xModel-KD: Cross-modal Knowledge Distillation for 3D Scene Perception using LiDAR

Thenukan Pathmanathan[1] Kanchan Keisham[2] Thangarajah Akilan[1]

[1]Lakehead University, Canada [2]Vellore Institute of Technology, India

Abstract

Point cloud segmentation is constrained by high annotation costs and inherent modality limitations. While 2D images provide rich texture, they lack geometric structure; 3D point clouds capture geometry but lack semantic richness. We propose **xModel-KD**, a training-only cross-modal knowledge distillation framework that transfers semantic knowledge from frozen 2D teachers into compact 3D networks with **zero inference overhead**. Through multi-scale contrastive alignment and KL-based distillation, our method achieves 2% **absolute mIoU improvement** over LiDAR-only baselines while using **23$\times$ fewer parameters** than competing methods.

Motivation & FOV Mismatch

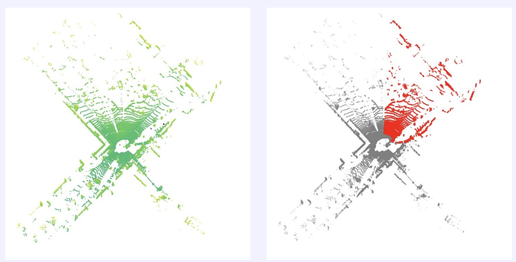
Key Challenges in Multi-modal 3D Segmentation

1. **FOV Mismatch**: A significant portion of LiDAR points lies outside the camera field-of-view, limiting direct cross-modal correspondence.
2. **Computational Overhead**: Multi-modal fusion introduces substantial latency and memory cost at inference.
3. **Limited Foundation Model Use**: 3D segmentation often trains from scratch, unlike 2D vision which benefits from large-scale pretraining.

Sensor Coverage Gap

- **LiDAR**: 360° geometric coverage
- **Camera**: Limited frontal FOV, typically ∼90–120°

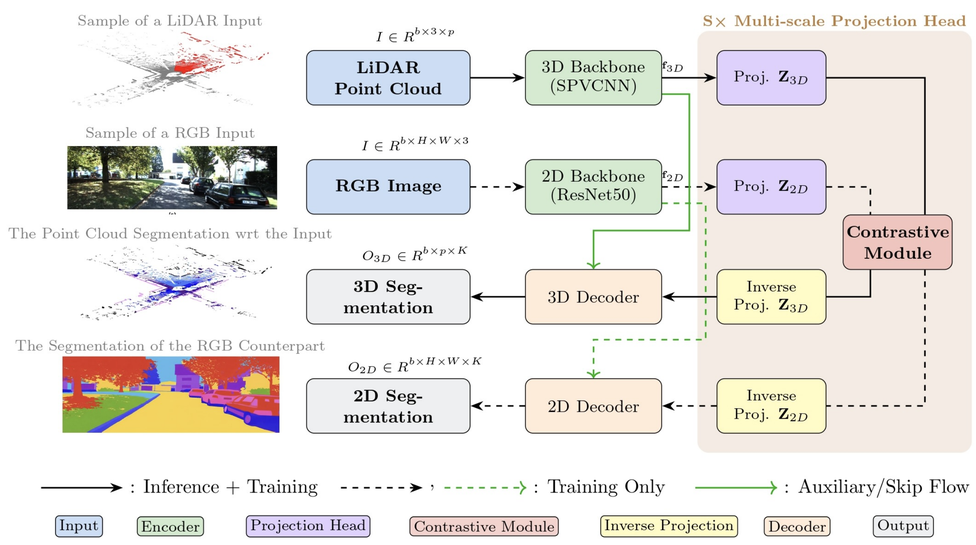
Traditional fusion methods rely mainly on points inside the camera FOV, leaving out-of-FOV geometry weakly supervised.



Full LiDAR point cloud Camera-FOV region marked in red

In SemanticKITTI, the LiDAR observes the full surrounding scene, while the camera only covers a limited forward-facing region. Our approach leverages **shared 3D parameters** learned under 2D supervision, allowing out-of-FOV points to benefit indirectly from cross-modal alignment.

Architecture



Layer Specification

Table 1. Layer specifications of the proposed `xModal-KD`.

Component	Input	Output
Encoders		
LiDAR input - $\mathbf{X}$	$[b, p, 4]$	–
RGB input - $\mathbf{I}$	$[b, 3, H, W]$	–
3D Backbone (SPVCNN)	$\mathbf{X}$	$[b, p, l_c]$
2D Backbone (ResNet50)	$\mathbf{I}$	$[b, H', W', l_h]$
Projection Head		
Proj. Head - $\mathbf{Z}_{3D}$	l_c	$\mathbf{d}_p$ - 128
Proj. Head - $\mathbf{Z}_{2D}$ (train only)	l_h	$\mathbf{d}_p$ - 128
Contrastive Module (train only)	$(\mathbf{Z}_{3D}, \mathbf{Z}_{2D})$	$\mathcal{L}_{contrast}$
Inverse Projection		
Inverse Proj. - $\mathbf{Z}_{3D}$	$\mathbf{d}_p$ - 128	l_h
Inverse Proj. - $\mathbf{Z}_{2D}$ (train only)	$\mathbf{d}_p$ - 128	l_c
Decoders and Outputs		
3D Decoder	$[b, p, l_h]$	$[b, p, K]$
2D Decoder (train only)	$[b, H', W', l_c]$	$[b, H, W, K]$

Note: b - batch size; p - number of data points in each LiDAR point cloud sample; H, W - spatial dimension of the inpput RGB image.

