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

Guhnoo Yun[1,2] Juhan Yoo[3] Kijung Kim[1,2] Jeongho Lee[1,2] Dong Hwan Kim[1,2]

[1]Korea Institute of Science and Technology
[2]Korea University [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 [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 [1], we utilized Score-CAM [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 `val2017` [5] and semantic segmentation on ADE20K [7], showcasing SPANet's ability to integrate seamlessly with dense prediction models like RetinaNet [4], Mask R-CNN [2], and Semantic FPN [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. 1, 2, 3

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 4

[3] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 1, 4

[4] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 4

[5] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 4

[6] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 24–25, 2020. 1, 3

[7] 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. 1, 4

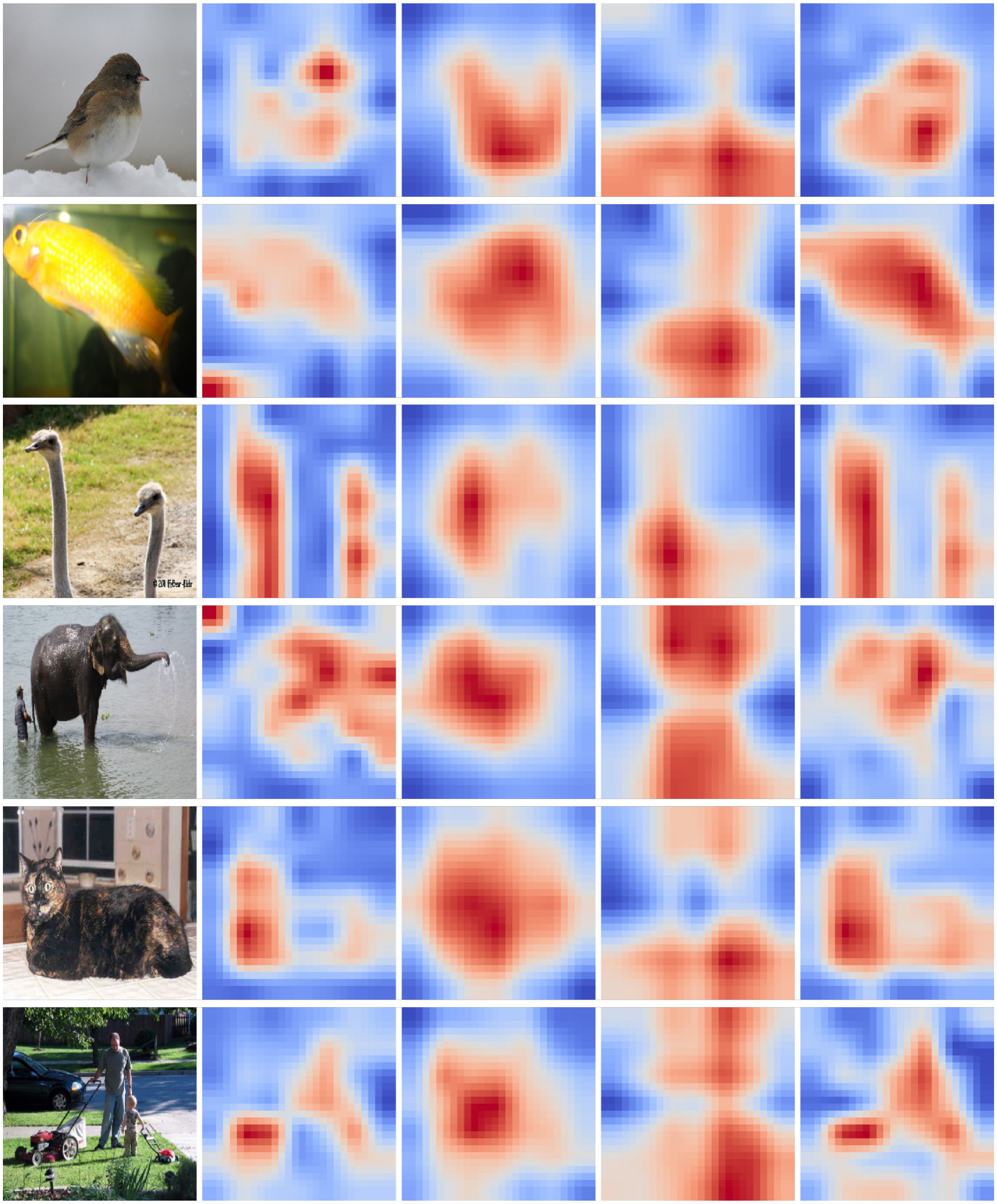


Figure 1: **Visualization of the results of SPGs and aggregated contexts at the last layer of SPANet-S trained on ImageNet-1K [1].** The columns from left to right are input images, SPG maps at balancing parameters of 0.7, 0.8, and 0.9, and contexts aggregated by addition.

