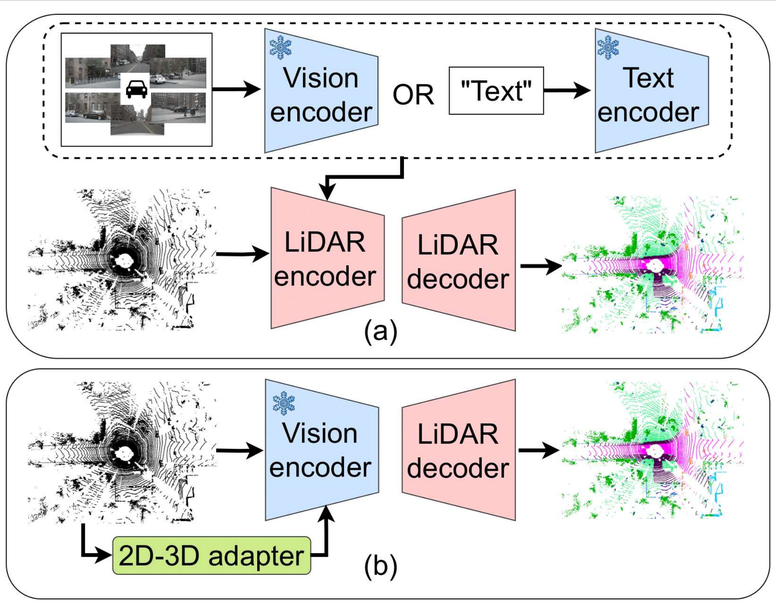
Label-Efficient LiDAR Scene Understanding with 2D-3D Vision Transformer Adapters

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

* Equal Contribution

Motivation



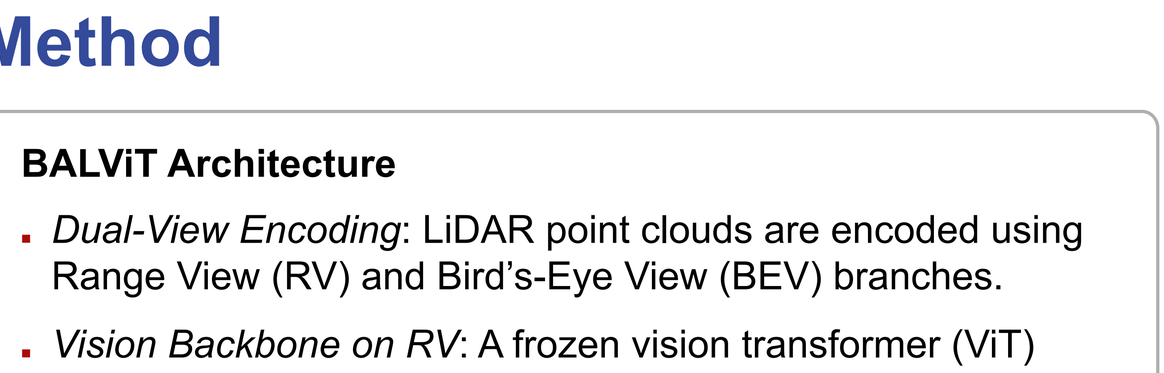
a) Vision or language foundation models are employed to distill knowledge into tailored LiDAR architectures.

b) Transferring a pre-trained vision model into the LiDAR domain using a 2D-3D adapter (ours).

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.
- *3D Positional Embeddings:* Provide spatial alignment between RV and BEV, enabling seamless cross-view attention in the adapter.
- *2D-3D Adapter:* Enables bidirectional injection of BEV and ViT-processed RV features via stacked parallel cross-attention for mutual refinement.
- *Parallel Decoding*: RV and BEV use independent decoders to predict semantic labels.



Inference Fusion

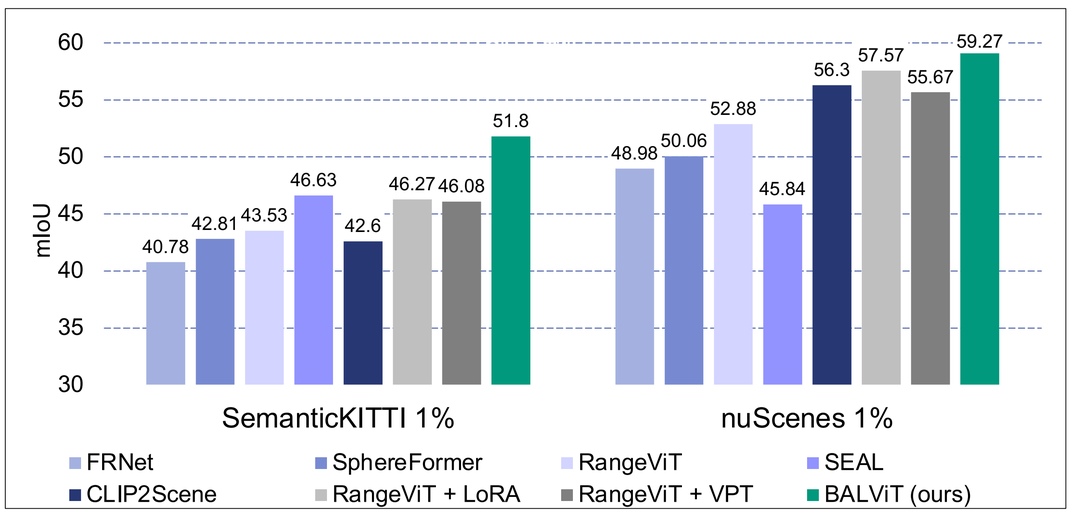
- The output is selected based on the highest logit confidence $\hat{y}$ from RV and BEV predictions, using a fixed threshold s.
- This ***simple uncertainty-aware selection*** favors fine-grained RV when confident and falls back to BEV for globally consistent context.

