(Supplementary)
Glimpse Attend and Explore: Self-Attention for Active Visual Exploration

Soroush Seifi* Abhishek Jha* Tinne Tuytelaars
PSI, ESAT, KU Leuven
Kasteelpark Arenberg 10, 3001 Leuven
firstname.lastname@esat.kuleuven.be
Code: https://github.com/soroushseifi/glimpse-attend-explore

1. Model architecture

In this section we present the detailed architecture of our network, figures 2,3,4,5. For all experiments, the input scene and glimpse sizes are fixed to $128 \times 256 \times 3$ and $48 \times 48 \times 3$. The glimpse input of the network is scaled down on locations further away from its center (retina-like glimpse), figure 1. Retina glimpses are used in previous literature to save on the total number of pixels processed from the scene [5, 6].

The prupose for channel reduction layers in all streams of the network is to bring down the memory usage of the network as well as the dimension of the features for the contrastive loss.

The ground truth stream shares all its parameters with the other streams of our network namely contrastive and self-attention streams. The training of all parameters happens through these two streams while we stop the gradient flow in the ground truth stream, figure 2. The ground truth stream is only employed during training time to generate ground truth features for the scene (F_I) in the main paper).

Contrastive module shares the reduction parameters on the bottleneck features level with the attention module and the ground truth stream. The fully connected layers in this level demand high memory making the channel reduction necessary. Finally, for classification, We employed a simple decoder that fits into both our model and the architecture proposed in [6].

Figure 1: **Retina and full resolution glimpse comparison:** Maintaining the consistency in training and evaluation with Attend and Segment [6], the input to the network is kept retina-like glimpse and not full image crops. For a retina-like glimpse the glimpse is kept sharp at the center and blurred away from the glimpse center.

*Equal contribution.

