Spatial Object Learning & Integrated Dimensions Multi-Modal Object Classification with RGB, Depth, and Geometry

Moses Adewolu, Terrence An Malayvong, Zihan Chen, Felipe Bergerman

June 15, 2025

1 Project Summary

This project addresses the challenge of object classification in real-world 3D environments by leveraging complementary data modalities: RGB images, depth maps, and 3D geometry. Traditional classification pipelines often rely on RGB data alone, which limits spatial understanding and robustness to occlusions and lighting variations. Our goal is to develop a multi-modal classification framework that fuses visual, depth, and geometric data into a unified, high-performing model. We will implement and evaluate different fusion strategies, benchmark performance across standard datasets, and propose an architecture optimized for multi-modal learning.

2 Approach

Our approach integrates three key data modalities:

- **RGB:** Captures color, texture, and fine-grained details; processed using ResNet or Vision Transformers (ViT).
- **Depth:** Provides spatial cues; represented as normalized depth maps and processed with CNN-based encoders.
- **3D Geometry:** Encoded as point clouds or voxel grids; processed via PointNet++ or volumetric CNNs.

We will evaluate the following fusion strategies:

1. **Early Fusion:** Concatenation of raw modalities before feature extraction.
2. **Late Fusion:** Independent feature extraction followed by decision-level integration.
3. **Cross-Attention Fusion:** Transformer-based modules to learn shared representations across modalities.

The framework will be implemented in PyTorch and trained on datasets with RGB-D and 3D annotations.

3 Related Work

Prior work has typically focused on RGB-D fusion or geometry-based learning in isolation. RGB-D classification models such as Deep Fusion Networks (Deng et al., 2020) utilize learned attention for

combining RGB and depth inputs. PointNet++ has been widely adopted for 3D point cloud processing. More recent work explores cross-modal attention using transformer architectures, enabling dynamic interaction between modalities. Our work builds upon these foundations by integrating all three data types in a unified framework and systematically evaluating the effectiveness of fusion strategies.

4 Contribution and Novelty

Our project differs from previous work in several key ways:

- Integration of RGB, depth, and 3D geometry within a single architecture.
- Comparative analysis of fusion strategies, including early, late, and transformer-based fusion.
- Extensive benchmarking across multiple datasets using standardized protocols.
- Ablation studies that quantify the contribution of each modality to overall performance.

5 Datasets and Experimental Setup

We will use the following publicly available datasets that provide aligned RGB, depth, and 3D geometry:

- **ModelNet40:** CAD-based 3D object dataset; RGB and depth maps rendered from mesh data.
- **SUN RGB-D:** Indoor scenes with aligned RGB, depth, and 3D point cloud annotations.
- **ScanNet:** RGB-D video sequences with full 3D reconstructions and semantic labels.

All models will be trained using PyTorch. Data preprocessing includes normalization, augmentation (e.g., rotation, jitter), and spatial alignment across modalities. We will use Open3D for visualization and diagnostics.

6 Algorithms and Training

We will implement a modular architecture combining the following components:

- CNN backbones (e.g., ResNet-18) for RGB and depth modalities.
- PointNet++ for 3D point cloud input.
- Transformer fusion modules for cross-modal interaction.
- Optimizers: AdamW with learning rate scheduling.
- Loss Function: Cross-entropy for classification.

Data augmentation will include random cropping, flipping, point dropout, and noise injection to improve generalization.

7 Evaluation

Evaluation will be both quantitative and qualitative:

- **Metrics:** Accuracy, per-class F1 score, and confusion matrices.
- **Visualizations:** t-SNE plots of learned embeddings, attention heatmaps, and example predictions.
- **Baselines:** Single-modality and dual-modality (RGB-D) models for comparison.

8 Risks and Mitigation

- **Modality Misalignment:** Inaccurate registration between RGB, depth, and 3D geometry. We will apply calibration and alignment heuristics.
- **Computational Demand:** Multi-branch and 3D models are compute-intensive. We'll leverage Georgia Tech's GPU cluster and optimize memory usage.
- **Fusion Design Complexity:** Cross-modal transformer design may be non-trivial. We'll begin with well-established variants and iterate.

9 Ethical Considerations

The use of RGB-D data, particularly in indoor settings, raises privacy and surveillance concerns. We will ensure that any person-identifiable data is anonymized or masked. Additionally, we will assess class imbalance and potential dataset biases, and report on fairness across object categories. Our project will be accompanied by a statement of limitations and responsible usage.

10 Team Members

Moses Adewolu, Terrence An Malayvong, Zihan Chen, Felipe Bergerman

11 References

[1] Qi et al., PointNet++: Deep hierarchical feature learning on point sets, NeurIPS 2017.

[2] Deng et al., Deep fusion networks for RGB-D object recognition, ICRA 2020.

[3] Song et al., SUN RGB-D: A RGB-D scene understanding benchmark suite, CVPR 2015.

[4] Armeni et al., 3D semantic parsing of large-scale indoor spaces, CVPR 2016.

[5] Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition, ICLR 2021.

[6] Qi et al., Frustum PointNets for 3D Object Detection from RGB-D Data, CVPR 2018.

[7] Loghmani et al., Recurrent Convolutional Fusion for RGB-D Object Recognition, ECCV 2018.

[8] Wang et al., Correlated and Individual Multi-Modal Deep Learning (CIMDL) for RGB-D Recognition, CVPR 2016.

[9] Wang and Gong, Adaptive Fusion for RGB-D Salient Object Detection, TIP 2019.

[10] Zeng et al., RGB-D Object Recognition Using Multi-Modal Deep Neural Network and Dempster–Shafer Evidence Theory, ICRA 2019.

[11] Chen et al., Exploring RGB + Depth Fusion for Real-Time Object Detection, arXiv 2019.

[12] Qi et al., PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation, CVPR 2017.

[13] Wu et al., Point-GR: Graph Residual Point Cloud Network for 3D Classification, arXiv 2024.

[14] Wang et al., REGNet: Ray-Based Enhancement Grouping for 3D Object Detection, ICCV 2021.

[15] Zhang et al., The Fusion Strategy of 2D and 3D Information Based on Deep Learning: A Review, Information Fusion 2021.

[16] Zhao et al., Point Transformer, ICCV 2021.