# PF3Det: A Prompted Foundation Feature Assisted Visual LiDAR 3D Detector

Kaidong Li[1], Tianxiao Zhang[1], Kuan-Chuan Peng[2], Guanghui Wang[3]

[1]University of Kansas, [2]Mitsubishi Electric Research Laboratories, [3]Toronto Metropolitan University

CVPR Nashville JUNE 11-15, 2025

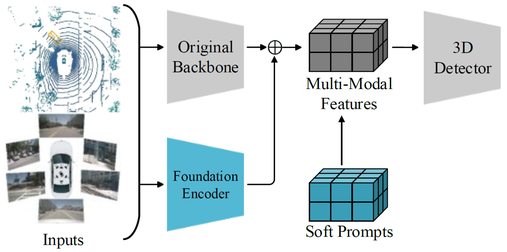
## Motivation

- Multi-modal foundation models demonstrate capability to be adapted for a wide range of downstream tasks, trained to extract common feature spaces from images, points, texts.
- Prompt tuning can adapt to different tasks, preserving the foundation models' capabilities.

## Related Work

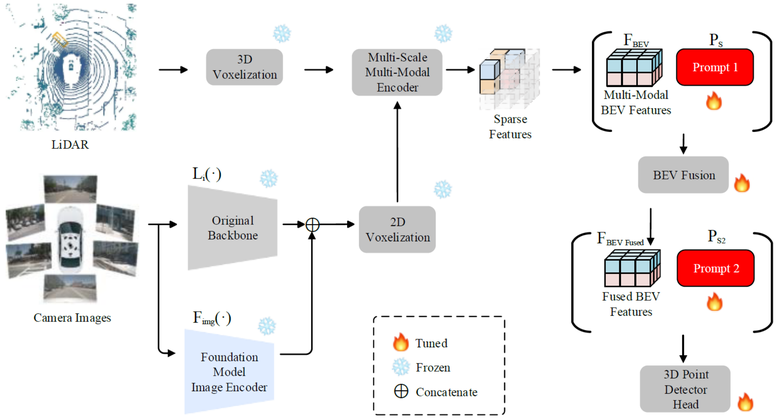
- **MSMDFusion** [21]: Deep fuses camera and LiDAR's features with two modules, Multi-Depth Unprojection (MDU) and Gated, Modality-Aware Convolution (GMA-Conv).
- **Multi-modal Large Language Models**: **CLIP** [42] has demonstrated impressive zero-shot generalization abilities, inspiring a range of applications. **ULIP** [52] creates language, image, and point triplet data, and learns to encode data from these three modalities into a unified feature space.
- **VPT** [20]: introduces a set of vision prompts at different layers to efficiently adapt large pre-trained models to downstream tasks.

## Contributions



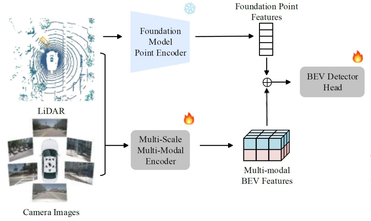

- We propose PF3Det, a visual LiDAR 3D detection model architecture that can efficiently learn to predict high-quality 3D objects with a small amount of data.
- Our proposed method takes multi-modal foundational features and bridges the modality domain gaps in the bird-eye-view (BEV) stage by incorporating the soft prompts for convolutional layers.
- Our proposed PF3Det achieves the SOTA performance on the 3D object detection task under limited training data. We run extensive experiments to explore different settings and provide guidance on how to select parameters.

## PF3Det Architecture



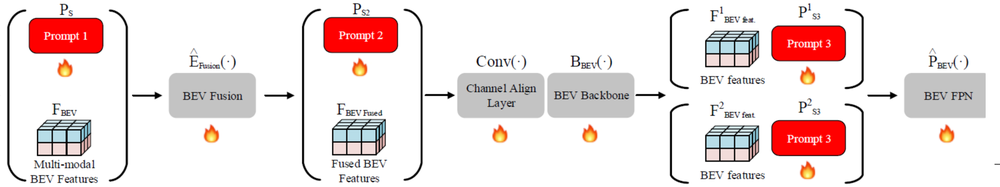

## Foundational Branch

- The foundational extractor F (·), helps the model extract more high-level information from the input images.




