Region-Based Representations Revisited

Supplementary Material

Table 13. Semantic Segmentation SAM Parameters

Parameter	Value
Points Per Side	32
Pred Iou Threshold	0.88
Stability Score Threshold	0.95
Stability Score Offset	1.0

6. Author Contributions

Michal developed and ran experiments for semantic segmentation and object retrieval. Ansel implemented the SLIC region generation, developed and ran experiments for semantic segmentation. Sethu and Heyi developed and ran experiments for semantic segmentation. Yao developed and ran experiments for activity classification. Yuqun developed and ran experiments for multi-view segmentation. Jae advised on region representation implementation and experimentation. Yuxiong advised on implementation and experimentation. Wilfredo advised on implementation. Derek guided the project and advised on all aspects: implementation, experimentation and paper writing. All authors contributed to paper writing.

7. Additional Experimental Parameters

We list experimental parameters and hyper-parameters for our experiments.

7.1. Semantic Segmentation

SAM parameters for the semantic segmentation experiments can be found in Table 13. Semantic segementation training hyper-parameters can be found in Table 16.

SLIC After viewing the generated superpixels with different hyperparamters for number of clusters and compactness, we chose 50 clusters with a compactness of 8 as this generated more semantically meaningful superpixels than those generated with a large number of clusters.

Table 14. **Multi-view Semantic Segmentation SAM Parameters.** Parameters not listed in the table follow the default values in SAM paper.

Parameter	Value
Points Per Side	16
Stability Score Threshold	0.85

7.2. Multi-view Semantic Segmentation

We utilize a different setting of SAM from (single-view) semantic segmentation because ScanNet [12] is less complicated and more diverse. To reduce the preprocessing time, we use a smaller "Points Per Side" number. The parameters of SAM for multi-view segmentation are shown in Table 14. Parameters not listed are the same as in Table 13.

During training, all models use an initial learning rate at 1e-5 with 50 training epochs. The optimizer is AdamW with 0 weight decay factor. Batch sizes of linear probe, transformer within images, and transformer within scenes are 256, 64, 1 respectively. Transformers have 3 layers and 8 heads, with 5 epochs of warm-up training.

7.3. Object-Based Image Retrieval

SAM regions were generated for the database images using the same parameters as the ones used for semantic segmentation (which are in Table 13). Ground truth masks from the train split of the COCO dataset [38] were used for query objects. Results are from the validation split.

Table 15. Activity Classification SAM Parameters

Parameter	Value
Points Per Side	8
Stability Score Threshold	0.85
Min Mask Region Area	500

7.4. Activity Classification

Similar to multi-view segmentation, activity-classification does not require as detailed features so several of the default SAM parameters are reduced as shown in Table 15.

During the training phase, we employed a transformer model with 3 layers and 16 heads, and trained for 40 epochs, where 2.5 were for warm-up. The learning rate was set at 1e-5, with a batch size of 32, and we utilized the AdamW optimizer, with a weight decay factor of 0.

8. Qualitative Results

Semantic Segmentation Additional qualitative results from the ADE20K dataset [60, 61] can be found in Figure 8. We show predictions from the DINOv2 patch-based model and DINOv2 region model, with regions generated by SAM and SLIC which are also shown. The effect of SAM and SLIC can be seen in the higher precision and clearer boundaries. Patch-based models undergo interpolation at the final stage resulting in uneven object segmentation.

Multi-view Semantic Segmentation Visualization of additional scene-level semantic segmentation are shown in Figure 9. We show predictions from a linear probe, transformer within image and transformer within scene. For better visualization, we only show the main 20 classes that Scan-

