universität freiburg

ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION ATLANTA 2025

robot learning

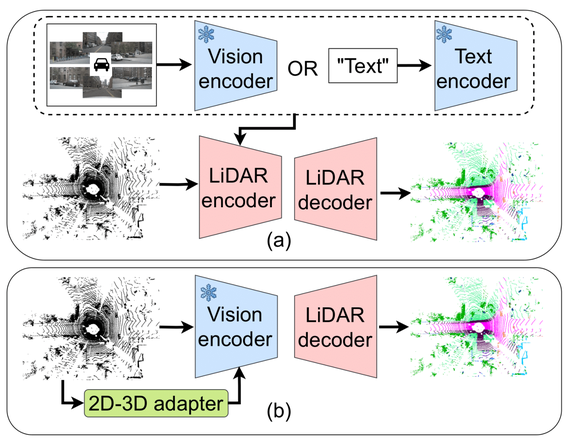
Label-Efficient LiDAR Scene Understanding with 2D-3D Vision Transformer Adapters

Julia Hindel*, Rohit Mohan*, Jelena Bratulic, Daniele Cattaneo, Thomas Brox, and Abhinav Valada

* Equal Contribution

Motivation

- Foundation models have advanced vision and language models, but LiDAR still lacks large-scale pretraining and robust foundation models.
- Existing methods rely on paired camera-LiDAR data and still require training 3D encoders from scratch to leverage foundation models.

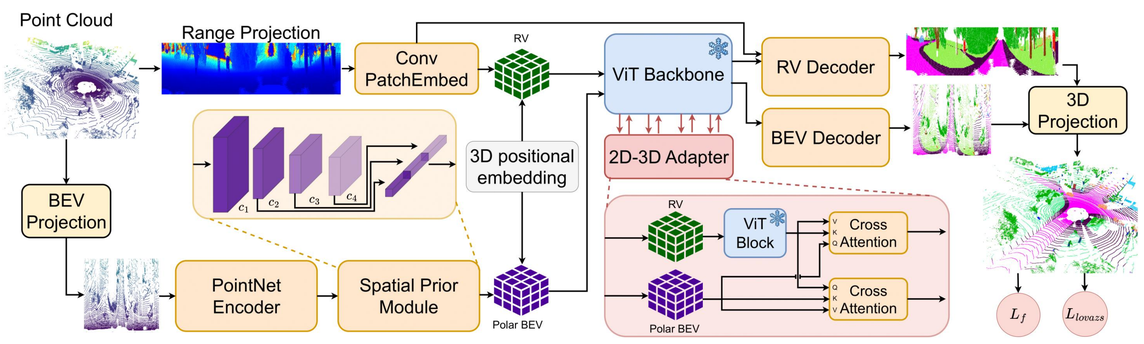


We propose BALViT, *a universal foundation model for LiDAR segmentation that leverages a* **novel 2D-3D adapter** *to enable* **efficient reuse of vision model features** *through structured 2D representations.*

Method

BALViT Architecture

- *Dual-View Encoding*: LiDAR point clouds are encoded using Range View (RV) and Bird's-Eye View (BEV) branches.
- *Vision Backbone on RV*: A frozen vision transformer (ViT) backbone processes RV features with a learnable patch embedding, leveraging rich pre-trained visual representations.
- *2D-3D Adapter:* Enables bidirectional injection of BEV and ViT-processed RV features via stacked parallel cross-attention for mutual refinement.
- *3D Positional Embeddings:* Provide spatial alignment between RV and BEV, enabling seamless cross-view attention in the adapter.
- *Parallel Decoding*: RV and BEV use independent decoders to predict semantic labels.



Inference Fusion

- The output is selected based on the highest logit confidence $\hat{y}$ from RV and BEV predictions, using a fixed threshold s.
- This ***simple uncertainty-aware selection*** favors fine-grained RV when confident and falls back to BEV for globally consistent context.

$$\text{Output} = \begin{cases} \hat{y}_{\text{RV}}, & \text{if } \hat{y}_{\text{RV}} > s \\ \hat{y}_{\text{RV}} & \text{if } \hat{y}_{\text{RV}} \leq s \text{ and } \hat{y}_{\text{BEV}} < s \\ \hat{y}_{\text{BEV}}, & \text{if } \hat{y}_{\text{RV}} \leq s \text{ and } \hat{y}_{\text{BEV}} > s \end{cases},$$

Results

Benchmark

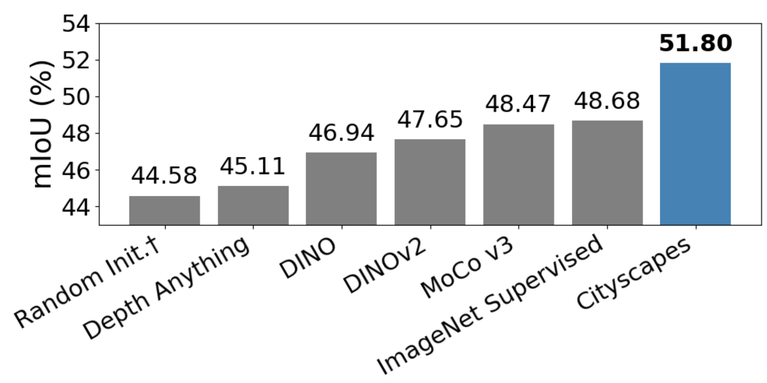
BALViT consistently outperforms baseline models on the SemanticKITTI and nuScenes datasets under low-label settings.

Method	SemanticKITTI 0.1%	SemanticKITTI 1%	nuScenes 0.1%	nuScenes 1%
SR-Unet18	-	39.50	-	30.30
FRNet	30.09	40.78	28.03	48.98
SphereFormer	29.21	42.81	30.42	50.06
RangeViT	28.74	43.53	27.79	52.88
SLidR	-	44.60	-	38.30
ST-SLidR	-	44.72	-	40.75
SEAL	-	46.63	-	45.84
CLIP2Scene	-	42.60	-	56.30
Frozen ViT backbone	29.97	45.91	28.72	54.70
Bias tuning	30.86	45.63	28.15	56.05
LoRA	31.65	46.27	28.27	57.57
VPT	31.07	46.08	29.68	55.67
ViT Adapter	29.55	45.01	27.50	56.06
BALViT (Ours)	**32.85**	**51.80**	**31.86**	**59.27**

Fully Supervised
Vision Distillation
Parameter Efficient Fine-Tuning

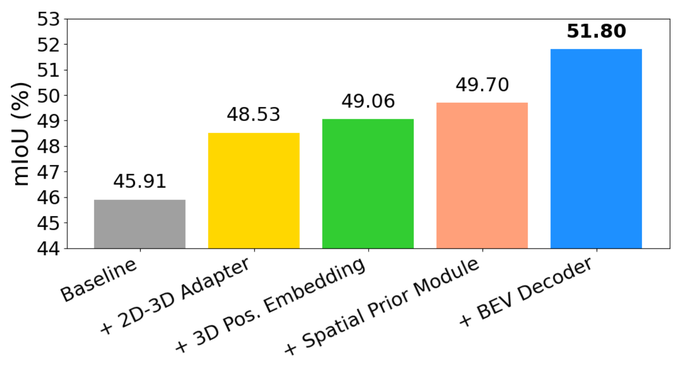
Effects of Vision Backbone Initialization

Cityscapes pretraining performs (1% SemanticKITTI), highlighting the value of domain-aligned vision backbones.



Ablation Study on BALViT Components

The 2D-3D adapter and BEV decoder contribute most to gains over the frozen ViT baseline (1% SemanticKITTI).



Qualitative