MixFeat3D: Improving 3D Semantic Segmentation with Feature Mixing

Chandrahas Aroori
Department of Computer Science
Boston University
Boston, MA, 02215
charoori@bu.edu

Sai Tejaswini Junnuri
Department of Computer Science
Boston University
Boston, MA, 02215
jteja@bu.edu

Deepti Ghadiyaram
Department of Computer Science
Boston University
Boston, MA, 02215
dghadiya@bu.edu

Abstract

There has been a lot of work involving using re-trained 2D vision models with 3D frameworks. We chose to build upon *Lexicon3D* Man et al. [2024], in which an ablation study demonstrated improved outcome in the semantic segmentation task by creating an ensembe of models involving LSeg, Stable Diffusion, and Swin3D. We investigated feature based fusion strategies, including additive and interleaved approaches, to refine the Mixture of Features paradigm. By extending the concept to encompass semantic segmentation, our goal is to identify optimal fusion techniques that outperform existing benchmarks, leveraging diverse pre-trained embeddings to improve understanding in 3D spaces.

1 Motivation

3D Semantic Segmentation using existing 2D models present challenges due to the fact that the models were not trained specifically on 3D data and use 2D abstractions to understand the 3D space.

In Lexicon3D, an ablation study conducted showed that combining three distinct & equally distinct feature embeddings—LSeg, StableDiffusion, and Swin3D—enhanced existing performance in semantic segmentation. Our conclusion from this was that leveraging multiple embeddings captures complementary information, enriching the understanding of visual data.

To build upon this we turned to,*Eyes Wide Shut* Tong et al. [2024], which explored Mixture of Features (MoF) techniques which have been shown to improve visual grounding in multimodal large language models (MLLMs) without degrading their instruction-following capabilities. Although we do not need the instruction-following capabilities, we found that these findings highlight the potential of MoF approaches synthesizing strengths from diverse feature representations.

We believe that individual models have distinct strengths, and by combining them, we can create a more robust model than any single one alone.

So we used the following types of models:

- **Image Models:** DINOv2 and LSeg are used for their strong semantic understanding.
- **Image-Text Models:** CLIP, while useful, does not show strong results in our context.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

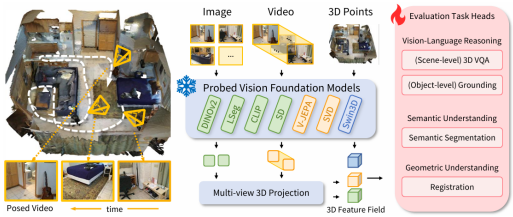


Figure 1: Framework as proposed by Lexicon3D

- **Diffusion-Based Models:** Stable Diffusion excels at preserving the local geometry and textures of scenes, thanks to generation-guided pretraining.

Motivated by these insights, we aimed to improve Lexicon3D's results by applying MoF techniques to combine semantic segmentations. Our work performs the following:

1. Extend the MoF technique to Lexicon3D's architecture.
2. Evaluate the impact of ensemble embeddings on Semantic 3D Segmentation
3. Visualize and interpret the enhanced feature representations in 3D space.

2 Related Work

Current approaches, including Lexicon3D, utilize foundational vision models to extract features from posed images, videos, and 3D point clouds. These features are integrated into a unified 3D embedding, supporting tasks like semantic segmentation. While Lexicon3D leverages a multi-view 3D projection module to merge 2D and 3D data, there is still potential to improve both performance and robustness.

The experimental results from Lexicon3D Man et al. [2024] indicate that image-based encoders, such as DINOv2 and LSeg, outperform video and 3D encoders in semantic segmentation tasks, likely due to their superior ability to capture fine-grained semantic features during training. By contrast, video encoders risk over-smoothing multi-view data, potentially hindering semantic understanding. Furthermore, the limited availability of 3D data for training foundation models contributes to the lower performance observed in 3D encoders like Swin3D.

3 Method

Our approach is based on the methodology provided by Lexicon3D, where all types of features are unified in the form of point clouds; therefore, semantic labels are predicted for each point within the point cloud in our setting. More specifically, given a 3D scene in the form of multi-view images and point clouds, the objective in this task is to predict the semantic label for every point in the cloud.

We obtain different sized embeddings from each model. We then project these different sized embeddings to a consistent embeddings space/ To merge different sized embeddings, we employ a linear probe using a single linear layer followed by a Sigmoid function to predict the probability distribution for each point in the cloud.

The features are then merged via either Additive Mixture or Interleaved Mixture of Featurees.