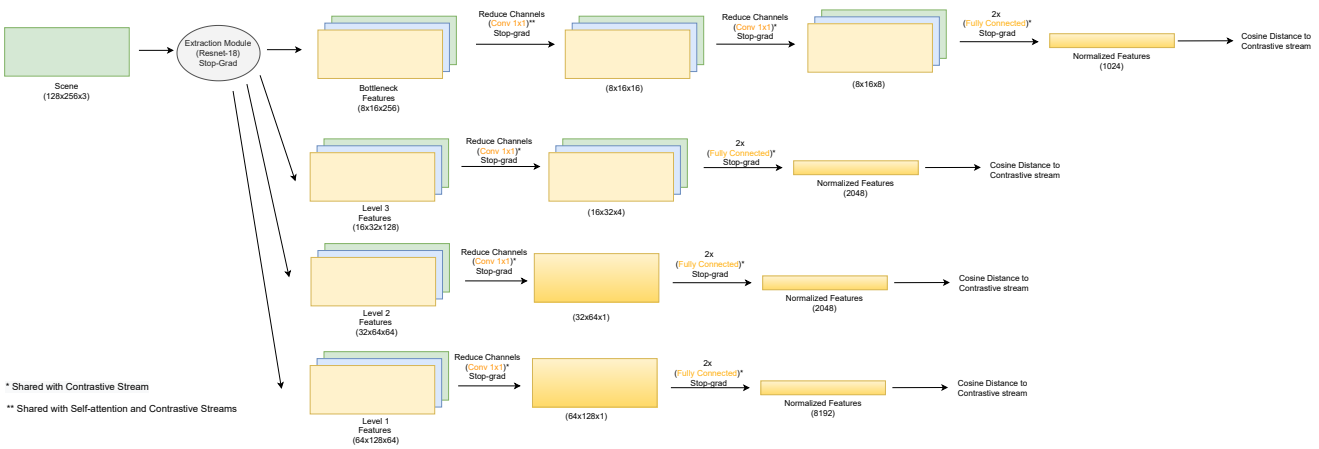


Figure 2: **Ground truth stream.**

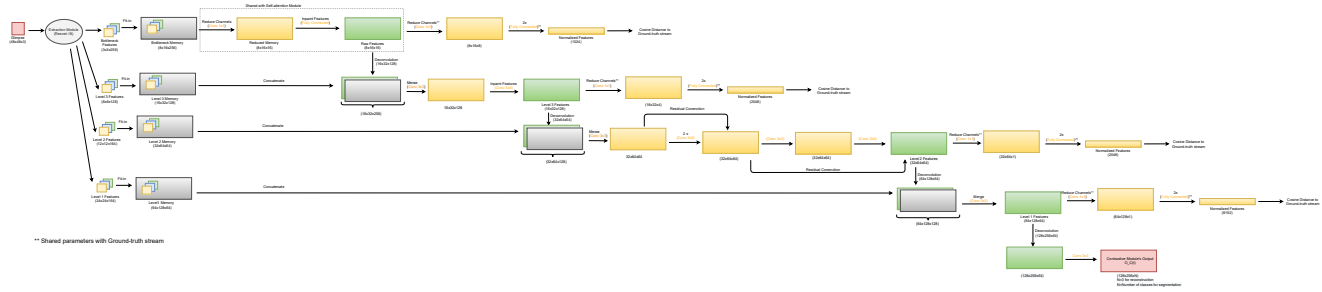


Figure 3: **Contrastive stream.**

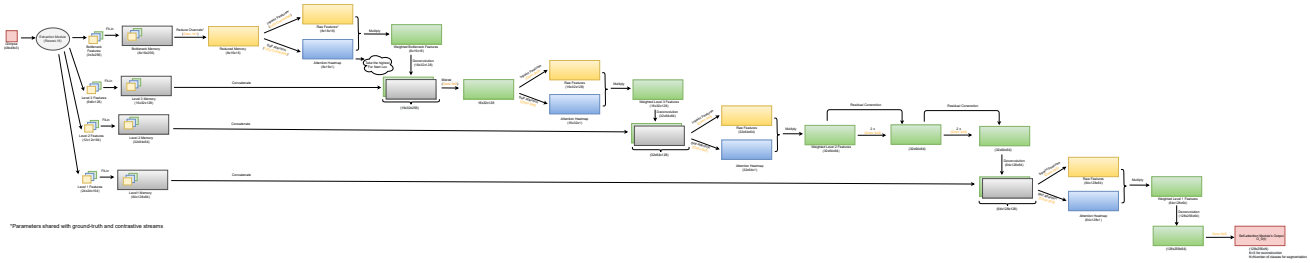


Figure 4: **Self-attention stream.**

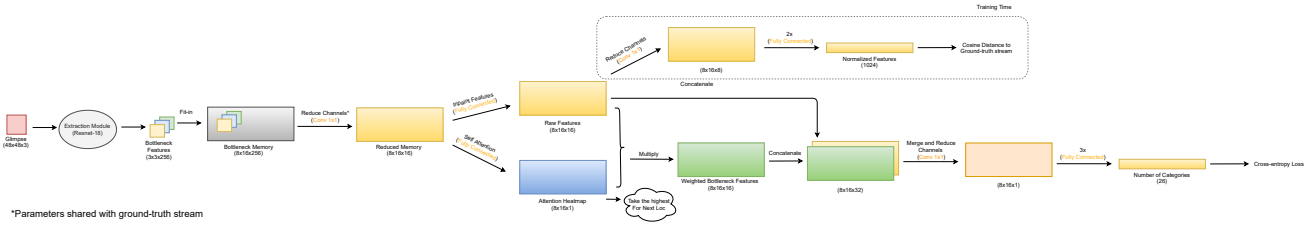


Figure 5: **Classification decoder.**

2. Average glimpse image

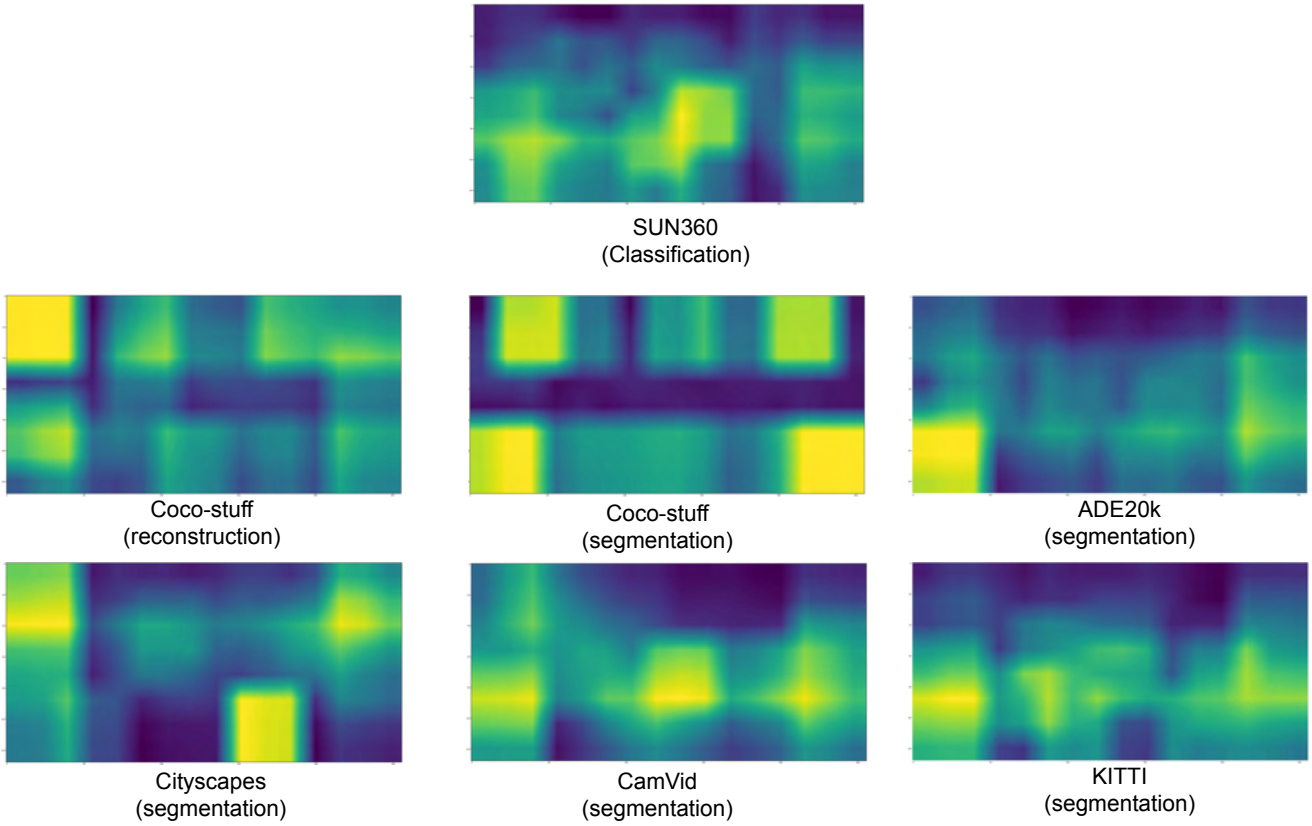


Figure 6: **Average glimpse image:** We observe that the patches with higher attendance rates are typically visited in the first steps of the exploration. We hypothesize these are the locations that the network fixes for each dataset to gain as much information about the general layout/type of the scene. For instance, the brighter corners For ADE20K [8] and COCO-Stuff [2] are attended in the second step to determine indoor/outdoor label of the scene (note that the first location is selected randomly.)

