UniDSeg: Unified Cross-Domain 3D Semantic Segmentation via Visual Foundation Models Prior

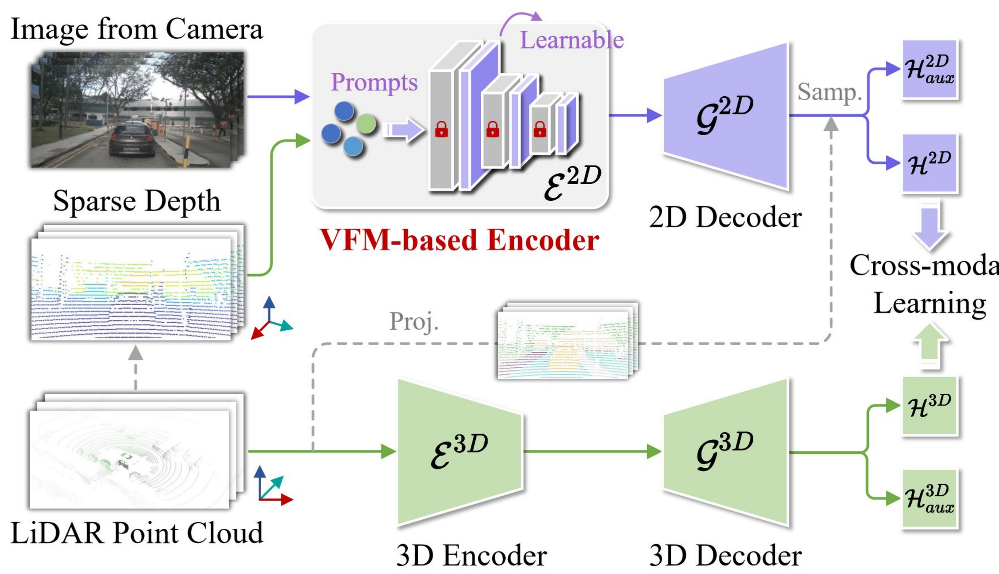
Yao Wu[1], Mingwei Xing[1], Yachao Zhang[1†], Xiaotong Luo[1], Yuan Xie[2], Yanyun Qu[1†]

[1]Xiamen University [2]East China Normal University

Motivation & Contribution

Currently, DA3SS and DG3SS methods have primarily focused on generalizing or adapting between synthetic and real scenes. This leaves a gap in exploring a universal framework, enabling the generalization and adaptation of 3SS models across datasets.