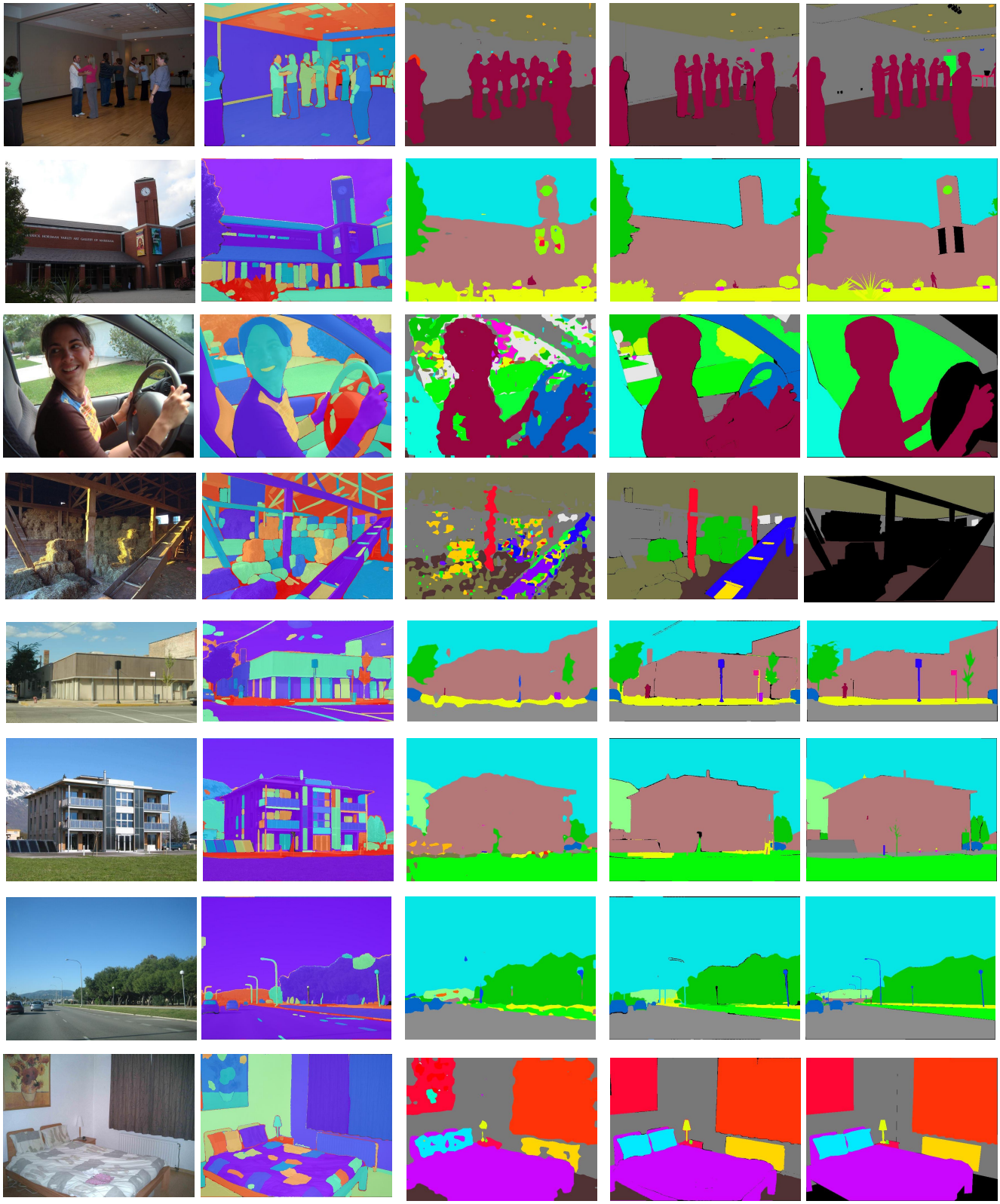

Figure 8. Semantic segmentation examples from ADE20K. The regions column shows masks from SAM and SLIC. The third column and fourth columns show pixel predictions from DINOv2 patch and region (with SAM and SLIC) based models respectively

Table 16. **Semantic Segmentation Training Hyper-Parameters** Models were trained until validation loss stopped decreasing.

Architecture	Initial LR		Batch Size		Epochs	
	Pascal-VOC	ADE20K	Pascal-VOC	ADE20K	Pascal-VOC	ADE20K
Linear (Regions)	5e-4	5e-4	32 regions	8192 regions	20	100
Linear (Patch)	1e-3	1e-3	8 images	16 images	20	20
MLP (hidden size: 1000)	5e-4	1e-4	32 regions	8192 regions	4	28
Transformer	1e-4	1e-4	2 images	2 images	4	8

Net [12] evaluate, and the remaining ones are marked as excluded labels.

Object-based Image Retrieval Visualizations of additional object-based image retrieval results can be found in Figure 10. Query objects of varying sizes are found in the database images. The second row contains an example where multiple regions are matched to the query region but only one is correct.

