Supplementary Material: Uncovering the Inner Workings of STEGO for Safe Unsupervised Semantic Segmentation

Alexander Koenig Maximilian Schambach Johannes Otterbach

Merantix Momentum

{firstname.lastname}@merantix.com

1. Full Configuration

We showed the model configuration in detail in our main contribution. We noticed that the original STEGO configuration varies for the three datasets. Tab. 1 shows some remaining noteworthy parameters in STEGO's configuration, which are the same for all datasets.

Parameter	Value
Loader crop type	Center
Extra clusters	0
Optimizers	Adam [2]
Linear and cluster probe learning rates	0.005
Segm. head learning rate	0.0005
Segm. head dropout probability	0.1
Feature samples	11
Negative samples	5

Table 1. Remaining model configuration for STEGO. These are the original parameters from the paper, also used in our study. Only the last four parameters are specific to the training of the segmentation head – the others also apply to our DINO, PCA, and RP baselines. Hamilton *et al.*'s [1] code repository contains more information on the parameters.

2. Accuracy Results

In the main paper, we report the mIoU results of the different frameworks. Here, we also provide supplementary information on the accuracy metric in Figure 1. Our observations from our previous discussion also translate to the accuracy plots. Hence, we solely display the results for completeness without further discussion.

3. Non-linear Dimension Reduction Baseline

Our previous analysis compared STEGO with the linear dimensionality reduction techniques, PCA, RP, and the DINO baseline. Since STEGO is a non-linear projection, comparing it to a non-linear dimension reduction method is interesting. Initially, we investigated Uniform Manifold Approximation and Projection (UMAP) [4], which builds a fuzzy topological representation of the data in the original space and, via cross-entropy, searches a new representation that approximates this topology in a lower dimensional space. Despite improved scalability of the UMAP algorithm over other non-linear dimension reduction algorithms like t-SNE [6], UMAP was prohibitively expensive to compute across all datasets and different embedding dimensions (*e.g.*, there are ≈ 0.8 billion 768-dimensional ViT training tokens for the Cocostuff dataset alone). A recently proposed optimization-free and faster algorithm, Hierarchical Nearest Neighbor Embedding (h-NNE) [5], approaches the problem by first building a clustering hierarchy of the data in high dimensions. Afterward, the method hierarchically projects the data into a lower dimensional space, preserving 1-nearest neighbor relationships.

We fit h-NNE on a randomly sampled subset of 1 million ViT training tokens, project the entire training and validation set into lower dimensions, and fit the linear and cluster probes on these projected embeddings. Figure 2 shows the results for the Cityscapes dataset. In Figure 2 (a,b), the h-NNE algorithm shows similar performance on the linear probing downstream task as STEGO, PCA, and RP across a wide range of dimensions. For the unsupervised cluster probe in Figure 2 (c), we see approximately equal mIoU performance compared to the PCA baseline, while the accuracy of the cluster probe trained on the h-NNE projections in Figure 2 (d) outperforms the PCA, RP baselines, although the variance of the h-NNE results appears more significant.

In summary, these preliminary results show that the non-linear projection with the h-NNE algorithm yields little to no benefit over the linear projection methods in the tested benchmarks. However, we assume that clustering the h-NNE projected output with k-means might not be the most suitable unsupervised downstream evaluation. The h-NNE algorithm already provides a hierarchy of clusterings. Hence, in future work, one could directly map the clusters detected by h-NNE to the human-interpretable labels with the Hungarian algorithm [3].

