Large-Scale Enhancements for PointNeXt in 3D Scene Understanding

Progress Report

E.A.B.T.Edirisinghe - 200151P

Literature Review

3D scene understanding is fundamental for robotics, AR/VR, and autonomous systems, where accurate semantic segmentation of large-scale indoor spaces is crucial. Point clouds were voxelized or projected to 2D views in earlier techniques, but these methods lost fine-grained geometry and produced discretization artefacts.

Point-based networks revolutionized the field by directly operating on unordered point sets. PointNet [1] introduced permutation-invariant pointwise MLPs with global pooling but lacked local structure awareness. PointNet++ [2] addressed this with hierarchical local neighborhood feature extraction, becoming a foundational architecture.

Subsequent research explored more powerful operators. KPConv [3] proposed kernel point convolutions to capture local geometric patterns. Sparse convolutional frameworks such as MinkowskiNet [4] enabled scalable training on massive indoor/outdoor scans. Transformer-based approaches, including Point Transformer [5] and Stratified Transformer [6], modeled long-range dependencies effectively, while RandLA-Net [7] demonstrated efficient large-scale point processing through random sampling and lightweight local aggregation.

Most recently, PointNeXt [8] revisited PointNet++ with modern training recipes and scaling strategies. By integrating inverted residual MLP blocks, separable MLPs, and strong data augmentation, PointNeXt achieved state-of-the-art performance on S3DIS, reporting 74.9% mIoU and 71.5% mIoU (Area-5 split). Crucially, the study showed that training strategies are as impactful as architectural novelty.

Despite its strong results, PointNeXt faces limitations in large-scale scene processing. Handling millions of points per indoor scan is memory-intensive and challenges throughput. Sliding-window inference can introduce boundary inconsistencies, while long-range context is often underutilized. Addressing these bottlenecks through scalable training, efficient neighborhood grouping, and enhanced context fusion motivates this research.

Methodology Outline

This research will focus on enhancing PointNeXt for large-scale indoor scene segmentation on the S3DIS benchmark.

1. Dataset

- Stanford Large-Scale 3D Indoor Spaces (S3DIS) dataset.
- 6 indoor areas, 271 rooms, 13 semantic classes.
- Standard protocols: Area-5 split for validation, 6-fold cross-validation for full benchmarking.

2. Baseline Model

- PointNeXt-S / PointNeXt-B .
- Full training pipeline replicated from official implementation [1].

3. Enhancement Strategies

The project will introduce systematic, incremental improvements in the following directions:

- Hyperparameter Optimization: fine-tuning LR, weight decay, smoothing, EMA/SWA.
- Loss Function Enhancements: explore class-balanced cross-entropy, Lovász-Softmax, and boundary-aware auxiliary losses.
- Architecture Tweaks: additional InvResMLP blocks at coarse stage, adaptive Δp normalization.
- Data Augmentation Improvements: scheduled color-drop probability, addition of normal vectors.

