Aggregating Bilateral Attention for Few-Shot Instance Localization –Supplementary Material–

He-Yen Hsieh, Ding-Jie Chen, Cheng-Wei Chang, and Tyng-Luh Liu
Institute of Information Science, Academia Sinica, Taiwan
heyen@iis.sinica.edu.tw, djchen.tw@gmail.com, johnnyccw.tw@gmail.com, liutyng@iis.sinica.edu.tw

To further realize our Aggregating Bilateral Attention (ABA) mechanism, this document provides complete experimental results and visualizations of attention heat maps in various model configurations/components for discussion.

One-shot object localization on MS-COCO dataset. Table 1 shows the comparison results with the previous methods for the one-shot object detection task on the MS-COCO dataset. The previous methods in this experiment include SiamMask and CoAE, as mentioned in the main paper. Comparing the results of 'CoAE' against 'CoAE (ABA)' in Table 1 show that all the dataset splits gain clear improvement while replacing the typical non-local block within the CoAE with our ABA mechanism. Such improvements were observed not only in the seen classes but also in the unseen classes, which again demonstrates the boosted attention quality by our aggregating bilateral attention .

The effects of p in the embedding norm ω^N. Figure 2 visualizes the attention heat maps deriving from the embedding norm of various p values. Recall that the embedding norm concerns the feature distance, implemented with p-norm, per query-support data pair. Note that the larger p values in ω^N prefer the larger attention heat maps, *i.e.*, including more medium-similarity pixels as each row in Figure 2. As a result, the norm operator retrieves the high-similarity pixels and recalls more medium-similarity pixels neighboring within an embedding space. Concretely, the bottom two sets of multi-instance query images show the ability of ω^N to recall the related-class (medium-similarity) instances, such as person-bicycle, person-horse, and person-chair, which usually coexists within the dataset. This class-similarity is modeled by the learned embeddings and our distance metric for calculating query-support similarity.

Examples of aggregating bilateral attention . Figure 1 and Figure 3 visualize the failed and successful cases' attention heat maps using our ABA mechanism on tackling the one-shot object detection task. In Figure 1, both rows show the co-exist object class should be the person, yet our ABA additionally retrieves the similar-class objects, *i.e.*, bike and car in the top row, or the white horse in the bottom row. Though it seems that ABA retrieves too many attended regions, its ability to associate similar objects indeed raises the potential of recalling the seen-class and unseen-class objects implied by the support image. Figure 3 shows more visualization results improved by our ABA compared with the typical dot-product-based non-local attention. In Figure 3, the four image sets from top to bottom show the cases of the non-local attention's failure to localize the large region of interest, failure to localize the correct regions of interest, failure to ignore the unrelated background regions, and attention everywhere. In contrast, the proposed aggregating bilateral attention mechanism shows the attention heat maps of better qualities.

Methods	Split				Average
	1	2	3	4	
SiamMask (seen)	38.9	37.1	37.8	36.6	37.6
CoAE (seen)	42.2	40.2	39.9	41.3	40.9
CoAE (ABA) (seen)	**49.8**	**48.1**	**45.7**	**48.7**	**48.1**
SiamMask (unseen)	15.3	17.6	17.4	17.0	16.8
CoAE (unseen)	23.4	23.6	20.5	20.4	22.0
CoAE (ABA) (unseen)	**25.6**	**26.0**	**21.6**	**21.3**	**23.6**

Table 1: Comparison on one-shot object detection task using MS-COCO dataset in terms of AP50 score (%).

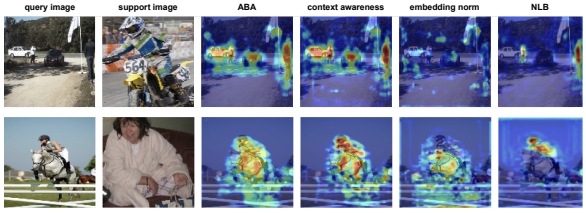

Figure 1: **Effects of ABA in failure cases.** The leftmost two columns show the input image pair, the rightmost column and middle three columns respectively show the attention heat maps of non-local block and our ABA mechanism.