3. Dataset bias

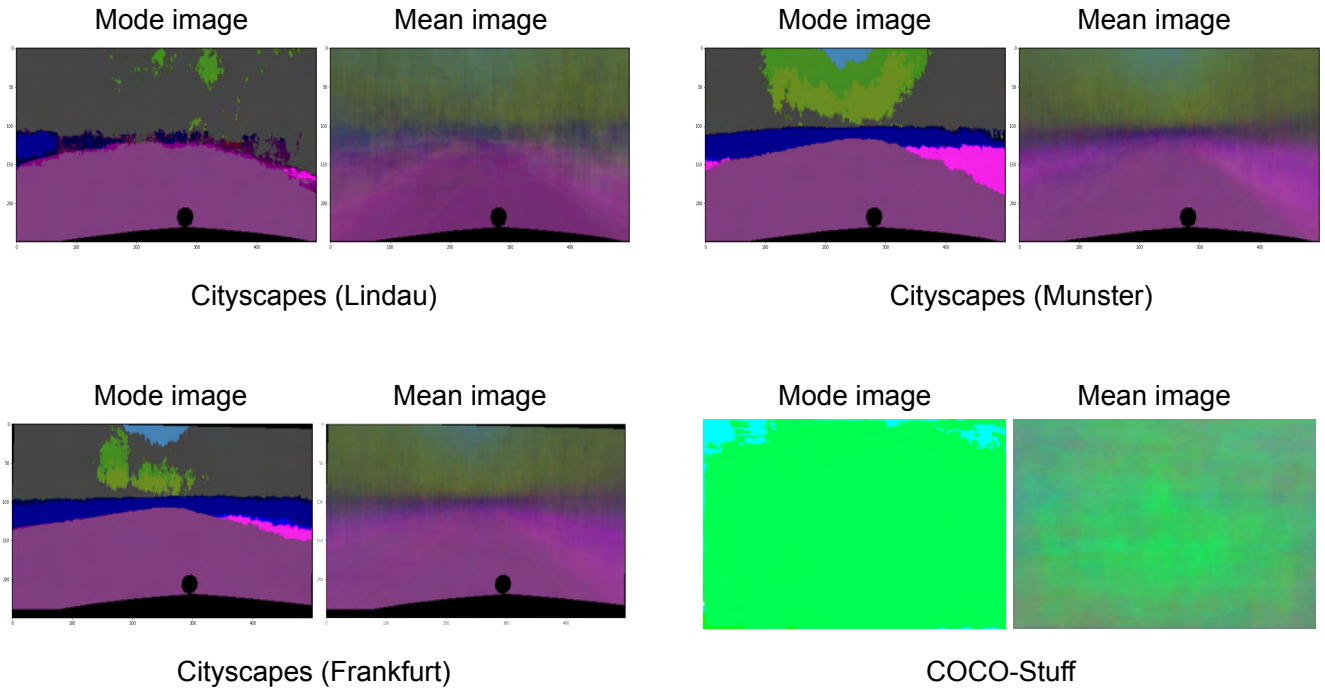


Figure 7: **Dataset bias:** Mode and mean image of datasets. On Cityscapes [3], we show the figures corresponding to the three 'cities' from the validation set, the high similarity of mode and mean images across different cities show high dataset bias. On COCO-Stuff (different from COCO-Things), we consider 91 classes including an unlabeled class consisting of 80 'things' classes combined under a single label. Since a large portion of many of the scenes consist of 'things' class instances, the dataset is biased towards this single 'unlabeled' class label as shown by green color in the figure. We observe the effect of this class-imbalance on the segmentation performance as well, as shown in figure 11, for COCO-stuff segmentation task. It is worth mentioning, compared to common segmentation methods working on the full observable scenes, our architecture is affected more by this class imbalance has due to the small size of glimpses and the partial observability setting. If the attended areas consist of objects from 'things' classes, the network does not benefit from those glimpses.