Key Contributions

- **Training-only framework** with zero inference overhead — 2D branch removed at test time.
- **Multi-scale contrastive alignment** for hierarchical feature transfer.
- **Lightweight architecture** 1.93M parameters (23× smaller than 2DPASS).
- **Strong performance** 69.1% mIoU on SemanticKITTI validation set.

Mathematical Formulation

- **Multi-Scale Feature Projection:** At each scale s, features are projected into a shared embedding space: $\mathbf{z}^s_{3D} = \text{Proj}^s_{3D}(\mathbf{f}^s_{3D})$, and $\mathbf{z}^s_{2D} = \text{Proj}^s_{2D}(\mathbf{f}^s_{2D})$.
- **Contrastive Loss (NT-Xent):** For matched point-pixel pairs with temperature τ:

$$\mathcal{L}^s_{\text{con}} = -\log \frac{\exp(\text{sim}(\mathbf{z}^s_{3D}, \mathbf{z}^s_{2D})/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{z}^s_{3D}, \mathbf{z}^{(j)}_{2D})/\tau)} \tag{1}$$

Aggregated:

$$\mathcal{L}_{\text{con}} = \frac{1}{S} \sum_{s=1}^{S} \mathcal{L}^s_{\text{con}} \tag{2}$$

- **Knowledge Distillation Loss:** For points $i \in \Omega$ within camera FOV, with projection $\pi(i)$:

$$\mathcal{L}_{\text{KD}} = \frac{1}{|\Omega|} \sum_{i\in\Omega} \text{KL}\left(P_{2D}(\pi(i)) \| P_{3D}(i)\right) \tag{3}$$

- **Total Loss:**

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{3D} + \mathcal{L}_{2D} + \lambda_{\text{con}}\mathcal{L}_{\text{con}} + \lambda_{\text{KD}}\mathcal{L}_{\text{KD}}, \tag{4}$$

where $\mathcal{L}_{3D}$ and $\mathcal{L}_{2D}$ combine cross-entropy and Lovász-Softmax losses.

Experimental Setup

- **Dataset**: SemanticKITTI
- **3D Backbone**: SPVCNN
- **2D Backbone**: ResNet50
- **Training**: 64 epochs, batch size 16, cosine LR decay
- **Inference**: LiDAR-only

Results on SemanticKITTI

Comparison with SOTA

Method	mIoU (%)	Params
SqueezeSegV2	39.7	–
DarkNet53Seg	49.9	–
RangeNet53++	52.2	–
Cylinder3D	68.9	–
RPVNet	70.3	–
2DPASS	**72.9**	45.6M
Baseline (LiDAR-only)	67.0	1.93M
xModel-KD (Ours)	**69.1**	**1.93M**

Ablation Study

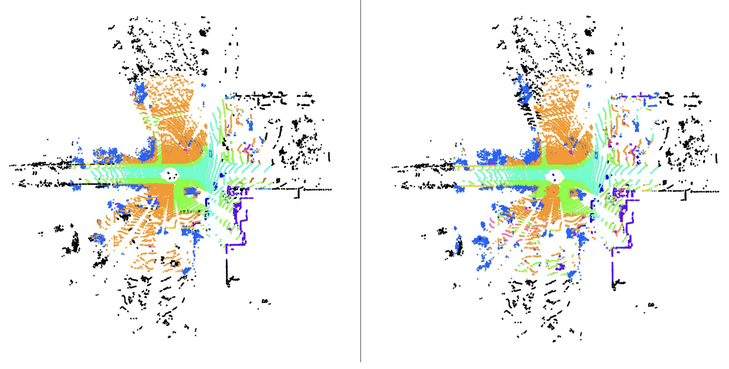
Configuration	MSCA	KD	mIoU (%)
Baseline	×	×	67.0
+ Contrastive	✓	×	67.8
+ KL-div	✓	×	68.4
xModel-KD	✓	✓	**69.1**

The Findings:

- **+2.1% mIoU** over baseline.
- Strong on hard classes (truck: 78.5%, bicycle: 59.0%)

Qualitative Results

Ground Truth `xModel-KD`**'s Prediction**



For the input sample `000000.bin` from SemanticKITTI validation set.

Legend: road, noise, parking, sidewalk, building, fence, traffic-sign.

The model produces predictions that closely match ground truth, with consistent spatial structures across different object classes.

Conclusion & Future Work

- Transfers rich 2D semantic priors to 3D networks without inference overhead
- Achieves **+2.1% mIoU improvement** over the LiDAR-only baseline with only 1.93M parameters
- Demonstrates that cross-modal learning can be fully decoupled from deployment

Future Work

- **Larger backbones**: Exploring stronger 3D backbones to further close the gap with fusion-based methods
- **Foundation model teachers**: Leveraging vision-language models as 2D teachers for richer semantic priors

Acknowledgment: This research was enabled in part by the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), DGECR/416-2022, and the support provided by the Digital Research Alliance of Canada (`https://alliancecan.ca/en`).