MMF-GNN Detailed Architecture with Mathematical Formulas

Multi-Modal Fusion Graph Neural Network

LiDAR Points
$\mathbf{P} \in \mathbb{R}^{N \times 4}$

$\mathbf{p}_i = [x_i, y_i, z_i, I_i]^T$

$\mathbf{P}_{norm} = \text{Normalize}(\text{Filter}(\mathbf{P}))$

$N = 512$, Point Cloud

[512,4]→[512,256]

LiDAR Embedding
$4 \to 128 \to 256$

$\mathbf{E}_L = \text{LN}(\text{Linear}(\text{LN}(\text{ReLU}(\text{Linear}(\mathbf{P}_{norm})))))$

Camera Features
$\mathbf{C} \in \mathbb{R}^{64 \times 130}$

$\mathbf{C} = \text{CNN}(\mathbf{I}_{cam})$

$\mathbf{C}_{proc} = \text{Pool}(\text{Extract}(\mathbf{C}))$

64×130, Feature Map

[64,130]→[64,256]

Camera Embedding
$130 \to 256$

$\mathbf{E}_C = \text{Dropout}(\text{ReLU}(\text{LN}(\text{Linear}(\mathbf{C}_{proc}))))$

GNSS Coords
$\mathbf{G} \in \mathbb{R}^3$

$\mathbf{G} = [\text{lat}, \text{lon}, \text{alt}]^T$

$\mathbf{G}_{norm} = \frac{\mathbf{G} - \boldsymbol{\mu}_G}{\boldsymbol{\sigma}_G}$

3D Coordinates

[3]→[256]

GNSS Embedding
$3 \to 256$

$\mathbf{E}_G = \text{Dropout}(\text{LN}(\text{Linear}(\mathbf{G}_{norm})))$

IMU Sequence
$\mathbf{M} \in \mathbb{R}^{T \times 6}$

$\mathbf{m}_t = [\mathbf{a}_t^T, \boldsymbol{\omega}_t^T]^T$

$\mathbf{M}_{filt} = \text{Filter}(\text{Align}(\mathbf{M}))$

$T = 10$, Temporal Sequence

[10,6]→[10,256]

IMU Embedding
$6 \to 256 \to 256$

$\mathbf{E}_M = \text{GRU}(\text{Dropout}(\text{ReLU}(\text{LN}(\text{Linear}(\mathbf{M}_{filt})))))$

V2X Messages
$\mathbf{V} \in \mathbb{R}^{K \times 8}$

$\mathbf{v}_k = [\mathbf{s}_k^T, \mathbf{u}_k^T]^T$

$\mathbf{V}_{valid} = \text{Validate}(\text{Parse}(\mathbf{V}))$

$K = 5$, Vehicle States

[5,8]→[5,256]

V2X Embedding
$8 \to 128 \to 256$

$\mathbf{E}_V = \text{Linear}(\text{Dropout}(\text{ReLU}(\text{LN}(\text{Linear}(\mathbf{V}_{valid})))))$

Multi-Modal Fusion Formula
$\mathbf{H}_{fused} = \text{Concat}([\mathbf{E}_L, \mathbf{E}_C, \mathbf{E}_G, \mathbf{E}_M, \mathbf{E}_V]) \in \mathbb{R}^{(N_L + N_C + 1 + T + K) \times 256}$

Transformer Block 1

Multi-Head Attention
$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{H}\mathbf{W}_Q, \mathbf{H}\mathbf{W}_K, \mathbf{H}\mathbf{W}_V$

$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$

$\text{MultiHead}(\mathbf{H}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_8)\mathbf{W}_O$

Feed Forward Network
$\text{FFN}(\mathbf{H}) = \max(0, \mathbf{H}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$

$\mathbf{W}_1 \in \mathbb{R}^{256 \times 1024}, \mathbf{W}_2 \in \mathbb{R}^{1024 \times 256}$

with Dropout($p = 0.2$) and GELU activation

Residual Connection & Layer Norm
$\mathbf{H}^{(l+1)} = \text{LayerNorm}(\mathbf{H}^{(l)} + \text{Sublayer}(\mathbf{H}^{(l)}))$

$\text{LayerNorm}(\mathbf{x}) = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}$

$\boldsymbol{\mu} = \frac{1}{d}\sum_{i=1} x_i, \boldsymbol{\sigma}^2 = \frac{1}{d}\sum_{i=1}(x_i - \mu)^2$

[N,256]→[N,256]

Transformer Block 2

Multi-Head Attention
$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{H}\mathbf{W}_Q, \mathbf{H}\mathbf{W}_K, \mathbf{H}\mathbf{W}_V$

$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$

$\text{MultiHead}(\mathbf{H}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_8)\mathbf{W}_O$

Feed Forward Network
$\text{FFN}(\mathbf{H}) = \max(0, \mathbf{H}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$