4. Visual comparison with Attend and Segment

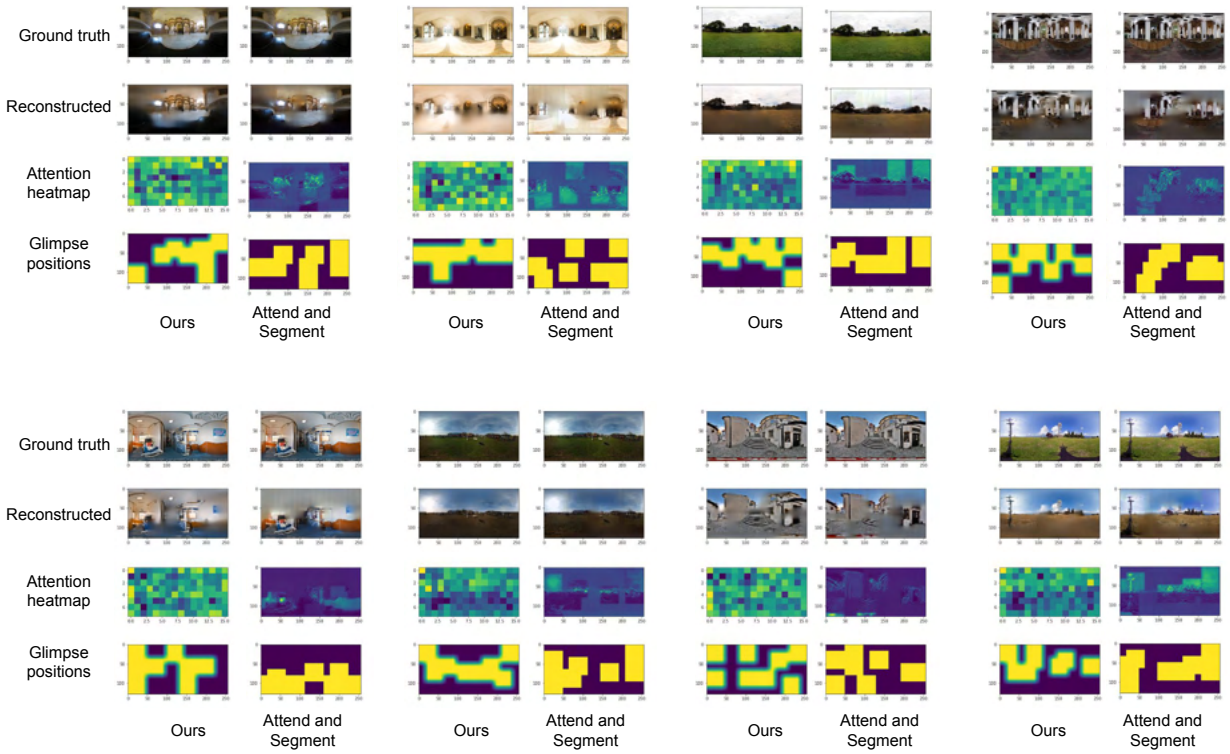


Figure 8: **Reconstruction comparison:** We compare the reconstruction performance against Attend and Segment model [6] on SUN360 dataset [7]. We observe that [6] reconstructed images contains visual patch-like artifacts, while our results provide a more consistent image reconstruction.

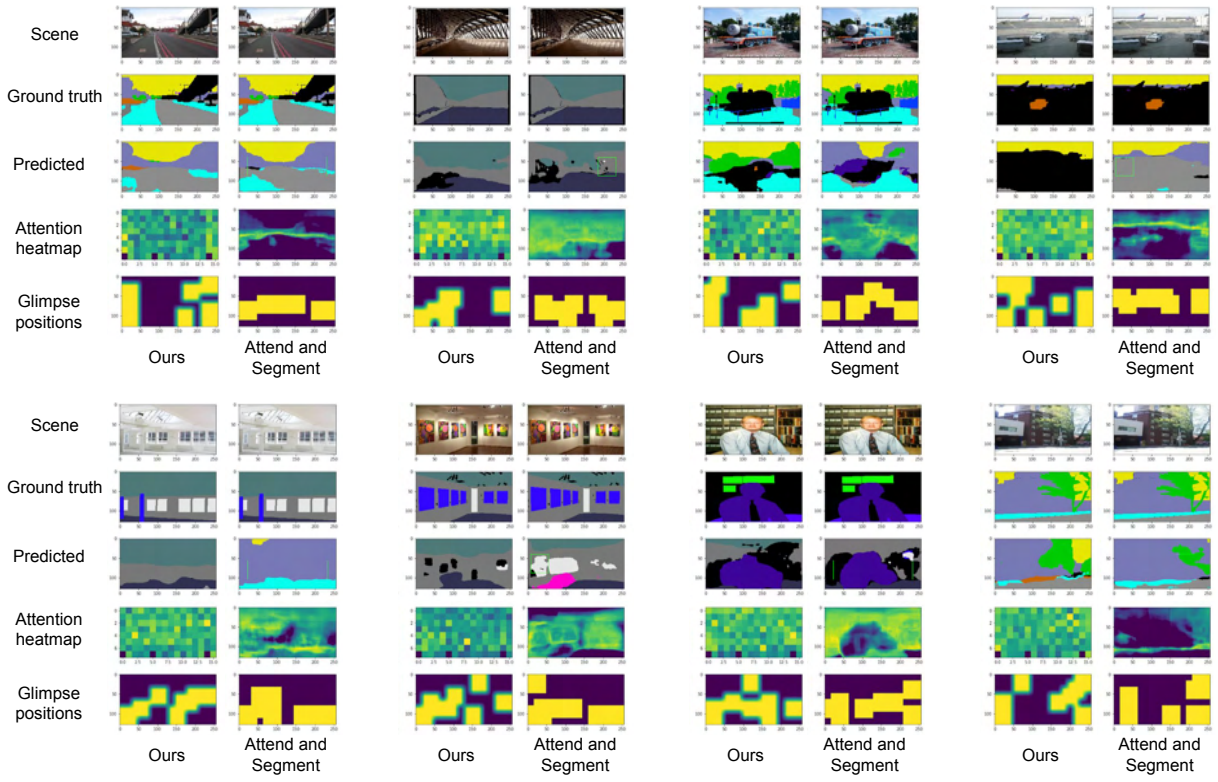
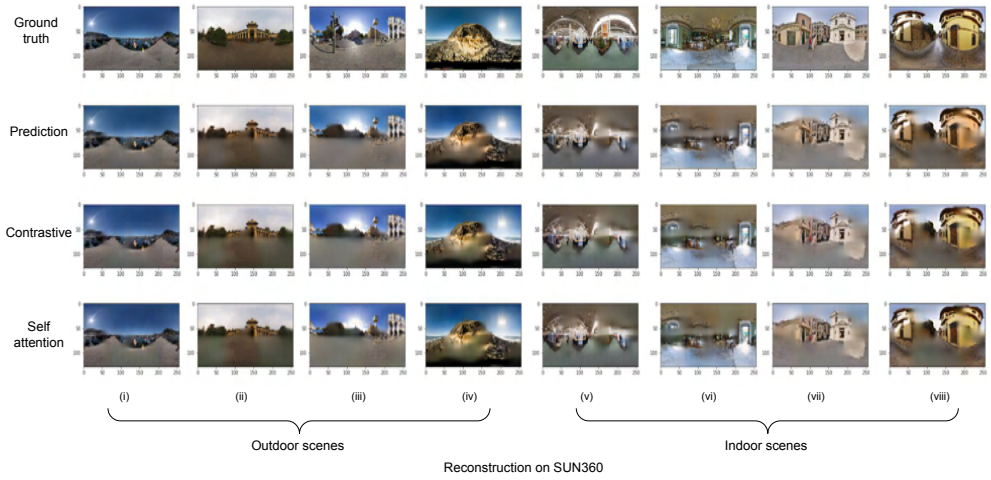
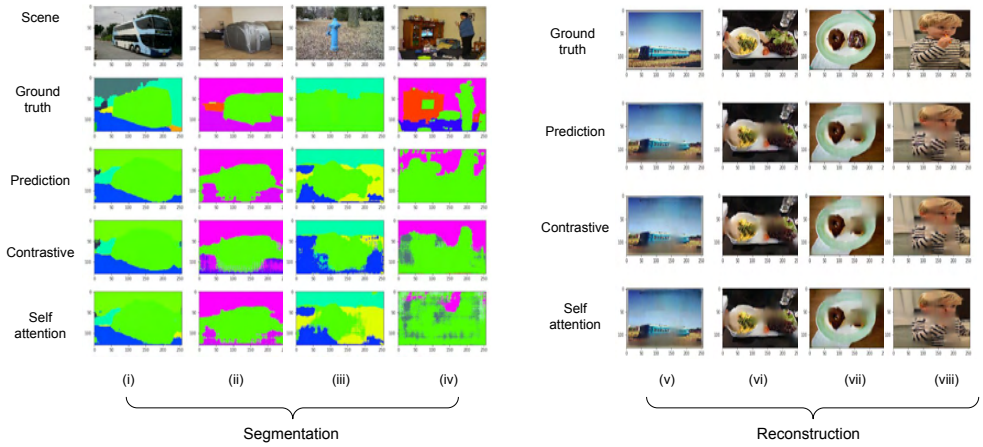