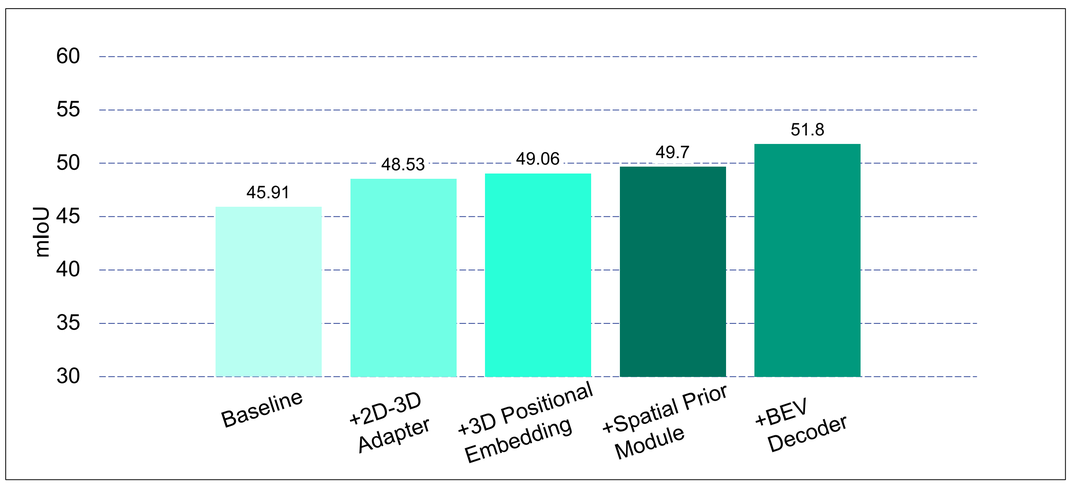
$$\text{Output} = \begin{cases} \hat{y}_{\text{RV}}, & \text{if } \hat{y}_{\text{RV}} > s \\ \hat{y}_{\text{RV}} & \text{if } \hat{y}_{\text{RV}} \leq s \text{ and } \hat{y}_{\text{BEV}} < s \\ \hat{y}_{\text{BEV}}, & \text{if } \hat{y}_{\text{RV}} \leq s \text{ and } \hat{y}_{\text{BEV}} > s \end{cases},$$

Results

Benchmark

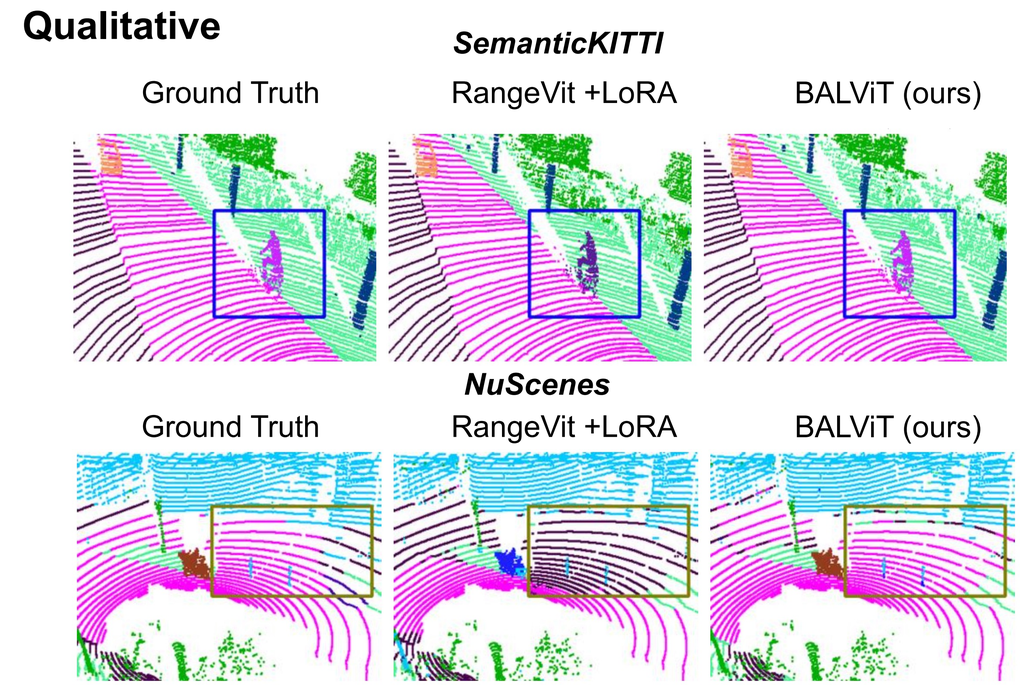
BALViT consistently outperforms baseline models on the SemanticKITTI and nuScenes datasets under low-label settings.





We observe better segmentation of large objects and boundary estimations.

Qualitative



Conclusion

- Outperforms existing supervised and self-supervised baselines on label-efficient training settings of 0.1% and 1% on the SemanticKITTI and nuScenes datasets.
- Vision models provide valuable priors for LiDAR semantic scene segmentation.
- Integrate pre-trained vision architectures in point cloud segmentation to leverage future advancements in foundation model research.