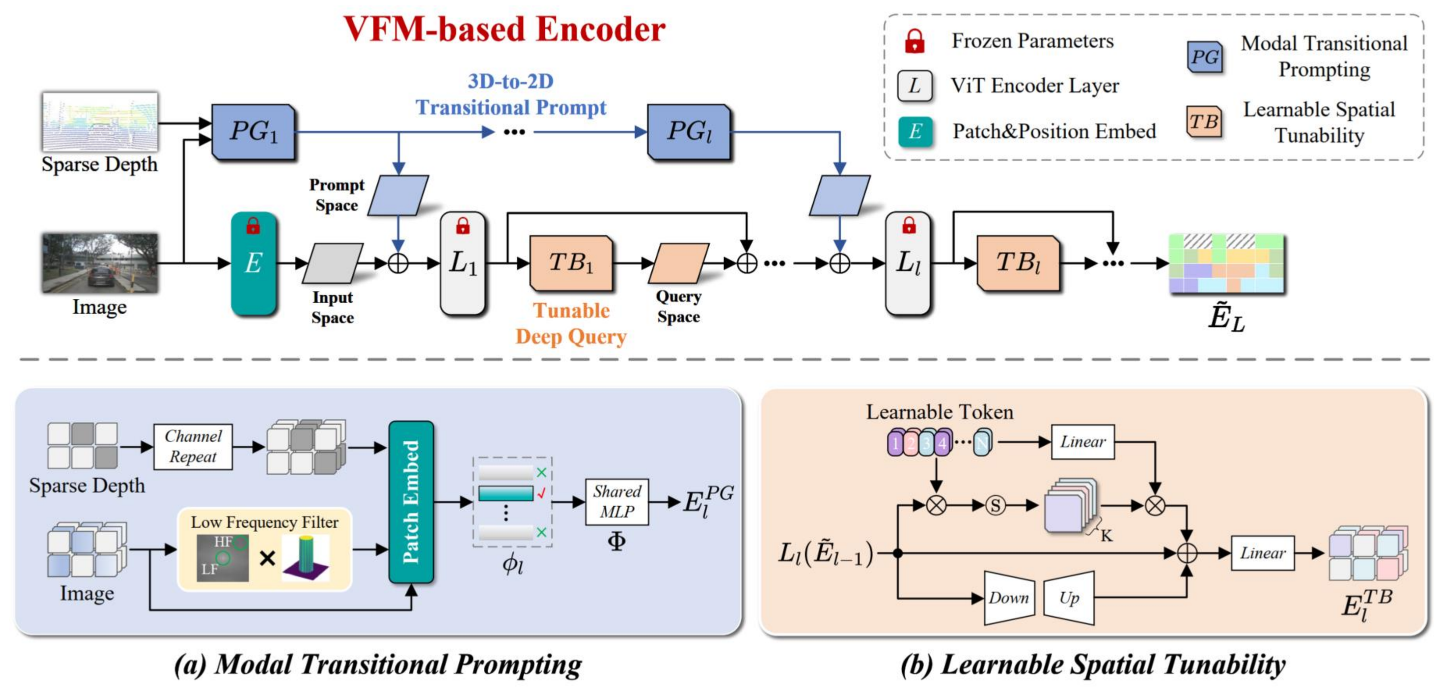
- Our method is groundbreaking in introducing the prompt-tuning concept into the universal model for DG3SS and DA3SS tasks.
- We propose a novel learnable-parameter-inspired mechanism to the off-the-shelf VFMs, which maximally preserves pre-existing target awareness in VFMs to further enhance its generalizability.



Method

Overview:

We introduce a novel task-specific VFM-based encoder, which is guided by point-level prompts from 3D information. We place layer-wise learnable blocks to take full advantage of semantic understanding of diverse levels and modalities, which inherits potential target information of VFMs into the current training model.



VFM-based Encoder:

It is designed to learn alternately between two lightweight modules: **Modal Transitional Prompting (MTP)** $PG_l(\cdot,\cdot)$ and **Learnable Spatial Tunability (LST)** $TB_l(\cdot)$.

$$\tilde{E}_l = L_l(\tilde{E}_{l-1}) + TB_l(L_l(\tilde{E}_{l-1})), \quad l = 1, 2, ..., L,$$

$$\tilde{E}_{l-1}[1:,:] = E_{l-1}[1:,:] + PG_l(X^{2D}, X^{Dep}),$$

MTP is designed to capture 3D-to-2D transitional prior and task-shared knowledge of this information from the prompt space, before being fed into layer L_l.

LST is introduced to bridge the discrepancy between the pre-training dataset and the target scene in the query space for seeking matched prompting after encoding in layer L_l.

$$E_l^{PG} = \Phi(\phi_l(E_0^{2D} \uplus E_0^{LF} \uplus E_0^{Dep})),$$

$$E_l^{TB} = \delta_2(L_l(\tilde{E}_{l-1})[1:,:] + W_{up}^{\top} \times (W_{down}^{\top} \times L_l(\tilde{E}_{l-1})[1:,:]) + J_l \times \delta_1(O_l)),$$

$$J_l = SoftMax(\frac{L_l(\tilde{E}_{l-1})[1:,:] \times O_l^{\top}}{\sqrt{D}}), \quad J_l \in \mathbb{R}^{M \times K},$$