Figure 9: **Segmentation comparison:** Here we visually compare our result against Attend and Segment model [6] for segmentation task on ADE20k dataset [8]

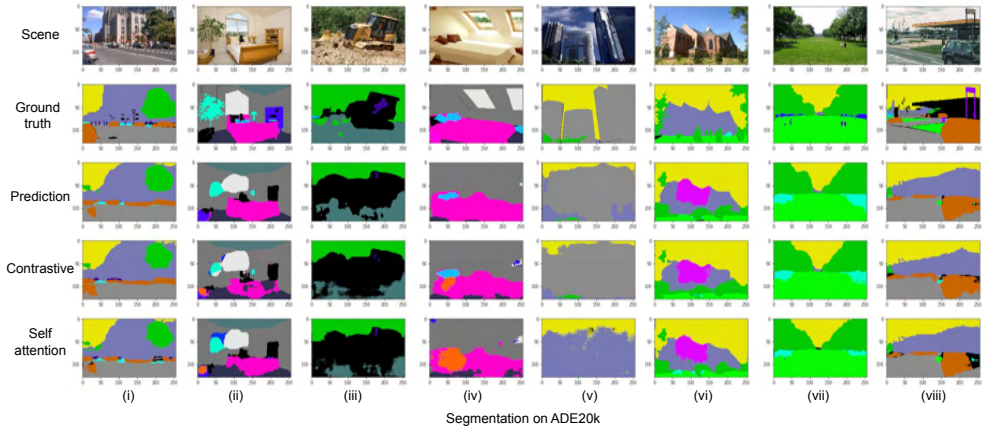
5. Prediction by different streams



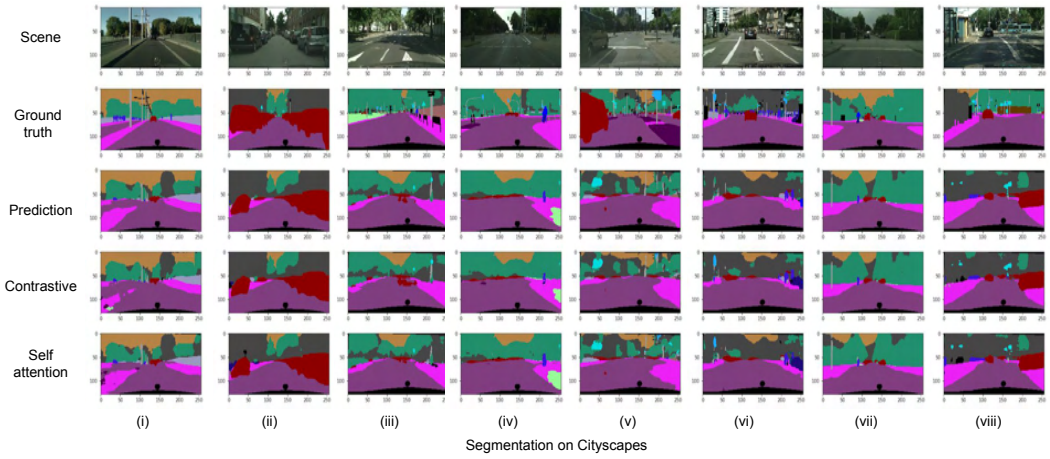
(a) Reconstruction of indoor and outdoor scenes from SUN360 dataset [7]