3.1 Feature Standardization & Mixing

Additive Mixture of Features: In this approach, features from different models are combined by directly adding them together. This allows for a richer feature space, where the strengths of each model contribute to a more comprehensive representation.

Interleaved Mixture of Features: In contrast, the interleaved mixture involves alternating or interleaving features from different models at different stages or layers of the model's architecture. This method allows the model to learn to utilize complementary features in a more dynamic and flexible way. However there may be some amount of information loss.1

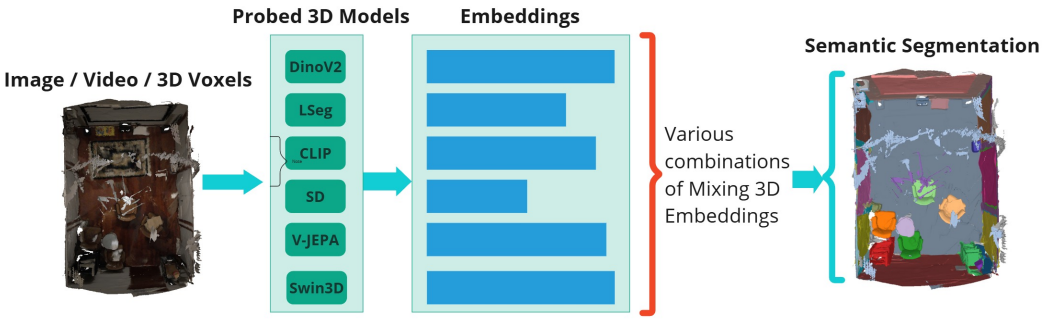


Figure 2: Methodology for 3D Feature Mixing

(a) Interleaved Mixture

(b) Additive Mixture

Figure 3: Feature Mixing

4 Experiments and Results

4.1 Experiment setup

Models used, training details:

Datasets: We have conducted our experiments on ScanNet V2 3D segmentation dataset. The dataset has 1,201 scenes for training and 312 scenes for validation. We have taken a subset of the dataset with 405 scenes for training and 62 scenes for validation.

Models: CLIP, LSEG, DINO, SD along with their combinations.

Training Setup: The models or combinations of them are passed through adapters individually to get to consistent length and various fusion strategies ar applied over them and the resulting embeddings are passed through a single linear MLP layer to get the predictions.

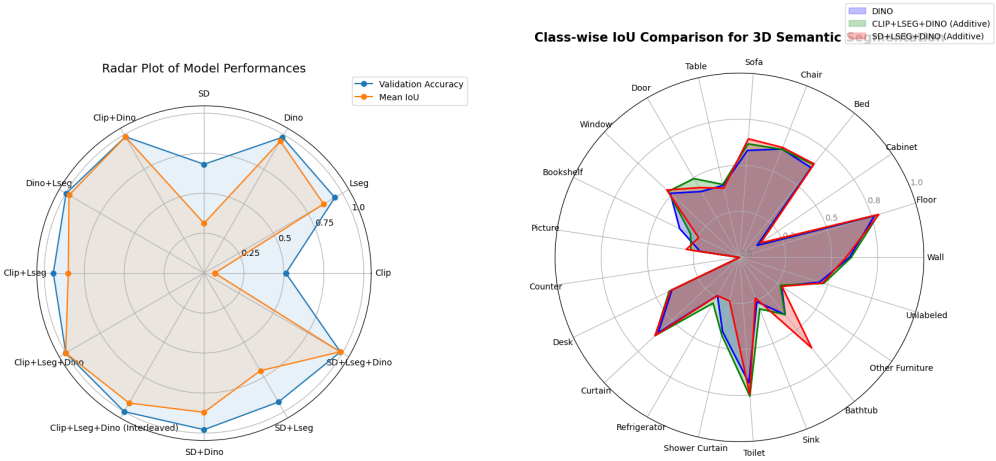
Evaluation Metrics: Validation accuracy and Mean IoU.

4.2 Results

- CLIP + LSEG + DINO (Additive): Has best overall performance with a Validation Accuracy of 71.81% and Mean IoU of 0.4462.
- SD + LSEG + DINO (Additive): Has competitive performance with a Validation Accuracy of 70.86% and Mean IoU of 0.4391.
- Additive fusion is comparatively better than interleaved fusion in this case as additive captures complementary strengths, whereas interleaved fusion likely under performs due to feature blending issues.
- The standalone models (CLIP, LSEG, DINO, SD) showcase individual strengths in global context, semantic alignment, and texture details but lack overall robustness. DINO shows the best overall individual performance.
- The performance cannot be compared to Lexicon 3D paper as the inference code and methodology is not provided and it is reported for the whole dataset.