$$O_l = O_{l,a} \times O_{l,b},$$

Experiments & Ablation Studies & Visualization

Multi-modal DA3SS & DG3SS

Task	Method	nuScenes:USA/Sing. 2D	3D	xM	nuScenes:Day/Night 2D	3D	xM	vKITTI/sKITTI 2D	3D	xM	A2D2/sKITTI 2D	3D	xM
	Source-only	58.4	62.8	68.2	47.8	68.8	63.3	26.8	42.0	42.2	34.2	35.9	40.4
DA	logCORAL [33]	64.4	63.2	69.4	47.7	68.7	63.7	41.4	36.8	47.0	35.1	41.0	42.2
DA	MinEnt [43]	57.6	61.5	66.0	47.1	68.8	63.6	39.2	43.3	47.1	37.8	39.6	42.6
DA	BDL [29]	62.0	64.8	70.4	47.0	69.6	63.0	21.5	44.3	35.6	34.7	41.7	45.2
DA	xMUDA [19]	64.4	63.2	69.4	55.5	69.2	67.4	42.1	46.7	48.2	38.3	46.0	44.0
DA	AUDA [30]	64.0	64.0	69.2	55.6	69.8	64.8	35.8	37.8	41.3	43.0	43.6	46.8
DA	DsCML [35]	65.6	56.2	66.1	50.9	49.3	53.2	38.4	38.4	45.5	39.6	45.1	44.5
DA	Dual-Cross [28]	64.7	58.1	66.5	58.5	69.7	68.0	40.7	35.1	44.2	45.9	40.0	48.6
DA	SSE [58]	64.9	63.9	69.2	62.8	69.0	68.9	45.9	40.0	49.6	44.5	46.8	48.4
DA	BFtD [45]	63.7	62.2	69.4	57.1	70.4	68.3	41.5	45.5	51.5	40.5	44.4	48.7
DA	MM2D3D [6]	71.7	66.8	<u>72.4</u>	70.5	70.2	**72.1**	53.4	50.3	56.5	42.3	46.1	46.2
DA	VFMSeg [51]	70.0	65.6	72.3	60.6	70.5	66.5	57.2	52.0	<u>61.0</u>	45.0	52.3	50.0
DA	**UniDSeg**	67.2	67.6	**72.9**	63.2	71.2	<u>71.2</u>	60.5	50.9	**62.0**	50.7	55.4	**57.5**
DG	xMUDA [19]	58.7	62.3	68.6	43.0	68.9	59.6	25.7	37.4	39.0	34.9	36.7	41.6
DG	MM2D3D [6]	69.7	62.3	<u>70.9</u>	65.3	63.2	<u>68.3</u>	37.7	40.2	<u>44.2</u>	39.6	35.9	<u>43.6</u>
DG	**UniDSeg**	66.5	64.5	**72.3**	57.0	70.5	**70.0**	57.6	44.7	**60.0**	48.8	46.3	**54.4**

VFM-based Encoder with Different Training Strategies for DG3SS

Strategy	Visual Backbone	Params	nuScenes:USA/Sing. 2D	3D	xM	nuScenes:Day/Night 2D	3D	xM	nuScenes:Sing./USA 2D	3D	xM
Finetune	CLIP:ViT-B	86.9M	62.4	64.1	69.6	53.3	70.7	68.8	65.7	67.9	72.9
Frozen	CLIP:ViT-B	0.0M	59.7	64.5	69.7	46.8	71.0	69.8	58.3	67.9	71.2
Ours	CLIP:ViT-B	1.82M	63.8	64.7	**71.5**	55.9	70.7	**70.0**	68.2	68.0	**74.0**
Finetune	CLIP:ViT-L	305M	65.5	64.5	70.4	54.9	70.7	67.3	69.9	67.8	74.5
Frozen	CLIP:ViT-L	0.0M	60.4	64.2	70.1	50.2	70.5	69.5	62.2	67.8	73.3
Ours	CLIP:ViT-L	4.70M	66.5	64.5	**72.3**	57.0	70.5	**70.0**	70.6	68.0	**75.1**