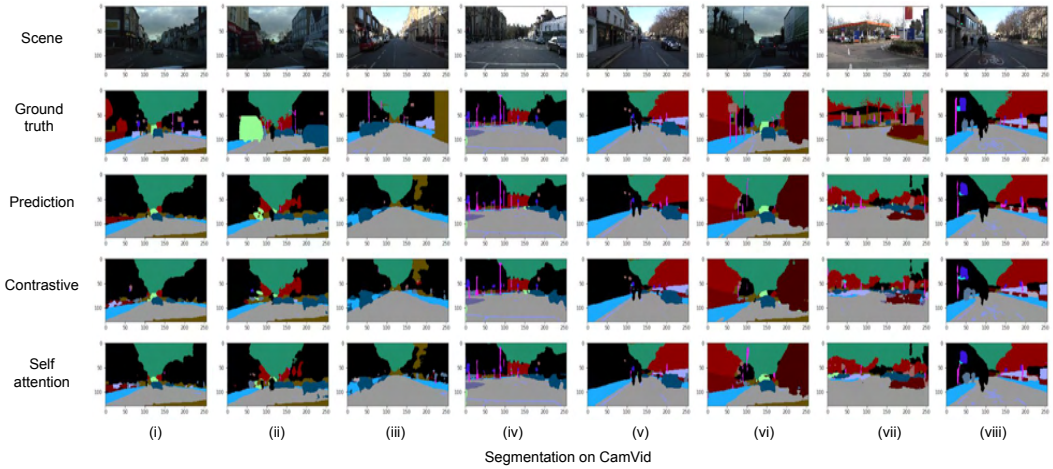
(b) Segmentation and reconstruction on COCO-Stuff dataset [2]



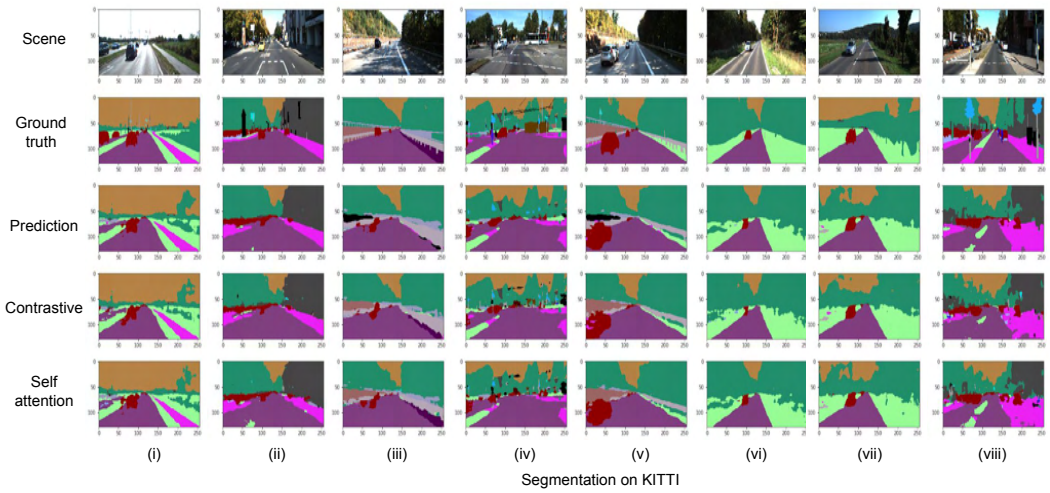
(c) Segmentation on ADE20k [8]



(d) Segmentation on Cityscapes dataset [3]



(e) Segmentation on CamVid dataset [1]



(f) Segmentation on KITTI dataset [4]

Figure 10: **Reconstruction and segmentation by self-attention stream, contrastive stream, and full model's output.**

