How Do Images Align and Complement LiDAR? Towards a Harmonized Multi-modal 3D Panoptic Segmentation

Yining Pan[1], Qiongjie Cui[1], Xulei Yang[2], Na Zhao[1]*

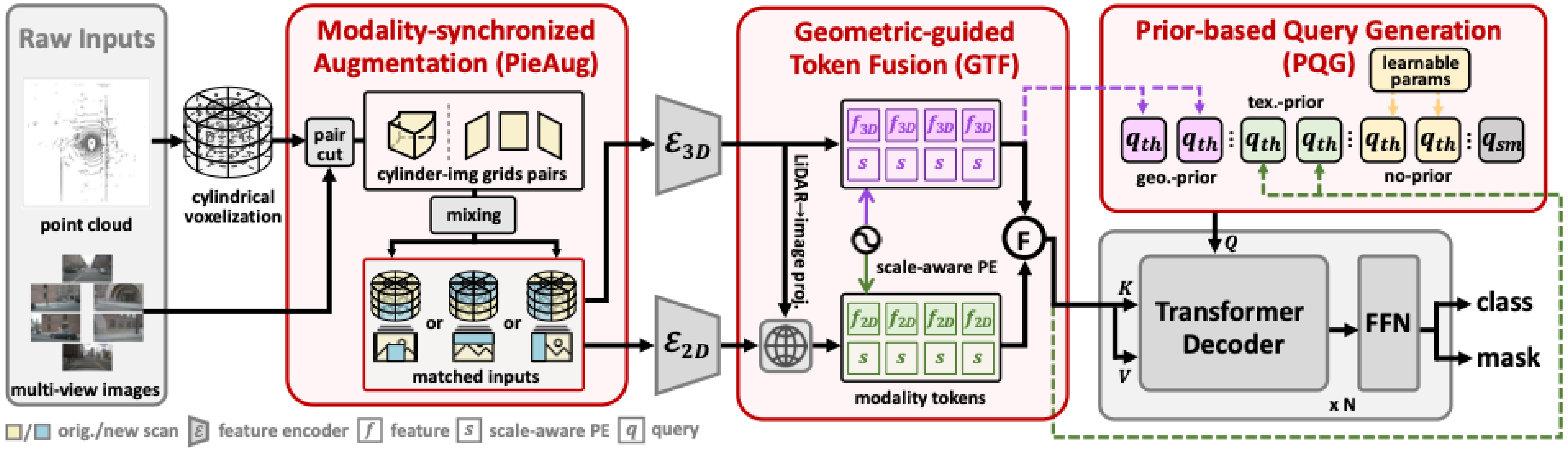
[1] Singapore University of Technology and Design; [2] Institute for Infocomm Research (I2R), A*STAR

Introduction

Motivation:
LiDAR inherently suffers from sparsity, limiting its effectiveness for small or distant objects. Images, providing dense texture details, naturally complement LiDAR. To leverage this synergy, we propose Image-Assists-LiDAR (IAL).

Key objectives:

- Propose a multimodal 3D panoptic segmentation framework without cumbersome post-processing.
- Introduce PieAug, a generalized approach for synchronized LiDAR-image augmentation.
- Design GTF and PQG modules to align and complement LiDAR and image features by generating effective tokens and queries.



Arxiv

GitHub

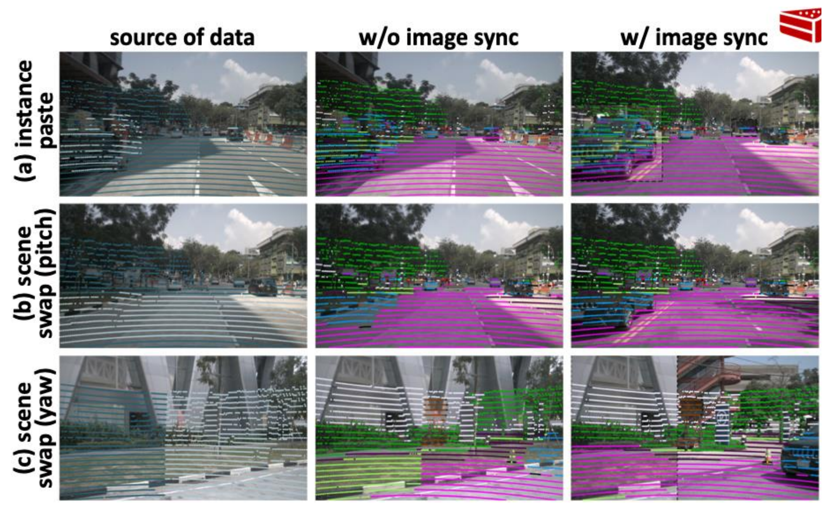
PieAug

Problem & Observation:

