1. Additional Ablation Studies

We further make ablation studies of AttentionShift with respect to the comparison with other clustering method, effect of supervisions in the instance segmentation branch, and statistic analysis of the learned features.

Comparison with Clustering Method  Note that AttentionShift is related to traditional clustering methods, we conduct an ablation study to replace the AttentionShift with the K-Means algorithm to obtain the key-points. As shown in Table 3, AttentionShift significantly outperforms both the vanilla K-Means and the K-Means equipped with our key-point filtering (K-Means*). Especially, AttentionShift outperforms by 6.2% (24.3% vs 18.1%) upon the challenging mAP75, indicating that the proposed key-point shift learns fine-grained semantics and therefore segments object more accurate.

Effect of Pseudo Mask Supervision  Table 2 shows the effect of pseudo mask supervision in the instance segmentation branch. “Mask” indicates each pseudo mask is generated by binarizing each value of the instance attention map to foreground class or background. “Point” denotes the pseudo mask is replaced by point sets as defined in Sec. 4 in the main document. It shows that point supervision achieves 1.0% improvement when using the baseline method. With the proposed AttentionShift, point supervision significantly outperforms the pseudo binary mask supervision by 6.7% (53.2% vs 45.5%). The results shows that the stable and extreme points (key-points) are able to represent the fine-grained semantic of object parts than the pseudo masks, and therefore facilitate the method to localize the full object extent.

Statistic Analysis of the Learned Features  Fig. 1 shows the feature similarity (cosine similarity) of the proposed AttentionShift and the baseline method. Compared with the baseline method, AttentionShift learns more discriminative features, indicated by lower feature similarity between background and foreground (Fig. 1(a)) or between
Figure 1. Statistic analysis of the feature similarity during training using AttentionShift and the baseline method. (a) shows the feature similarity between foreground and background. (b) shows the feature similarity among objects from different categories (inter-class similarity). (c) shows the feature similarity among objects from the same category (intra-class similarity).

foregrounds of different categories (Fig. 1(b)) during the whole training phase. Specifically, by introducing key-point shift, AttentionShift first learns more diversity semantics (results in lower intra-class feature similarity, Fig. 1(c)) at the early training epochs. In the final epoch, AttentionShift reduces the semantic bias and learns comparably compact intra-class features than the baseline method.

**Backbone Fairness** We conduct experiments to show the backbone we use is not superior to other backbones in this task. We replace the backbone of BESTIE with ViT-S and get $mAP_{25} = mAP_{50} = 36.2\%$, and $mAP_{75} = 17.3\%$. ViT-S does not outperform HRNet48, it shows that the performance improvement does not come from the backbone.

### 2. Additional Visualization Analysis

We provide additional visualization results of Fig. 4 and Fig. 5 in the main document. The results are shown in Fig. 3 and Fig. 2 respectively.

We also visualize the instance segmentation results of the proposed AttentionShift, as shown in Fig. 4.

### References


Figure 2. Visualization of iterative estimation of part-based attention maps using the proposed token querying and key-point shift procedure. (Best viewed in color)
Figure 3. Visualization of the self-attention maps and instance masks.
Figure 4. Visualization of instance segmentation results on Pascal VOC 2012 \textit{val} and MS-COCO 2017 \textit{test-dev}.