6. Negative Results

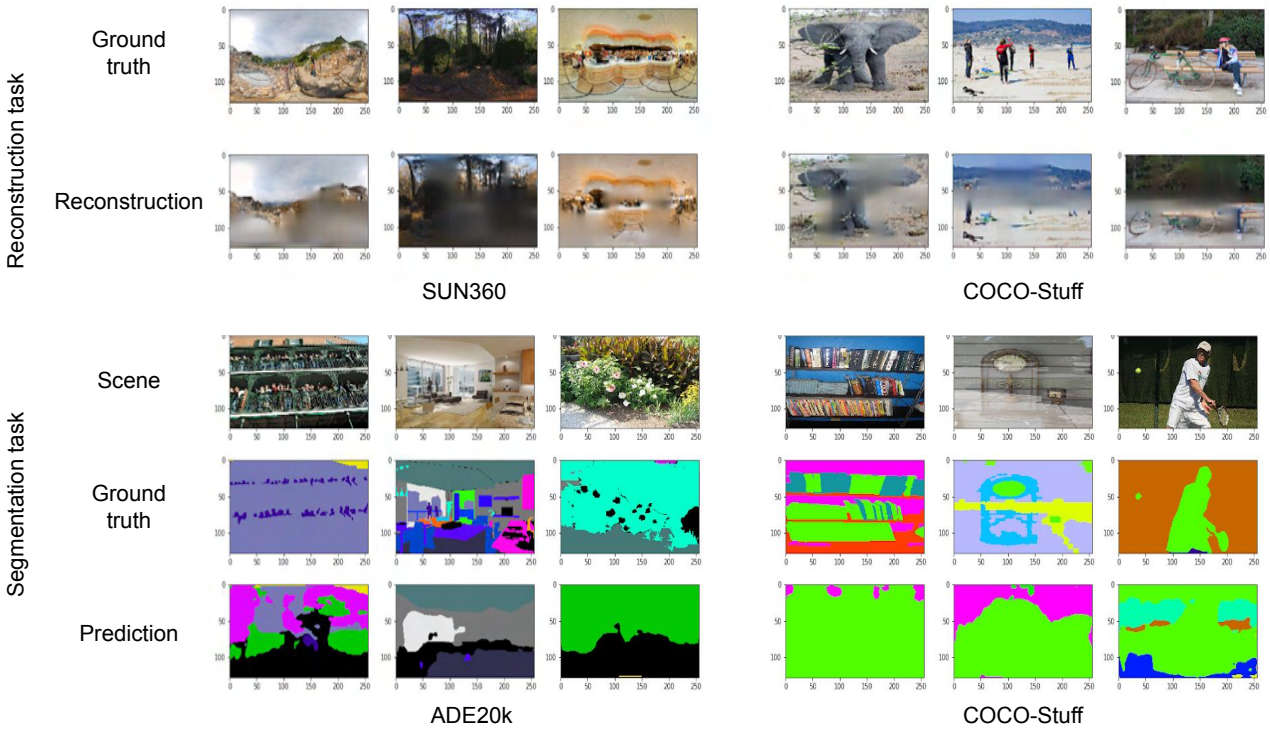
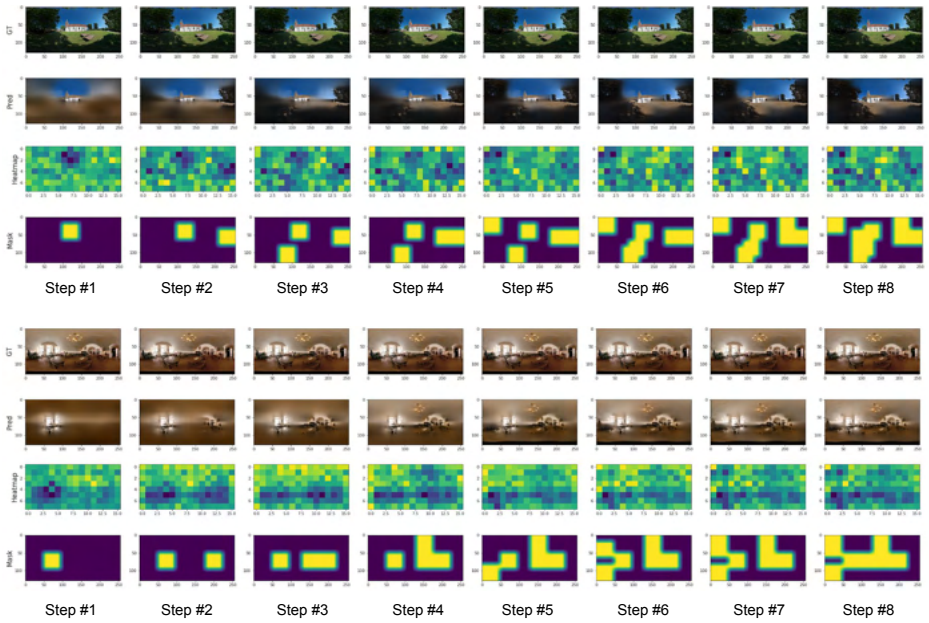
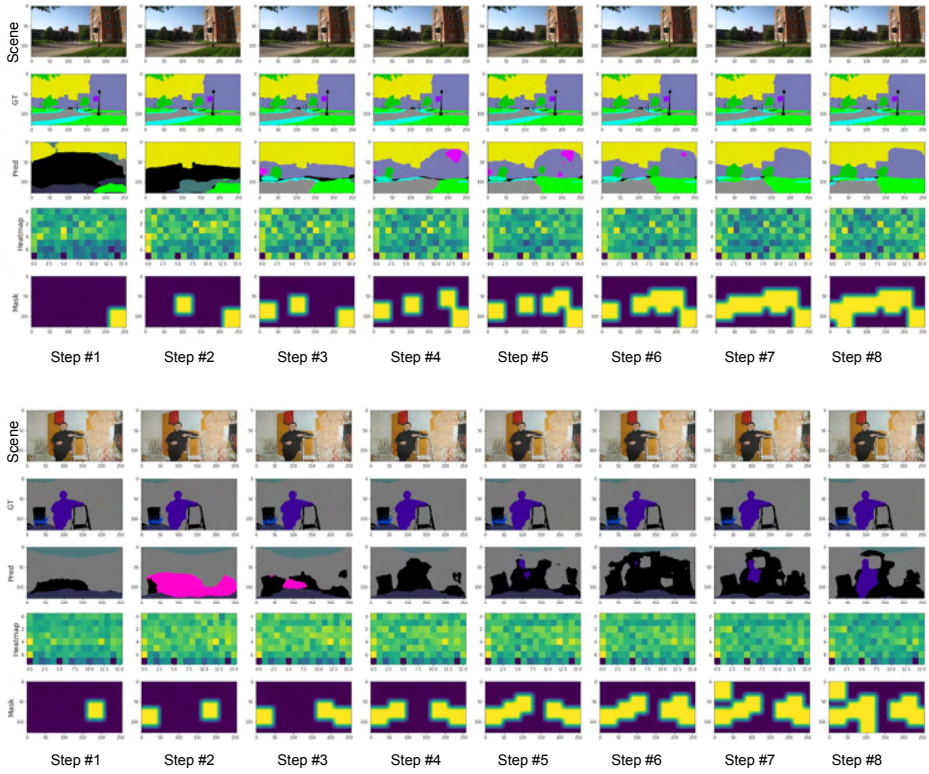


