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

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.