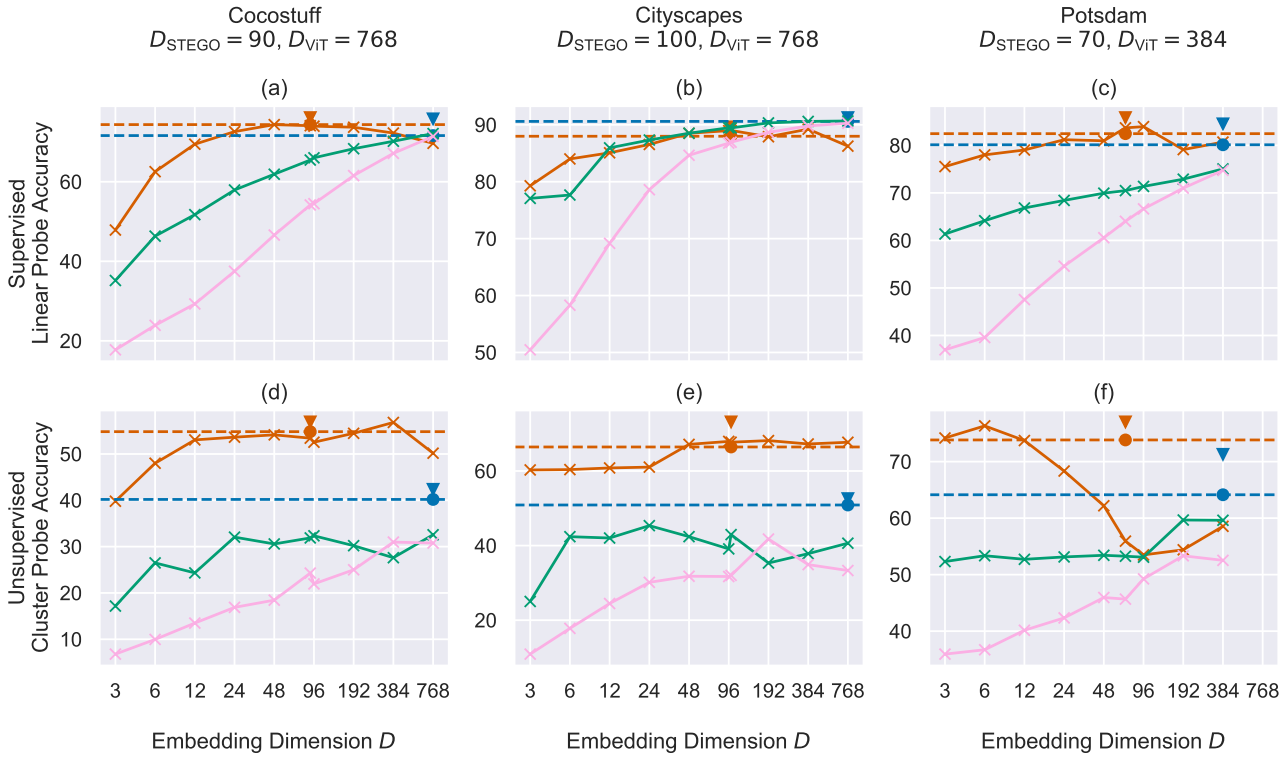

Figure 1. Validation accuracy of different dimensionality reduction techniques. Readers are referred to Figure 2 for a color-coded legend.

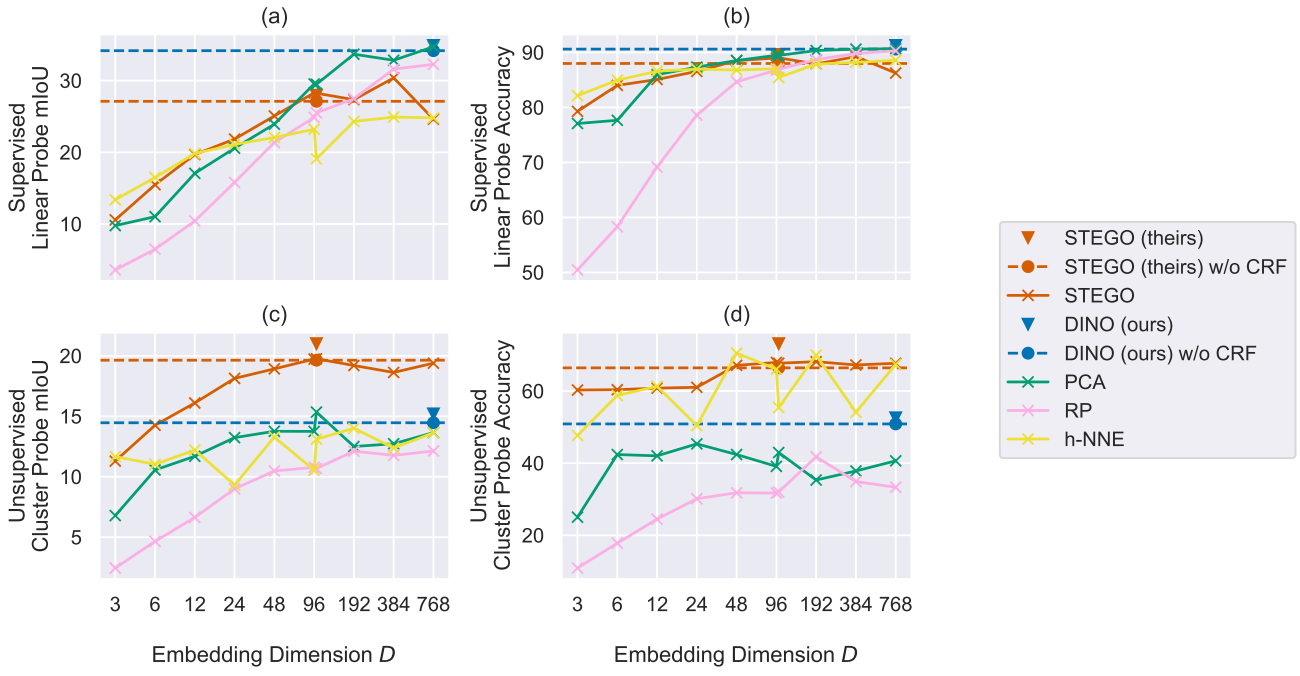

Figure 2. Cityscapes validation results from our main contribution and Figure 1 with overlaid h-NNE [5] results.

References

[1] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022. 1

[2] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 1

[3] Harold W Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 1955. 1

[4] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 2018. 1

[5] M. Saquib Sarfraz, Marios Koulakis, Constantin Seibold, and Rainer Stiefelhagen. Hierarchical nearest neighbor graph embedding for efficient dimensionality reduction. In *Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008. 1