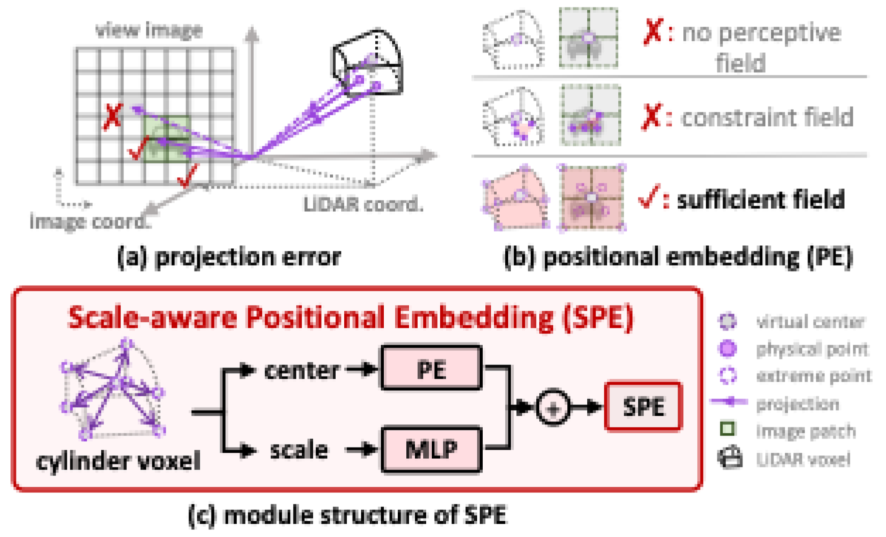
- Pioneer methods only augment on the LiDAR side, causing the misalignment between modalities.

Solution: *"Things are as easy as sharing a pie."*

- **Cut and Make a Pair**: Each cylindrical voxel is paired with corresponding image grids $\langle v_i, g_i \rangle$.
- **Mixing**: Determine the binary mask $\mathbf{S}$ to organize the LiDAR-image pairs: $\mathbf{V}^{\text{aug}} = \mathbf{V}^{\text{org}} \otimes (1-\mathbf{S}) + \mathbf{V}^{\text{new}} \otimes \mathbf{S}$. This paradigm can be transformed to different modes:
  - Instance Pasting: $\mathbf{S} = \bigcup_{r=1,\theta=1,z=1}^{R\times\Theta\times Z} \mathbb{1}[(r,\theta,z) \in \mathcal{C}]$
  - Scene Swapping: $\mathbf{S}(r,\theta,z) = \begin{cases} 1, & \text{if } \theta \in \mathcal{O} \\ 0, & \text{otherwise} \end{cases}$
- **Remarks**. PieAug can generalize most LiDAR-only augmentations like instance copy-paste, LaserMix, PolarMix etc., but achieve LiDAR & image synchronized augmentation.



Geometric-Guided Token Fusion



Problem & Observation:

- The projection error caused by virtual points, more severe when the voxel size grows.
- Ignore the effect of perceptive field for both 3D and 2D features. E.g., not indicate or constrained field limited by physical points.

Solution:

- Project physical points and aggregate the representation for accurate alignment.
- Apply scale embedding for both modalities. The scale is determined by virtual points.

Prior-Based Query Generation

Problem & Observation:

- Learnable queries tend to converge to easier samples.
- Giving query a positional hint helps model locating.

Table 1. Preliminary study of positional embedding for objects of thing classes. We conduct the experiment on our LiDAR branch. "GT" denotes using the ground truth center position, while "Noise" denotes adding Gaussian noise with a kernel size of 3 to the GT center position. "th" and "st" is the thing and stuff classes.

Modality	GT	Noise	PQ	mIoU	PQ$^{\text{th}}$	PQ$^{\text{st}}$
LiDAR			77.0	75.9	77.8	75.7
LiDAR	✓		83.2	82.3	88.5	74.4
LiDAR	✓	✓	81.8	79.8	86.8	73.6

Solution:

- **Geometric-Prior Query**
  - LiDAR feature excels in precise location prediction.
  - Predict BEV center heatmap and average the height.
- **Texture-Prior Query**
  - Texture feature is denser for hard samples, e.g., small and remote objects.
  - Extract the 2D mask and Clustering 3D points within the mask frustum.
- **No-Prior Query**
  - instances without advanced priors exhibit specific representation paradigm.
  - Apply learnable queries to learn this paradigm.

All thing queries and semantic queries are combined and updated through transformer.

