Supplementary Material: Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers

Sixiao Zheng[1*] Jiachen Lu[1] Hengshuang Zhao[2] Xiatian Zhu[3] Zekun Luo[4] Yabiao Wang[4]
Yanwei Fu[1] Jianfeng Feng[1] Tao Xiang[3, 5] Philip H.S. Torr[2] Li Zhang[1†]
[1]Fudan University [2]University of Oxford [3]University of Surrey
[4]Tencent Youtu Lab [5]Facebook AI
https://fudan-zvg.github.io/SETR

In this supplementary material, we provide more experimental details, further ablation studies on the additional visualizations.

A. Visualizations

Position embedding Visualization of the learned position embedding in Figure 1 shows that the model learns to encode distance within the image in the similarity of position embeddings.

Features Figure 3 shows the feature visualization of our SETR-*PUP*. For the encoder, 24 output features from the 24 transformer layers namely $Z^1 - Z^{24}$ are collected. Meanwhile, 5 features ($U^1 - U^5$) right after each bilinear interpolation in the decoder head are visited.

Attention maps Attention maps (Figure 4) in each transformer layer catch our interest. There are 16 heads and 24 layers in T-large. Similar to [1], a recursion perspective into this problem is applied. Figure 2 shows the attention maps of different selected spatial points (red).

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint*, 2020. 1

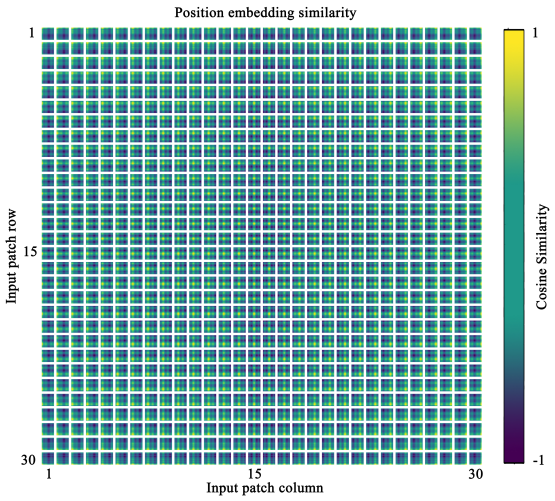

Figure 1. Similarity of position embeddings of SETR-*PUP* trained on Pascal Context. Tiles show the cosine similarity between the position embedding of the patch with the indicated row and column and the position embeddings of all other patches.

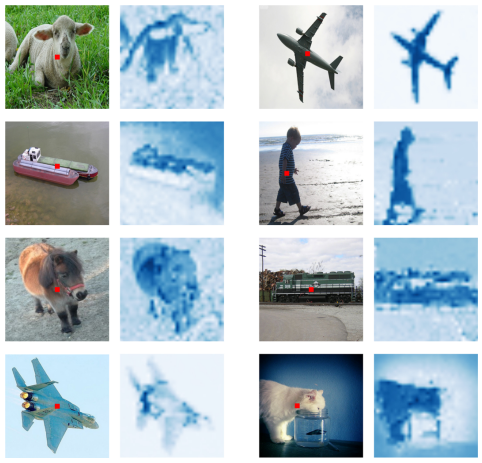

Figure 2. The first and third columns show images from Pascal Context. The second and fourth columns illustrate the attention map of the picked points (red).

*Work done while Sixiao Zheng was interning at Tencent Youtu Lab.

†Li Zhang (lizhangfd@fudan.edu.cn) is the corresponding author with School of Data Science, Fudan University. Yanwei Fu is with the School of Data Science, MOE Frontiers Center for Brain Science, and Shanghai Key Lab of Intelligent Information Processing, Fudan University. Jianfeng Feng is with the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University.

Figure 3. Visualization of output feature of layer $Z^1 - Z^{24}$ and $U^1 - U^5$ of SETR-*PUP* trained on Pascal Context. Best view in color. **First row:** The input image. **Second row:** Layer Z^1-Z^{12}. **Third row:** Layer Z^{13}-Z^{24}. **Fourth row:** Layer $U^1 - U^5$.

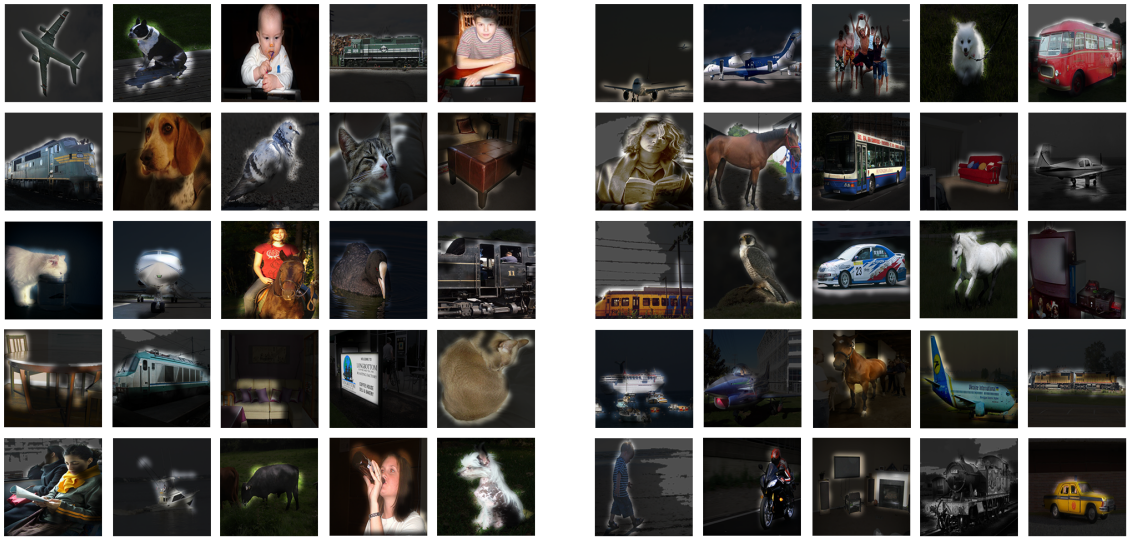

Figure 4. More examples of attention maps from SETR-*PUP* trained on Pascal Context.