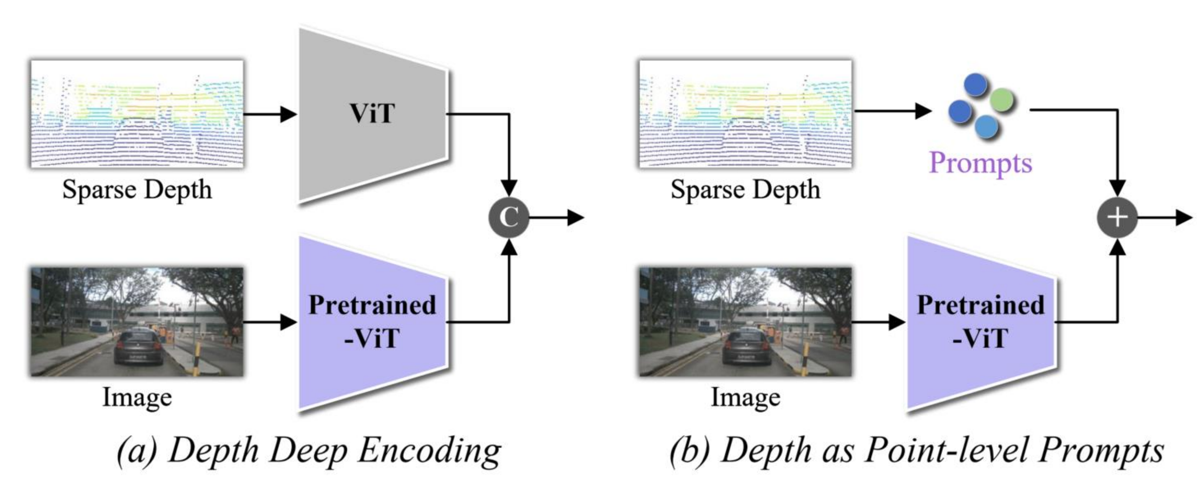
Strategy	Visual Backbone	Params	vKITTI/sKITTI 2D	3D	xM	A2D2/sKITTI 2D	3D	xM	A2D2/nuScenes 2D	3D	xM
Finetune	CLIP:ViT-B	86.9M	54.9	41.5	55.8	43.0	43.8	51.5	55.4	50.1	60.2
Frozen	CLIP:ViT-B	0.0M	49.1	42.0	54.4	35.3	43.8	48.7	51.2	49.4	58.1
Ours	CLIP:ViT-B	1.82M	55.6	43.6	**58.0**	43.2	44.6	**52.0**	56.3	50.3	**61.0**
Finetune	CLIP:ViT-L	305M	57.4	43.5	58.7	46.9	44.3	53.0	57.2	50.8	61.3
Frozen	CLIP:ViT-L	0.0M	54.0	42.9	58.4	41.8	44.4	51.4	53.7	50.0	59.6
Ours	CLIP:ViT-L	4.70M	57.6	44.7	**60.0**	48.8	46.3	**54.4**	58.0	50.7	**61.9**

Source-free DA3SS

Task	Method	Source-free	nuScenes:USA/Sing. 2D	3D	xM	nuScenes:Day/Night 2D	3D	xM	A2D2/sKITTI 2D	3D	xM
DA	Baseline	✓	58.4	62.8	68.2	47.8	68.8	63.3	34.2	35.9	40.4
DA	Consistency	✓	58.7	63.2	68.1	50.4	66.8	63.6	37.1	36.5	41.8
DA	Pseudo-Label	✓	58.9	62.7	68.5	48.3	69.0	63.2	37.6	36.6	41.5
DA	SUMMIT† [41]	✓	61.6	66.2	68.4	53.8	68.9	68.2	42.9	43.7	46.8
DA	**UniDSeg**	×	67.2	67.6	72.9	63.2	71.2	71.2	50.7	55.4	57.5
DA	**UniDSeg**	✓	69.3	71.7	73.5	62.6	70.7	68.7	49.6	59.1	58.6