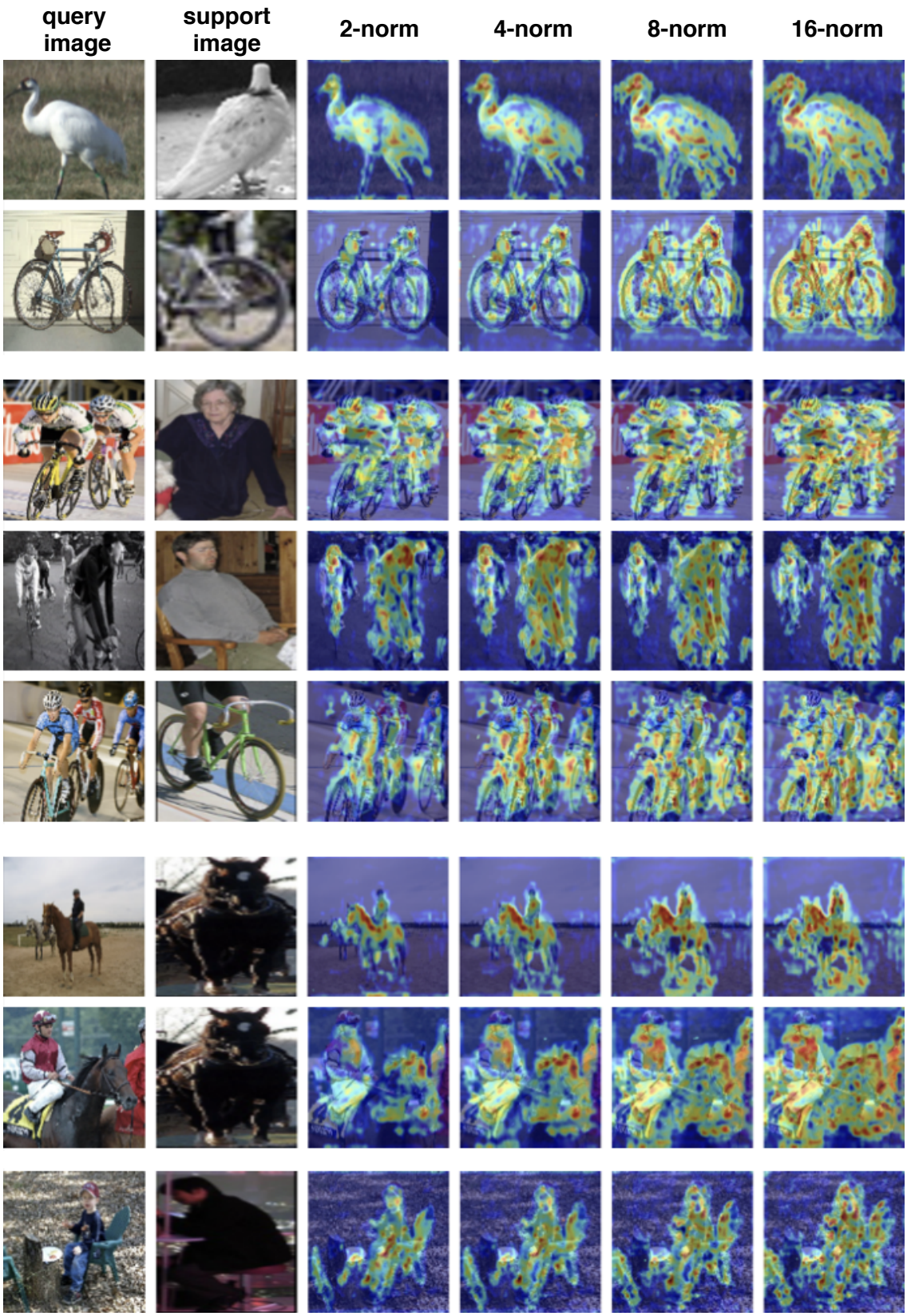

Figure 2: **Effects of Embedding Norm.** The leftmost two columns show the input image pair, the remained columns from left to right show the attention heat maps of our embedding norm using $p = 2$, $p = 4$, $p = 8$, and $p = 16$, respectively.

Figure 3: **Effects of ABA in successful cases.** The leftmost two columns show the input image pair, the rightmost column and middle three columns respectively show the attention heat maps of non-local block and our ABA mechanism.