할당

- Semantic-Filtered Region Contrastive Loss(SFR)
  - Contrastive learning을 통한 false positive 필터링



- **Method - Details**

- Region Caption Prompts(RCP)
  - 이미지 내의 각 영역의 구체적인 텍스트 설명을 생성 (위치, 관계, 색상 속성 등)
- Region Semantic Assignment(RSA)
  - 이미지, 포인트 내의 영역에 가장 적합한 카테고리를 할당
- Semantic-Filtered Region Contrastive Loss(SFR)
  - Contrastive learning을 통한 false positive 샘플 필터링

원본 이미지가 아닌 SAM을 통과한 이미지를 사용하여 general이 아닌 specific semantic 정보에 대한 prompts를 추출하도록 Pretrain

Limitation of <CLIP2Scene>

First, the aforementioned textual embeddings are straightforwardly generated from the category names **(e.g., car, pedestrian, etc)** of the dataset with fixed and short templates **(e.g., a photo of a {}, etc)**

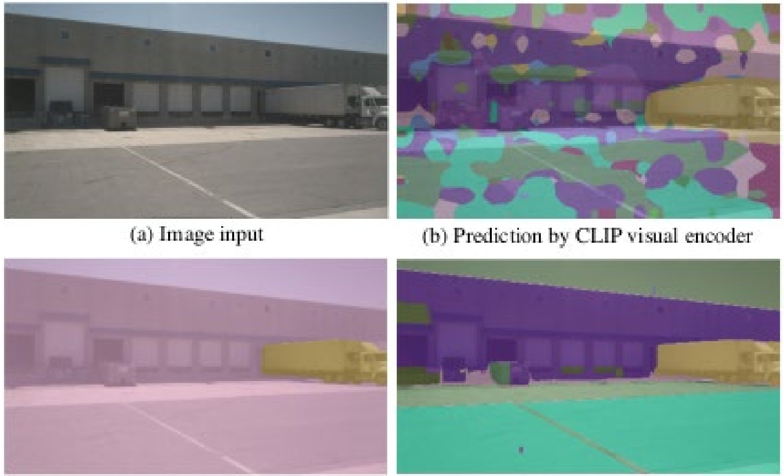
- Lack of realistic and detailed description

-> Region Caption Prompts(RCP) 도입 김범준

- **Method - Details**

- Region Caption Prompts(RCP)
  - 이미지 내의 각 영역의 구체적인 텍스트 설명을 생성 (위치, 관계, 색상 속성 등)
- Region Semantic Assignment(RSA)
  - 이미지, 포인트 내의 영역에 가장 적합한 카테고리를 할당
- Semantic-Filtered Region Contrastive Loss(SFR)
  - Contrastive learning을 통한 false positive 샘플 필터링

→ SAM 마스크 내부 픽셀 중 가장 많은 class를 채택
→ SAM 마스크와 point를 align 하여 가장 많은 class를 채택

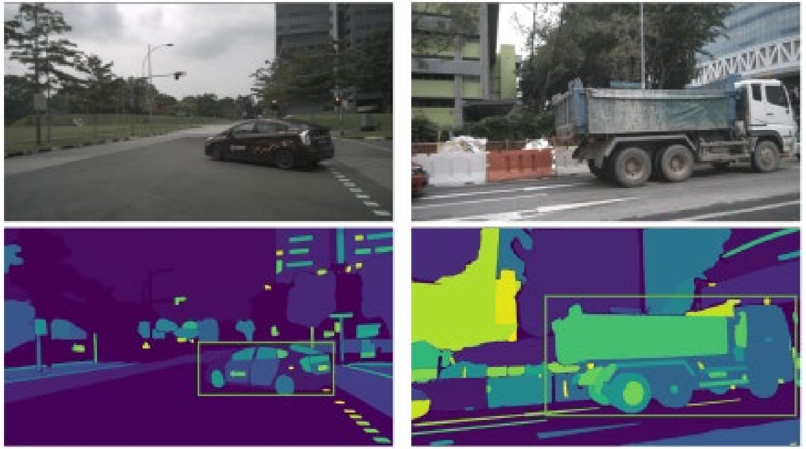


(a) Image input
(b) Prediction by CLIP visual encoder
(c) Examples of region prediction
(d) Prediction by our method with RSA

