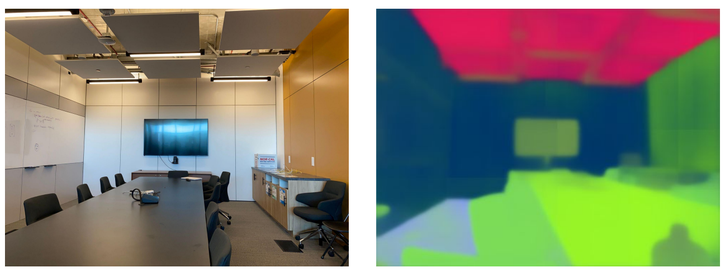
Smoothness in Distilled Feature Fields

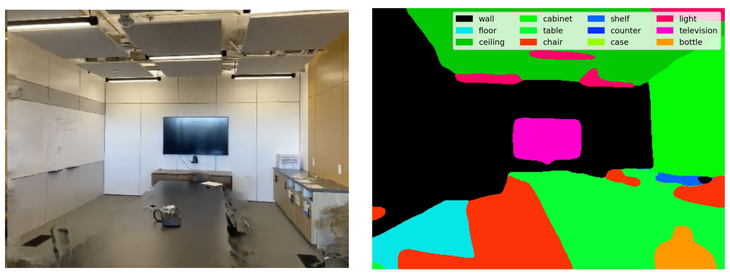
Sruthi Srinivasan, Umangi Jain
Department of Computer Science, University of Toronto

Motivation

- Feature fields for 3D scenes are representations encoding dense information about the scene, useful for diverse downstream applications
- Feature field distillation utilizes knowledge from large scale 2D image extractors
- However, naive distillation can contain unwanted high-frequency artifacts, hampering fine-grained control and resulting in imprecise scene decomposition
- In this work, we generate smoother DFFs using segmentation masks and explicit regularizers, and test the features on scene editing



Visualization of LSeg feature



Application of feature fields in editing (left) and segmentation (right)

Related Work

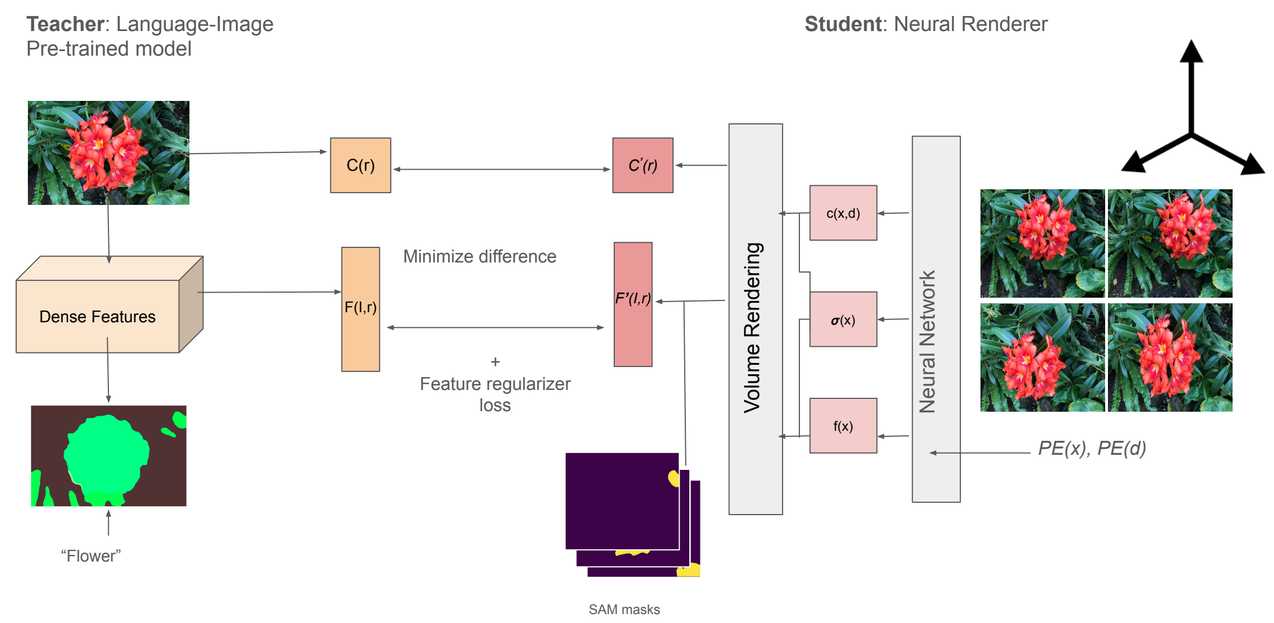
- Unlike label distillation from 2D images to 3D scenes, DFFs significantly increases applicability
- Recent success in rendering feature representations for 3D scenes has facilitated label-free scene understanding and decomposition[1][2]
- Kobayashi *et al.*[1] addressed the noise in DFFs and proposed an ad hoc coarse sampling
- However, high-frequency artifacts still persist, indicating that DFFs can benefit from smoothness
- Emerging progress in 2D image segmentation, such as Segment Anything Model (SAM)[4], can provide masks, that can facilitate edge-preserving blurring
- Total Variation (TV) and Bilateral filtering are also useful baselines to enforce smoothness while preserving edges

References

[1] Kobayashi *et al.,* Decomposing nerf for editing via feature field distillation, 2022
[2] Tschernezki *et al.,* Neural feature fusion fields: 3d distillation of self-supervised 2d image representations, 2022
[3] Li et al., Language-driven Semantic Segmentation, 2022
[4] Kirillov *et al.,* Segment anything, 2023
[5] Müller *et al.,* Instant Neural Graphics Primitives with a Multiresolution Hash Encoding, 2022

Proposed Methodology

- Feature fields are learnt in 3D space using a 2D pre-trained teacher model and a student model in the 3D space, optimized in conjunction with radiance field
- Hierarchical sampling in NeRF-like models induces high frequency artifacts in rendered feature, as scene decomposition is a lower spatial frequency task
- To mitigate the noise, we propose smoothing the segments of the rendered features, guided by the segmentation masks from SAM model
- We also test against anisotropic total variation and bilateral filtering baselines
- The teacher model used is LSeg[3] and neural renderer is Instant-NGP[5]



Our proposed distillation improvement framework

Experimental Results

- Qualitative comparison of the LSeg features (after PCA^{+}). Explicit smoothness regularizers and using SAM for edge-preservation shows sharper decomposition

Kobayashi et al.[1] | Ours (Total Variation) | Ours (SAM-guided)

- Downstream task of editing (extraction, deletion, colourization), using the 3D consistent features, displays smoother segments

Closest training scene | Vanilla distillation | Ours

Task: Extract "apple" and "banana"

- Learning from 2D features does not hurt the geometry (see PSNR), and the scene features have high similarity with the teacher supervision (see cosine similarity)

	PSNR	SSIM	Cos Similarity
Baseline[1]	24.05	0.5425	0.9696
Anisotropic total variation	24.31	0.5479	0.9792
Bilateral filtering	23.97	0.5372	0.9799
SAM-guided smoothening	24.11	0.5469	0.9768

- While the proposed metrics are better than the baseline, the quality of features also depends on the downstream application

$^{+}$Please not that it is not valid to compare the colours here, as they are obtained from different PCA instances