Figure 11: **Negative results for segmentation and reconstruction task.**

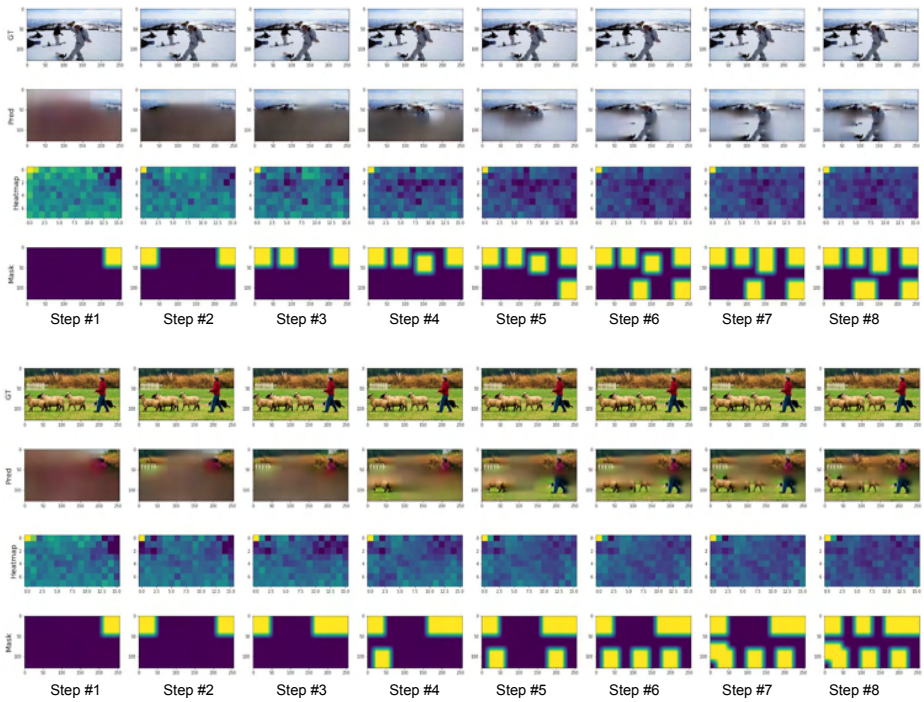
7. Step-by-step Glimpse selection



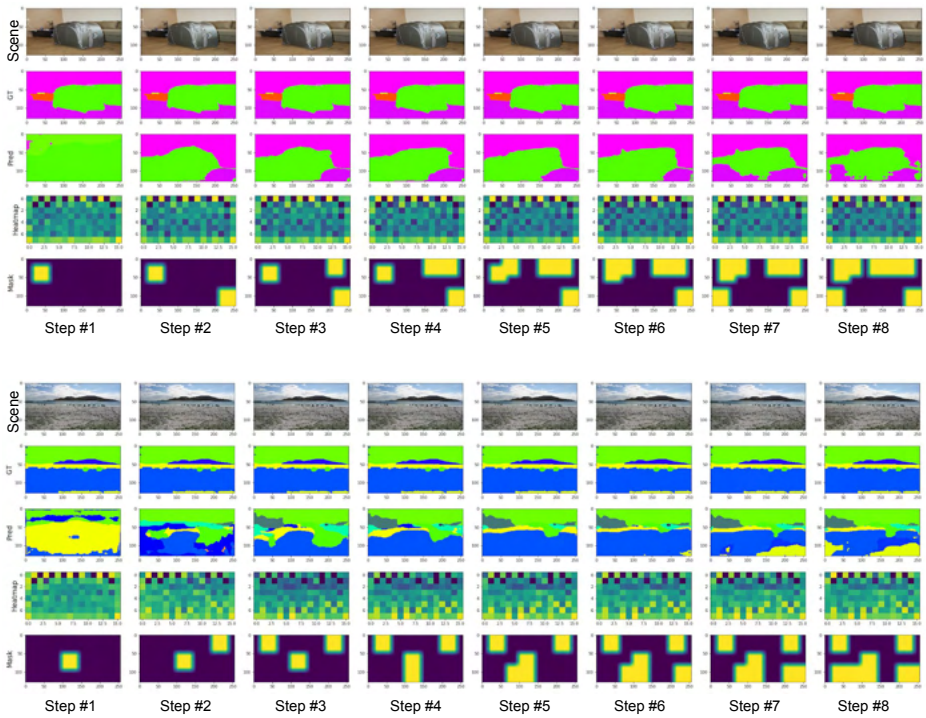
(a) Reconstruction on SUN360 dataset [7]



(b) Segmentation on ADE20k [8]



(c) Reconstruction on COCO-Stuff [2]



(d) Segmentation on COCO-Stuff [2]

Figure 12: **Glimpse-Attend-and-Explore:** Step-by-step glimpse selection and execution of reconstruction and segmentation task.

References

[1] Gabriel J Brostow, Jamie Shotton, Julien Fauqueur, and Roberto Cipolla. Segmentation and recognition using structure from motion point clouds. In *European conference on computer vision*, pages 44–57. Springer, 2008. 8

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 3, 7, 11

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 4, 8

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 8

[5] Soroush Seifi and Tinne Tuytelaars. Where to look next: Unsupervised active visual exploration on 360° input. *arXiv e-prints*, pages arXiv–1909, 2019. 1

[6] Soroush Seifi and Tinne Tuytelaars. Attend and segment: Attention guided active semantic segmentation. pages 305–321, 2020. 1, 5, 6

[7] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 5, 7, 10

[8] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 3, 6, 7, 10