CLIP: Lack of precise edges by pixel-level
vs
RSA: Semantic consistency by region-level

- **Method - Details**

- Region Caption Prompts(RCP)
  - 이미지 내의 각 영역의 구체적인 텍스트 설명을 생성 (위치, 관계, 색상 속성 등)
- Region Semantic Assignment(RSA)
  - 이미지 내의 영역에 가장 적합한 카테고리를 할당
- Semantic-Filtered Region Contrastive Loss(SFR)
  - Contrastive learning을 통한 false positive 샘플 필터링



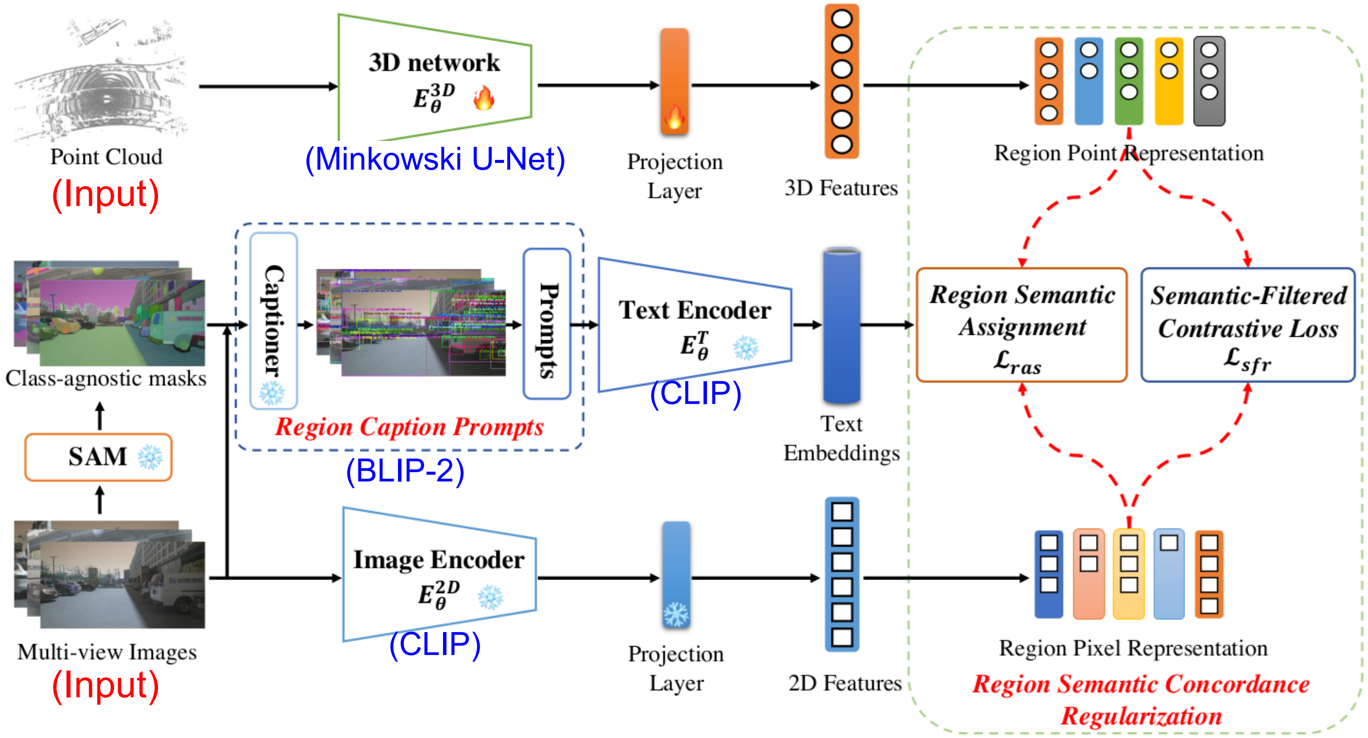
SAM

$$\mathcal{L}_{sfr}(\mathbf{P},\mathbf{Q}) = -\frac{1}{M}\sum_{i=0}^{M}\log\left[\frac{e^{((\mathbf{p}_i\cdot\mathbf{q}_i)/\tau)}}{\sum_{j\neq i}\mathbf{T}_{ij}\cdot e^{\left((\mathbf{p}_i\cdot\mathbf{q}_j)/\tau\right)}+e^{((\mathbf{p}_i\cdot\mathbf{q}_i)/\tau)}}\right]. \quad (2)$$