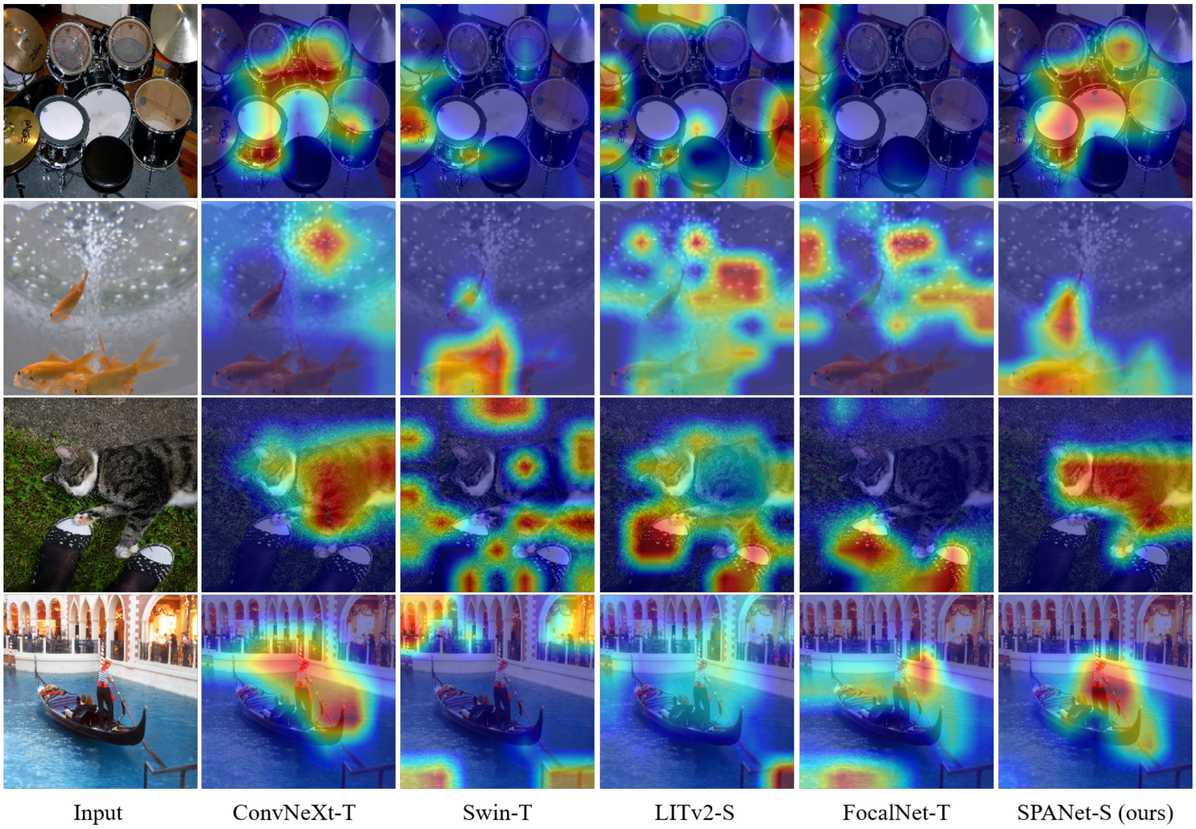
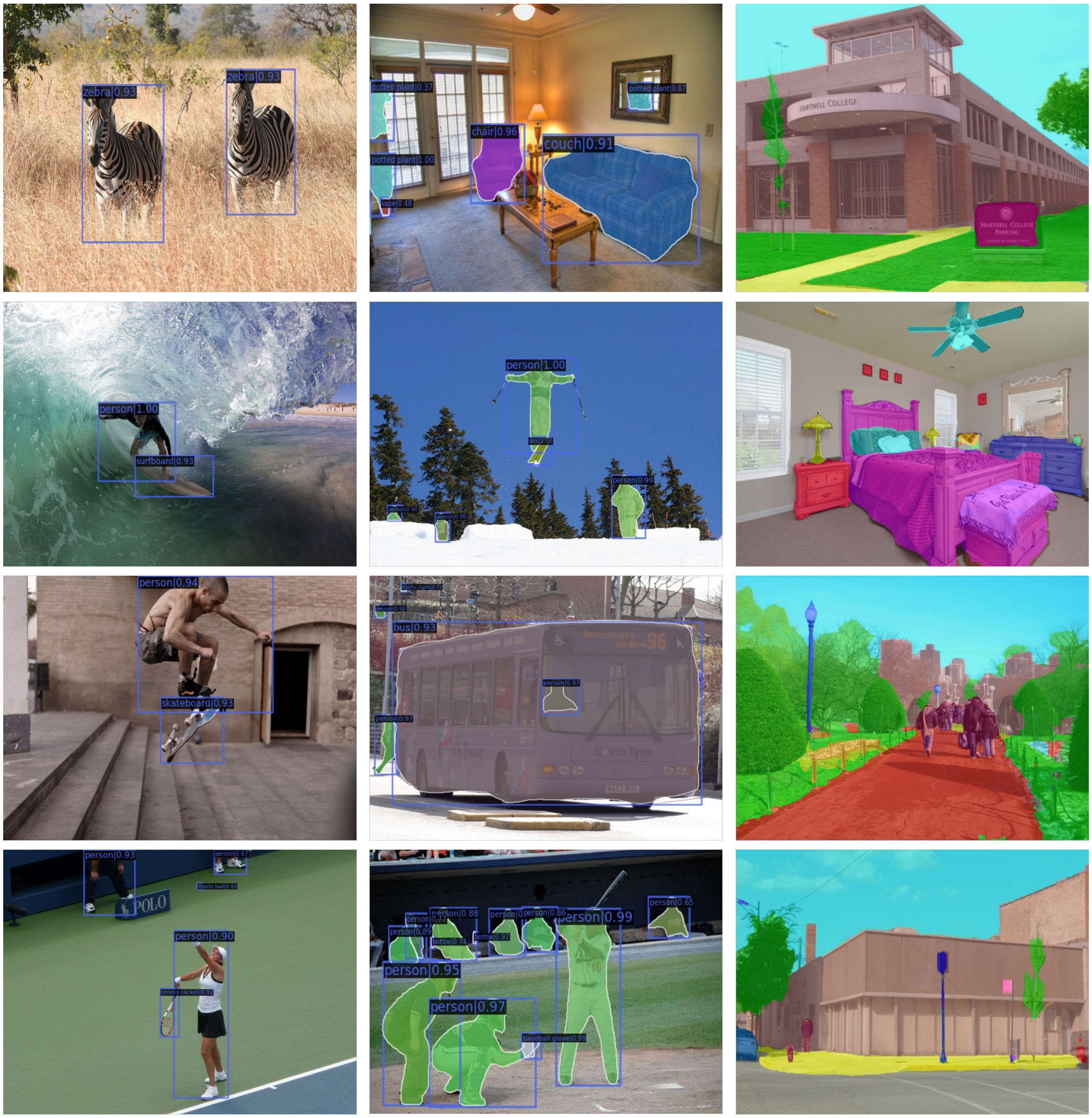


Figure 2: **Score-CAM [6] activation maps of the models trained on ImageNet-1K [1].** The source images are from validation set.



Object Detection on COCO Instance Segmentation on COCO Semantic Segmentation on ADE20K

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.