4. Training Procedure

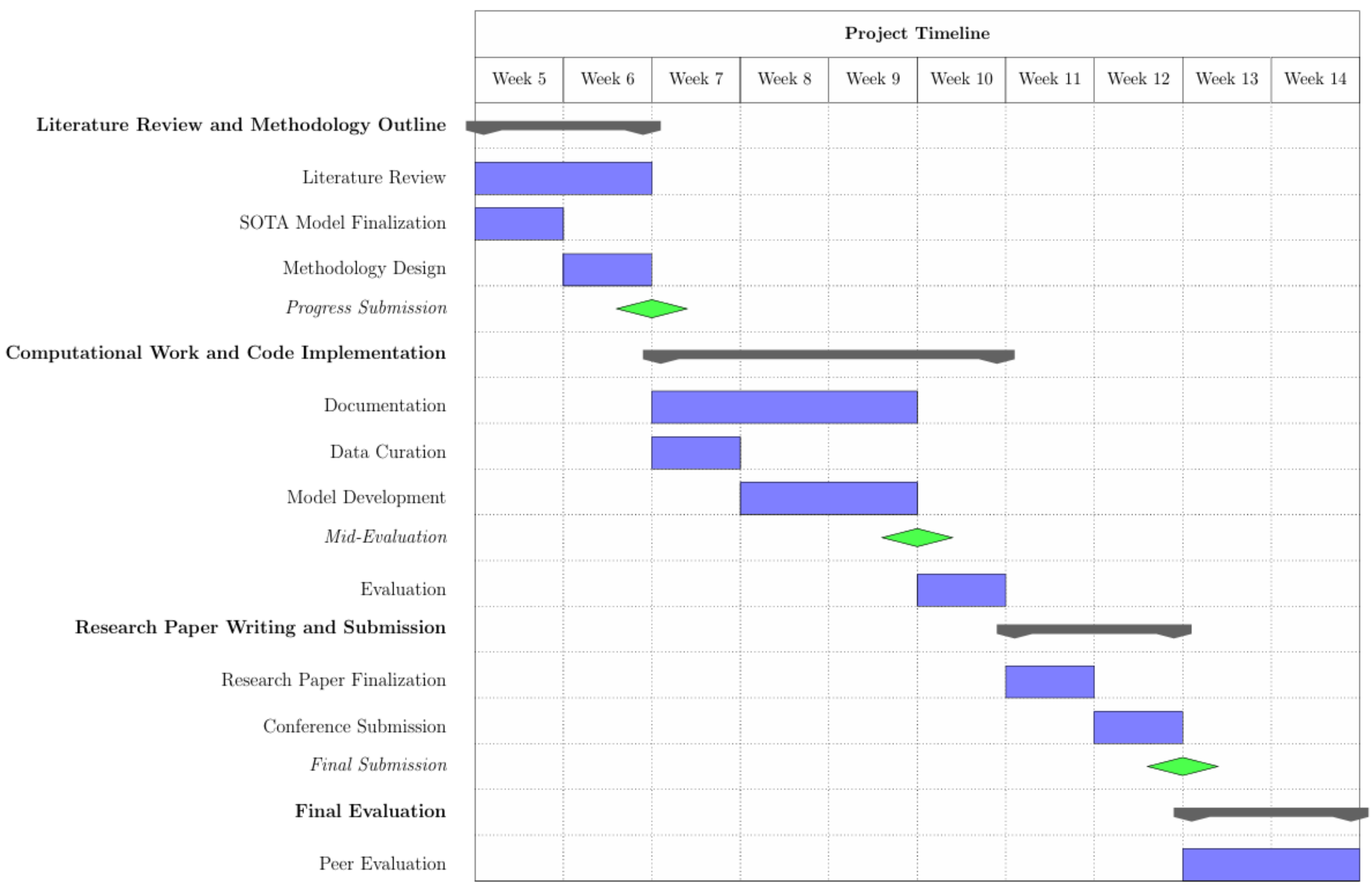
- Optimizer: AdamW with cosine LR scheduling and warmup.
- Label smoothing applied.
- Mixed precision training with gradient checkpointing for scalability.

5. Evaluation Metrics

- Primary: mean Intersection over Union (mIoU), Overall Accuracy (OA).

- Secondary: throughput (samples/sec), GPU memory usage.
- Robustness tests: varying voxel size, color removal, reduced point budgets.

Project Timeline



References

[1] Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017). PointNet: Deep learning on point sets for 3D classification and segmentation. CVPR.
[2] Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017). PointNet++: Deep hierarchical feature learning on point sets in a metric space. NeurIPS.
[3] Thomas, H., et al. (2019). KPConv: Flexible and deformable convolution for point clouds. ICCV.
[4] Choy, C., Gwak, J., & Savarese, S. (2019). MinkowskiNet: Generalized sparse convolutional neural networks. CVPR.
[5] Zhao, H., et al. (2021). Point Transformer. ICCV.
[6] Lai, X., et al. (2022). Stratified Transformer for efficient 3D point cloud learning. CVPR.
[7] Hu, Q., et al. (2020). RandLA-Net: Efficient semantic segmentation of large-scale point clouds. CVPR.
[8] Qian, G., et al. (2022). PointNeXt: Revisiting PointNet++ with improved training and scaling strategies. NeurIPS.
[9] Armeni, I., et al. (2016). 3D Semantic parsing of large-scale indoor spaces. CVPR.