4.3 Qualitative Analysis

- Strength of CLIP + LSEG + DINO:



(a) Model Wise Overall Performance

(b) Class Wise Segmentation Performance

Figure 4: 3D Segmentation Performance

Shows better segmentation in global-context-heavy classes, particularly Shower Curtain (IoU: 0.4319) and Chair (IoU: 0.6300) as compared to the model with SD

Outperforms SD-based combinations in classes requiring structural understanding and semantic context alignment.

- Strength of SD + LSEG + DINO:

  Better performance compared to the above model in texture-heavy classes like Cabinet (IoU: 0.1402) and Bathtub (IoU: 0.6280)

  Highlights SD's effectiveness in capturing localized and detailed features as compared to CLIP based combinations

4.4 Current weaknesses and failure cases

Additive and interleaved combinations can sometimes dilute model-specific strengths, resulting in inconsistent performance across complex or texture-heavy classes.

Sparse Coverage: Varying 2D-to-3D mapping masks across models lead to missing embeddings for some points, affecting overall segmentation accuracy.

5 Introspection

What was the most fun part of working on this project?

- Chandrahas: Trying to get the code provided by Lexicon3D working was the most fun part.
- Tejaswini: Learning to understand these strengths of these 2D vision models and rationalize their performances when mixed together

What was the most challenging part of working on this project?

- Chandrahas: Working with 3D data was challenging for me.
- Tejaswini: Getting the 3D fusion features for the dataset.

What was your key takeaway from this project?

- Chandrahas: Trying to build atop a given paper is difficult and requires thorough understanding of the original paper and the weaknesses.
- Tejaswini: How the 2D vision models work and how model ensembling is powerful esp for models with varied training objectives.

References

Yunze Man, Shuhong Zheng, Zhipeng Bao, Martial Hebert, Liang-Yan Gui, and Yu-Xiong Wang. Lexicon3d: Probing visual foundation models for complex 3d scene understanding, 2024. URL https://arxiv.org/abs/2409.03757.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

6 Appendix

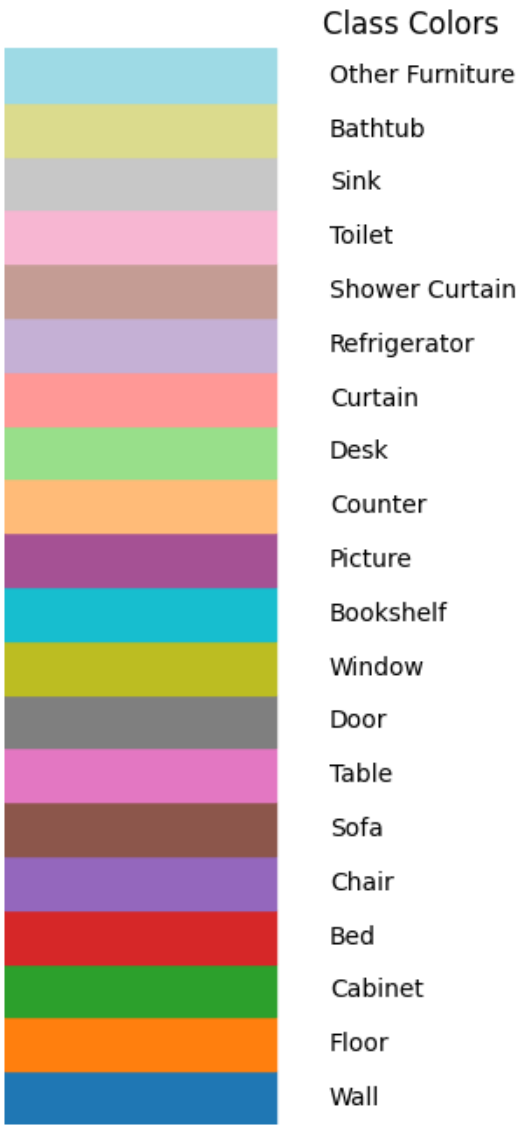


Figure 5: Label Image

Figure 6: CLIP+DINO+LSEG for a scene where Chair, Shower Curtain is better with predicted segmentation on the right

Figure 7: SD+DINO+LSEG for a scene where Other Furniture, Bathtub is better with predicted segmentation on the right