## Multi-modal Soft-prompt Adapter

- Two levels of soft prompts are added at different levels of the detector at $F_{BEV}$ and $F_{BEV\ Fused}$ .




## Qualitative Results

| Exp. ID | Method | Concat-FM | FM Backbone | Prompt Tuning | Prompt Tuning Layers | NDS | mAP |
|---|---|---|---|---|---|---|---|
| 1 | MSMDFusion [21] | ✗ | N/A | ✗ | N/A | 69.23 | 64.50 |
| 2 | **PF3Det** | ✓ | ResNet50 [14] | ✗ | N/A | 68.36 | 62.86 |
| 3 | **PF3Det** | ✓ | ViT-L [10] | ✗ | N/A | 70.21 | 66.15 |
| 4 | **PF3Det** | ✗ | N/A | ✓ | 1 | 69.90 | 65.66 |
| 5 | **PF3Det** | ✗ | N/A | ✓ | 2 | 70.00 | 66.69 |
| 6 | **PF3Det** | ✓ | ViT-L [10] | ✓ | 2 | **70.42** | **66.92** |

Table 1. Overall experimental results on nuScenes validation dataset. MSMDFusion results are reproduced by training on 5% of nuScenes dataset. When we experiment with the foundation model (FM), we use CLIP [46] with either ResNet50 or ViT-L as its backbone. Acronyms: Exp.: experiment; Concat-FM: concatenation with FM features.

### Foundational branch ablation study

| Foundation Model | Modality | Foundation Model Backbone | Channel Size | Upsample | NDS | mAP |
|---|---|---|---|---|---|---|
| CLIP [42] | image | ViT-L [10] | 768 | ✗ | **70.21** | **66.15** |
| CLIP [42] | image | ResNet50 [14] | 1024 | ✗ | 68.36 | 62.86 |
| ULIP [52] | point cloud | PointBERT [55] | 50 | ✗ | 68.42 | 64.33 |
| ULIP [52] | point cloud | PointBERT [55] | 100 | ✗ | 67.89 | 63.66 |
| ULIP [52] | point cloud | PointBERT [55] | 50 | ✓ | 69.50 | 65.53 |
| ULIP [52] | point cloud | PointBERT [55] | 100 | ✓ | 69.35 | 65.74 |

Table 2. Ablation Study with different foundation feature (FM) modality encoders. The column *Channel Size* stands for foundation feature channel size. The column *Upsample* indicates whether upsampling convolutions are used.

### Multi-modal soft-prompt adapter experiment

| Method | *Channel* | NDS | mAP |
|---|---|---|---|
| MSMDFusion [21] | - | 69.23 | 64.50 |
| **PF3Det** Single-Level Prompts | 10 | 69.22 | 64.63 |
| | 50 | 69.68 | 65.27 |
| | 100 | **69.90** | **65.66** |
| | 150 | 69.47 | 65.16 |
| | 200 | 69.70 | 65.37 |
| | 500 | 69.79 | 65.53 |
| | 1000 | 69.83 | 65.61 |

| | | | |
|---|---|---|---|
| **PF3Det** Multi-Level Prompts | [100, 150, -, -] | **70.00** | **66.69** |
| | [100, 100, -, -] | 68.97 | 64.50 |
| | [100, 50, -, -] | 69.38 | 64.94 |
| | [150, 75, 38, 75] | 67.66 | 64.20 |
| | [50, 25, 12, 25] | 67.45 | 64.07 |

Table 3. **Detection metrics on different prompt setups.** The column *Channel* indicates the number of channels for each soft prompts. Multi-level prompts reloaded here means the final model is first loaded from a pre-trained single level prompted model and then goes through the training process. Other models in the table are trained from TransFusion [3] pre-trained weights. The results are reported in %.