Experiments

Our IAL method achieves SOTA results on outdoor panoptic segmentation:

- Rank **1st** on the nuScenes-panoptic leaderboard.
- **Highest** performance on nuScenes and SemanticKITTI validation sets.
- Outperforms LCPS and Panoptic-FusionNet by up to **5.1%** on PQ.

nuScenes

Method	M.	PQ	PQ†	RQ	SQ	PQ$^{\text{th}}$	RQ$^{\text{th}}$	SQ$^{\text{th}}$	PQ$^{\text{st}}$	RQ$^{\text{st}}$	SQ$^{\text{st}}$	mIoU
P.-PolarNet (Zhou et al., 2021)	L	67.7	71.0	78.1	86.0	65.2	74.0	87.2	71.9	84.9	83.9	69.3
P.-PHNet (Li et al., 2022a)	L	74.7	77.7	84.2	88.2	74.0	82.5	89.0	75.9	86.9	86.8	79.7
LCPS (Zhang et al., 2023)	L	72.9	77.6	82.0	88.4	72.8	80.5	90.1	73.0	84.5	85.5	75.1
P.-PCSCNet (Song et al., 2024)	L	72.7	75.4	84.8	86.4	71.2	82.9	86.6	75.1	84.2	84.2	69.8
P3Former (Xiao et al., 2025)	L	75.9	78.9	84.7	89.7	76.9	83.3	92.0	75.4	87.1	86.0	76.8
IAL (our LiDAR branch)	L	77.0	79.6	85.1	90.2	77.8	83.8	92.6	75.7	87.3	86.2	75.9
LCPS (Zhang et al., 2023)	L+C	79.8	84.0	88.5	89.8	82.3	89.6	91.7	75.6	86.5	86.7	80.5
P.-FusionNet (Song et al., 2024)	L+C	77.2	79.3	87.2	87.8	77.5	87.7	88.2	76.2	85.9	86.0	73.4
IAL (ours)	L+C	**82.3**	**84.7**	**89.7**	**91.5**	**85.3**	**90.6**	**94.1**	**77.3**	**88.2**	**87.2**	**80.6**

Method	M.	PQ	PQ†	RQ	SQ	PQ$^{\text{th}}$	RQ$^{\text{th}}$	SQ$^{\text{th}}$	PQ$^{\text{st}}$	RQ$^{\text{st}}$	SQ$^{\text{st}}$	mIoU
P.-PolarNet (Zhou et al., 2021)	L	63.6	67.1	75.1	84.3	59.0	69.8	84.3	71.3	83.9	84.2	67.0
P.-PHNet (Li et al., 2022a)	L	80.1	82.8	87.6	91.1	82.1	88.1	93.0	76.6	86.6	87.9	80.2
LCPS (Zhang et al., 2023)	L	72.8	76.3	81.7	88.6	72.4	80.0	90.2	73.5	84.6	86.1	74.8
IAL (our LiDAR branch)	L	75.1	77.7	83.0	90.1	75.0	80.9	92.4	75.2	86.5	86.4	73.3
4DFormer (Athar et al., 2023)	L+C	78.0	81.4	86.6	89.7	80.0	87.8	90.9	74.6	84.5	87.6	80.4
LCPS (Zhang et al., 2023)	L+C	79.5	82.3	87.7	90.3	81.7	88.6	92.2	75.9	86.3	87.3	78.9
IAL (ours)	L+C	**82.0**	**84.3**	**89.3**	**91.6**	**84.8**	**90.2**	**93.8**	**77.5**	**87.8**	87.8	79.9

- Ablation studies for proposed modules.

Table 5. Ablation study of the proposed modules in our framework. "PIE" denotes the PieAug module.

PIE	GTF	PQG	PQ	PQ†	RQ	SQ	mIoU
			75.7	78.1	84.4	88.3	73.8
✓			78.4	81.0	86.9	90.0	78.2
✓	✓		81.1	83.5	89.0	90.9	80.2
✓	✓	✓	**82.3**	**84.7**	**89.7**	**91.5**	**80.6**

SemanticKITTI

Method	M.	PQ	PQ†	RQ	SQ	mIoU
P.-PolarNet	L	59.1	64.1	70.2	78.3	64.5
DS-Net	L	57.7	63.4	68.0	77.6	63.5
EfficientLPS	L	59.2	65.1	69.8	75.0	64.9
P.-PHNet	L	61.7	–	–	–	65.7
CenterLPS	L	62.1	67.0	72.0	80.7	–
LCPS	L	55.7	65.2	65.8	74.0	61.1
P3Former	L	62.6	66.2	72.4	76.2	–
IAL (LiDAR)	L	62.0	65.1	71.9	76.0	64.9
LCPS	L+C	59.0	**68.8**	68.9	79.8	63.2
IAL (ours)	L+C	**63.1**	66.3	**72.9**	**81.4**	**66.0**

More results can be found in our paper.

Qualitative Results

The visualization of panoptic prediction (left) and error map (right). IAL showcases significant performance improvements in: 1. distinguishing closed objects; 2. detecting distant objects; 3. recognizing FP and FN objects.