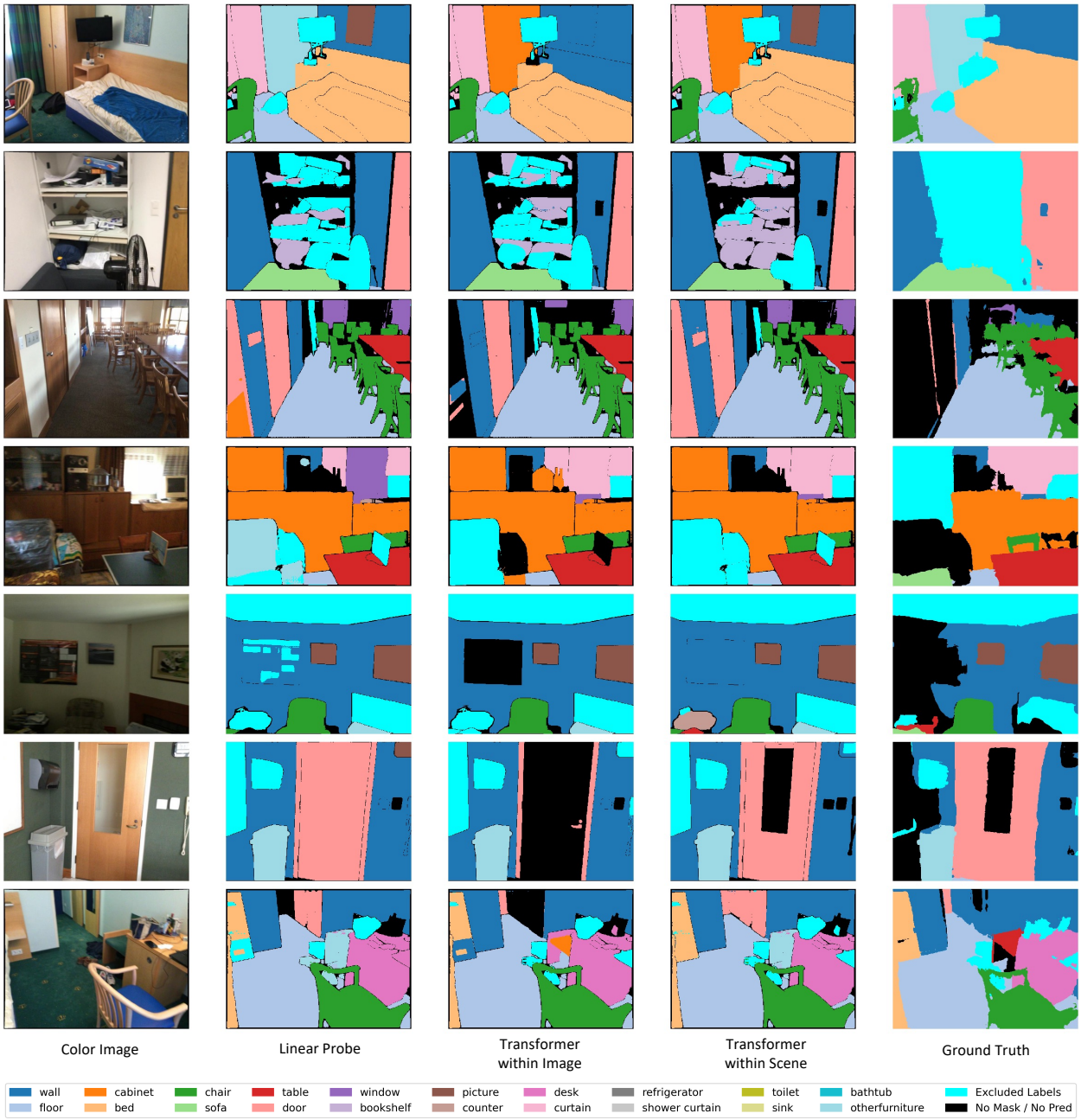

Figure 9. **Additional qualitative results for scene-level semantic segmentation.** From left to right: color images, prediction from linear prob, prediction from transformer within image, prediction from transformer within scene, and ground truths.

Query Object | Database Image | Similarity Heatmap

Figure 10. Additional object retrieval results using our region representation. An object mask can sometimes (incorrectly) match multiple regions in a database image, as shown in row 2.