Cosine similarity

$$\epsilon > \Phi(t_i, x_i) - \Phi(t_i, x_j)$$

• Method - VLM2Scene



- **Experiments**

Method	Reference	nuScenes LP	nuScenes 1%	nuScenes 5%	nuScenes 10%	nuScenes 25%	nuScenes 100%	KITTI 1%
Random	N/A	8.10	30.30	47.84	56.15	65.48	74.20	39.50
PointContrast (Xie et al. 2020)	ECCV20	21.90	32.50	-	-	-	-	41.10
DepthContrast (Zhang et al. 2021)	ICCV21	22.10	31.70	-	-	-	-	41.50
PPKT (Liu et al. 2021)	arXiv21	35.90	37.80	53.74	60.25	67.14	74.52	44.00
SLidR (Sautier et al. 2022)	CVPR22	38.80	38.30	52.49	59.84	66.91	74.79	44.60
CLIP2Scene (Chen et al. 2023)	CVPR23	-	33.05	52.18	59.87	66.87	74.63	43.10
ST-SLidR (Mahmoud et al. 2023)	CVPR23	40.48	40.75	54.69	60.75	67.70	75.14	44.72
VLM2Scene (Ours)		**51.54**	**47.59**	**58.08**	**63.08**	**68.39**	**75.42**	**47.37**

Table 1: Performance comparison with other methods pre-trained on nuScenes and fine-tuned on nuScenes, and SemanticKITTI. LP indicates linear probing with frozen backbones. We report the mIoU scores for evaluation.

Experiments

Method	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
Random	30.3	0.0	0.0	8.1	65.0	0.1	6.6	21.0	9.0	9.3	25.8	89.5	14.8	41.7	48.7	72.4	73.3
PointContrast	32.5	0.0	1.0	5.6	67.4	0.0	3.3	31.6	5.6	12.1	30.8	91.7	21.9	48.4	50.8	75.0	74.6
DepthContrast	31.7	0.0	0.6	6.5	64.7	0.2	5.1	29.0	9.5	12.1	29.9	90.3	17.8	44.4	49.5	73.5	74.0
PPKT	37.8	0.0	2.2	20.7	75.4	1.2	13.2	45.6	8.5	17.5	38.4	92.5	19.2	52.3	56.8	80.1	80.9
SLidR	38.3	0.0	1.8	15.4	73.1	1.9	19.9	47.2	17.1	14.5	34.5	92.0	27.1	53.6	61.0	79.8	82.3
CLIP2Scene	33.1	0.0	1.9	10.4	70.2	1.5	9.1	41.3	0.0	20.0	28.3	87.8	15.6	37.1	52.7	74.8	77.6
ST-SLidR	40.8	0.0	2.7	16.0	74.5	3.2	25.4	50.9	20.0	17.7	40.2	92.0	30.7	54.2	61.1	80.5	82.9
Ours	**47.6**	0.0	**7.3**	**49.0**	**77.7**	**17.1**	**30.3**	**53.2**	**40.7**	**20.2**	**51.9**	**92.5**	**36.2**	**57.6**	**62.3**	**82.2**	**83.0**

Table 2: Per-class 3D semantic segmentation IoU performance on the nuScenes vaild set when fine-tuning with 1 % labels.

• Experiments

Methods	Components			nuScenes	
	RCP	SFR	RSA	1%	5%
Baseline				38.8	51.6
Ours	✓			43.4	54.6
		✓		43.8	55.0
			✓	42.1	53.6
	✓	✓		46.5	56.9
		✓	✓	45.4	56.3
	✓	✓	✓	**47.6**	**58.1**

Table 3: Ablation Study of each component.

Strategies	Methods	nuScenes	
		1%	5%
RCP	only template prompts	45.4	56.3
	only RCP	46.5	57.1
	Ours	**47.6**	**58.1**
RSC	w point-level	43.4	54.6
	w super-pixel	44.1	55.1
	w/o semantic filtering	45.0	55.9
	Ours	**47.6**	**58.1**

Table 4: Experimental results for different strategies.