$\mathbf{W}_1 \in \mathbb{R}^{256 \times 1024}, \mathbf{W}_2 \in \mathbb{R}^{1024 \times 256}$

with Dropout($p = 0.2$) and GELU activation

Residual Connection & Layer Norm
$\mathbf{H}^{(l+1)} = \text{LayerNorm}(\mathbf{H}^{(l)} + \text{Sublayer}(\mathbf{H}^{(l)}))$

$\text{LayerNorm}(\mathbf{x}) = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}$

$\boldsymbol{\mu} = \frac{1}{d}\sum_{i=1} x_i, \boldsymbol{\sigma}^2 = \frac{1}{d}\sum_{i=1}(x_i - \mu)^2$

[N,256]→[N,256]

Transformer Block 3

Multi-Head Attention
$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{H}\mathbf{W}_Q, \mathbf{H}\mathbf{W}_K, \mathbf{H}\mathbf{W}_V$

$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$

$\text{MultiHead}(\mathbf{H}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_8)\mathbf{W}_O$

Feed Forward Network
$\text{FFN}(\mathbf{H}) = \max(0, \mathbf{H}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$

$\mathbf{W}_1 \in \mathbb{R}^{256 \times 1024}, \mathbf{W}_2 \in \mathbb{R}^{1024 \times 256}$

with Dropout($p = 0.2$) and GELU activation

Residual Connection & Layer Norm
$\mathbf{H}^{(l+1)} = \text{LayerNorm}(\mathbf{H}^{(l)} + \text{Sublayer}(\mathbf{H}^{(l)}))$

$\text{LayerNorm}(\mathbf{x}) = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}$

$\boldsymbol{\mu} = \frac{1}{d}\sum_{i=1} x_i, \boldsymbol{\sigma}^2 = \frac{1}{d}\sum_{i=1}(x_i - \mu)^2$

[N,256]→[N,256]

Transformer Block 4

Multi-Head Attention
$\mathbf{Q}, \mathbf{K}, \mathbf{V} = \mathbf{H}\mathbf{W}_Q, \mathbf{H}\mathbf{W}_K, \mathbf{H}\mathbf{W}_V$

$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}$

$\text{MultiHead}(\mathbf{H}) = \text{Concat}(\text{head}_1, \ldots, \text{head}_8)\mathbf{W}_O$

Feed Forward Network
$\text{FFN}(\mathbf{H}) = \max(0, \mathbf{H}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2$

$\mathbf{W}_1 \in \mathbb{R}^{256 \times 1024}, \mathbf{W}_2 \in \mathbb{R}^{1024 \times 256}$

with Dropout($p = 0.2$) and GELU activation

Residual Connection & Layer Norm
$\mathbf{H}^{(l+1)} = \text{LayerNorm}(\mathbf{H}^{(l)} + \text{Sublayer}(\mathbf{H}^{(l)}))$

$\text{LayerNorm}(\mathbf{x}) = \frac{\mathbf{x} - \boldsymbol{\mu}}{\boldsymbol{\sigma}} \odot \boldsymbol{\gamma} + \boldsymbol{\beta}$

$\boldsymbol{\mu} = \frac{1}{d}\sum_{i=1} x_i, \boldsymbol{\sigma}^2 = \frac{1}{d}\sum_{i=1}(x_i - \mu)^2$

Global Pooling with Attention
$\boldsymbol{\alpha} = \text{softmax}(\tanh(\text{LN}(\mathbf{H}_{final}\mathbf{W}_{pool})))$

$\mathbf{h}_{global} = \sum_{i=1}^{N} \alpha_i \mathbf{h}_i \in \mathbb{R}^{256}$

$\mathbf{W}_{pool} \in \mathbb{R}^{256 \times 1}$, Dropout($p = 0.2$)

Output Head Transformation
$\mathbf{z} = \text{LN}(\text{GELU}(\text{LN}(\mathbf{h}_{global}\mathbf{W}_{out1})))\mathbf{W}_{out2}$

$\mathbf{W}_{out1} \in \mathbb{R}^{256 \times 256}, \mathbf{W}_{out2} \in \mathbb{R}^{256 \times 256}$

Output: $\mathbf{z} \in \mathbb{R}^{256}$ (Fused Representation)

MMF-GNN Mathematical Specifications:

- Total Parameters: $\theta_{MMF} = 8.2M$ (52.9% of total)
- Input Dimensions: $\mathbf{X} = \{\mathbf{P}, \mathbf{C}, \mathbf{G}, \mathbf{M}, \mathbf{V}\}$
- Embedding Dimension: $d_{model} = 256$
- Transformer Blocks: $L = 4$ with Multi-Head Attention
- Attention Heads: $h = 8$, Head Dimension: $d_k = d_v = 32$
- Feed Forward Dimension: $d_{ff} = 1024$
- Dropout Rate: $p_{drop} = 0.2$ (Attention + FFN)
- Layer Normalization: Pre-norm with $\varepsilon = 1e-6$
- Output: $\mathbf{z}_{MMF} \in \mathbb{R}^{256}$ (Fused Representation)
- Computational Complexity: $\mathcal{O}(N^2 \cdot d + N \cdot d^2)$