SPANet: Frequency-balancing Token Mixer using Spectral Pooling Aggregation Modulation  
– Supplementary Material –

Guhnoo Yun\textsuperscript{1,2} Juhan Yoo\textsuperscript{3} Kijung Kim\textsuperscript{1,2} Jeongho Lee\textsuperscript{1,2} Dong Hwan Kim\textsuperscript{1,2}

\textsuperscript{1}Korea Institute of Science and Technology  \textsuperscript{2}Korea University  \textsuperscript{3}Semyung University

\{doranlyong, plan100day, kape67, gregorykim\}@kist.re.kr
unchinto@semyung.ac.kr

S1. Visualizations of Context Aggregation Results from the SPAM

Figure 1 presents the visualized activation maps of each SPG and the context aggregated by addition, at the last layer of SPANet-S trained on ImageNet-1K \cite{1}. The second to fourth columns demonstrate that SPGs learn different contexts at balancing parameters of 0.7, 0.8, and 0.9, respectively. As shown in the second column where the balancing parameter is 0.7 which means relatively weak low-pass filtering (i.e., relatively strong high-pass filtering), SPG concentrates on the overall shapes of objects. The final column depicts the contexts that have been adaptively gathered from the various balancing levels. It shows that the proposed SPAM efficiently captures contextual information of objects, leading to improved performance.

S2. Visualizations of Classification Results

In order to visualize the results of different models trained on ImageNet-1K \cite{1}, we utilized Score-CAM \cite{6}. As shown in Figure 2, SPANet-S exhibits superior semantic object localization and aggregation capabilities compared to the other models, although all models achieve correct object classification.

S3. Visualizations of Detection and Segmentation Results

In Figure 3, we also present qualitative results for object detection and instance segmentation on COCO val 2017 \cite{5} and semantic segmentation on ADE20K \cite{7}, showcasing SPANet’s ability to integrate seamlessly with dense prediction models like RetinaNet \cite{4}, Mask R-CNN \cite{2}, and Semantic FPN \cite{3}. Our results demonstrate the superior quality of SPANet in achieving high-quality results in these tasks.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. \textsuperscript{1, 2, 3}


Figure 1: Visualization of the results of SPGs and aggregated contexts at the last layer of SPANet-S trained on ImageNet-1K [1]. The columns from left to right are input images, SPG maps at balancing parameters of 0.7, 0.8, and 0.9, and contexts aggregated by addition.
Figure 2: **Score-CAM [6]** activation maps of the models trained on ImageNet-1K [1]. The source images are from validation set.
Figure 3: Qualitative evaluation results of object detection and instance segmentation on COCO val2017 [5], and semantic segmentation on ADE20K [7]. The results, ordered from left to right, are generated by SPANet-S backbone equipped with RetinaNet [4], Mask R-CNN [2], and Semantic FPN [3], respectively.