Fully-supervised 3SS Results on the *SemanticKITTI* Validation Set

Method	car	bicycle	motorcycle	truck	bus	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	mIoU (%)
MinkowskiNet [7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	61.1
SPVCNN [42]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	63.8
Cylinder3D [62]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	65.9
2DPASS† [52]	95.3	47.1	73.7	81.8	56.0	73.5	87.6	2.1	92.4	45.2	78.6	1.0	90.8	61.8	88.4	69.5	75.5	58.1	51.8	64.7
2DPASS† [52] w/ TTA	96.6	52.2	77.9	91.1	68.2	77.9	92.0	0.2	94.0	50.6	81.4	1.2	91.8	66.3	89.6	72.0	77.3	63.0	53.5	68.2
Ours	96.2	47.2	70.1	84.3	64.5	74.1	89.5	2.1	92.6	46.6	79.1	3.2	90.9	62.8	88.3	69.9	75.1	58.6	52.1	**65.6**$_{+0.9}$
Ours w/ TTA	97.0	52.3	73.4	92.6	71.1	78.3	92.3	0.0	94.1	51.3	81.8	3.3	92.1	67.4	89.5	72.0	77.0	63.8	54.6	**68.6**$_{+0.4}$



(a) Depth Deep Encoding *(b) Depth as Point-level Prompts*

Role of Depth	Params	nuScenes:Sing./USA 2D	3D	xM
Deep Encoding	86.9M	66.1	67.7	73.0
Point-level Prompts	0.48M	67.8	67.9	73.8

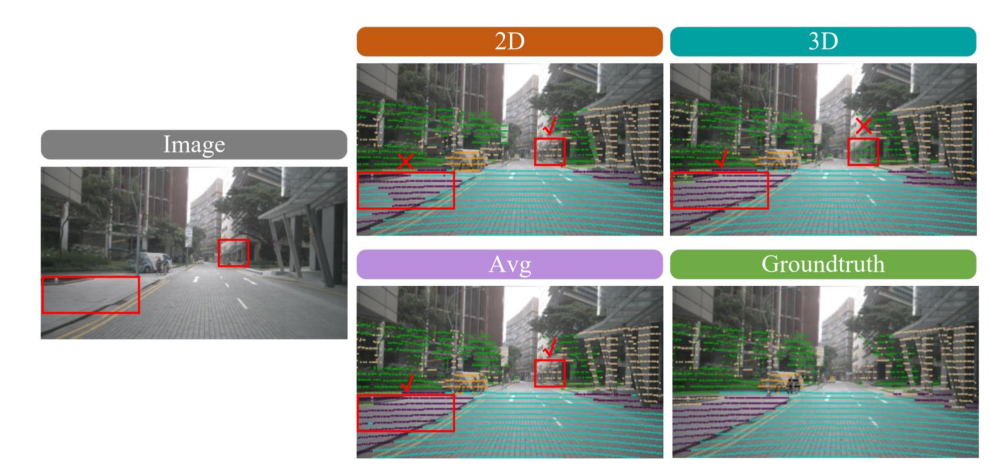
Computation Cost

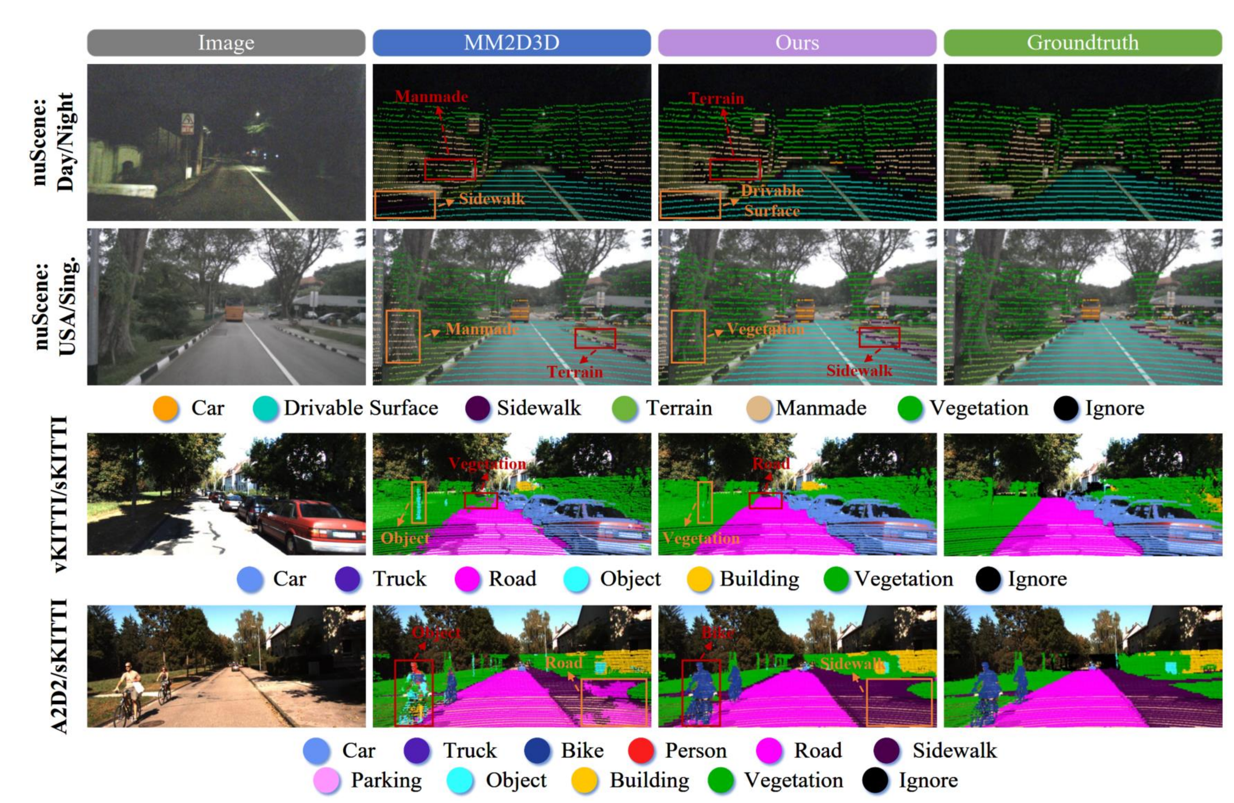
2D Backbone	Full Params	Trainable Params	Cost	MTP	LST
CLIP:ViT-B	86.9M	1.82M	**2.09%**	0.48M	1.34M
CLIP:ViT-L	305M	4.70M	**1.54%**	1.78M	2.92M
SAM:ViT-L	307M	4.34M	**1.41%**	1.42M	2.92M

Different Components

MTP	LST	nuScenes:Sing./USA 2D	3D	xM	A2D2/sKITTI 2D	3D	xM
Frozen VFM		58.3	67.9	71.2	35.3	43.8	48.7
✓		63.9	67.8	72.5	40.4	44.0	50.5
	✓	65.7	67.8	73.3	41.8	44.2	51.1
✓	✓	68.2	68.0	74.0	43.2	44.6	52.0

Visualization





Different 2D and 3D Backbones

3D Backbone	DA3SS	USA/Sing. 2D	3D	xM
SparseConvNet	xMUDA	64.4	63.2	69.4
SparseConvNet	UniDSeg	67.2	67.6	72.9
MinkowskiNet	xMUDA	65.9	64.0	69.7
MinkowskiNet	UniDSeg	67.5	68.6	73.1

Task	2D Backbone	USA/Sing. 2D	3D	xM
DG	CLIP:ViT-L	66.5	64.5	72.3
DG	SAM:ViT-L	66.8	64.7	72.6
DA	CLIP:ViT-L	67.2	67.6	72.9
DA	SAM:ViT-L	67.8	68.8	73.3

Contact us

Github.io WeChat

Email: wuyao@stu.xmu.edu.cn