SUPPLEMENTARY DOCUMENT

Where to Look?: Mining Complementary Image Regions for Weakly Supervised Object Localization

A. Experimental Studies on CUB-200-2011

In this section, we provide details of our experimental study on the CUB-200-2011 dataset. We show that by introducing complementary images, even a simple mechanism of addition (late addition fusion) of the outputs from classifiers X and $\tilde{X}$ gives promising results, when compared to HaS [3]. However, as evident from table I, our proposed method has significant gains in both the localization metrics. Our method effectively uses the information lost in regional dropout in one of the input images to both classify and localize objects much efficiently, using both the inputs, X and $\tilde{X}$.

Method	Top-1 Loc (%)	GT-Loc (%)
**Hide and Seek [3]	57.86	68.27
Late Addition Fusion	60.65	72.28
Ours	**64.70**	**77.35**

Table I: Comparison of the proposed approach with a simple late addition-based fusion of the outputs of two classifiers X and $\tilde{X}$. We use ResNet50 as the backbone architecture for the above experiment. ** indicates values computed on our own for the ResNet50 backbone to ensure fair comparison.

We also demonstrate the effect of different patch sizes $\{16, 32, 44, 56\}$ and a combination of these (*Mixed* approach as stated in table II) on the classification accuracy of our model. We observe that classification accuracy is highest when the patch size is 56, i.e., when the maximum part of the input image is visible during training. However, this is not the case with localization accuracies. Both Top-1 Loc and GT-Loc perform well when we randomly sample the patch sizes from $\{16, 32, 44, 56\}$ for each input image in every epoch during training, as also discussed subsection **4.3** of our paper.

Patch Size	Top-1 Clas (%)	Top-1 Loc (%)	GT-Loc (%)
16	68.21	55.98	67.52
32	69.23	57.68	68.57
44	70.59	56.20	69.90
56	**72.59**	57.56	70.50
Mixed	71.65	**58.12**	**71.28**

Table II: Effect of different patch sizes on classification and localization accuracies. The *Mixed* approach, similar to that in [3], performs best on the localization metrics. For the above experiment, we have used VGG16 as the backbone CNN architecture.

We also demonstrate the effect of the individual modules in our proposed architecture on classification and localization accuracies, with CUB-200-2011 dataset in table III (ablation studies). CAAM refers to our proposed Channel-wise Attention Module, SSAM refers to Spatial Self-Attention Module and L_{at_fuse} refers to our Attention-based Fusion Loss function. These modules are explained in detail in section **3** of our main manuscript.

CAAM	SSAM	L_{at_fuse}	Top-1 Loc (%)	GT-Loc (%)
✓	✓	✗	63.05	75.66
✓	✗	✓	62.11	75.06
✗	✓	✓	59.54	73.63
✓	✓	✓	**64.70**	**77.28**

Table III: **Ablation Studies.** Effect of each module in the architecture on localization accuracy. For the above experiment, we have used ResNet-50 as the backbone CNN architecture. We observed that the proposed CAAM is the most significant module in terms of accurate localization of objects, followed by SSAM and our L_{at_fuse} loss function.

We also evaluate our method on the recently proposed $MaxBoxAccv2$ metric [1]. Table IV gives a detailed study of the $MaxBoxAccv2$ metric on CUB200-2011 and ILSVRC 2016 datasets with ResNet50 as the backbone CNN architecture.

Method	CUB-200-2011	ILSVRC 2016
CAM [7]	73.2	64.2
HaS-32 [3]	78.1	63.2
ACoL [5]	72.7	61.7
SPG [6]	71.4	63.5
ADL [2]	73.5	64.2
CutMix [4]	67.8	63.9
Ours	**78.5**	**64.8**

Table IV: **Evaluating our method on MaxBoxAccv2.** We evaluate our model on the recently proposed $MaxBoxAccv2$ metric [1] on ResNet50 as the backbone CNN.

B. Qualitative Illustration of Performance

Using visual illustrations, we now show the effect of our proposed Attention-based Fusion Loss, as explained in subsection **3.6** of our paper. From figure I, we observe how our model looks at complementary body parts of the birds (e.g., head, torso, wings, tail) to analyze and provide better performance. After applying our proposed Attention-based Fusion Loss on the localization maps of two classifiers X and $\tilde{X}$, our model is able to localize the integral objects. E.g., in row (5) of figure I, the bird is **Scissor-Tailed-Flycatcher**. We can clearly observe that one of the classifiers focuses more on the face of the bird, which is its most-discriminating part, whereas the other classifier highlights its scissor-tail. Finally, in the last column of the same row, we can see how our model captures both the face and the scissor-tail of the bird, leading to its effective localization.

For figures II and III, column (a) refers to the input image, column (b) refers to the attention map generated by our model, column (c) refers to the bounding box predicted by our model (here, the bounding boxes in Green are the ones predicted by our model whereas the bounding boxes in Red refer to that of the ground truth) and column (d) denotes the bounding box overlayed on the attention map.

Figure II shows our qualitative results on ILSVRC 2016 dataset. Observe that in cases of **King-snake** and **Joystick** (rows (2) and (3) respectively), our model skillfully looks only at the object of interest present in the hand while localizing it correctly. Also, our model is good at localizing places (e.g., in case of row (4), it localizes the **Fountain** class). Also, in row (6), our model does an excellent job in localizing **School-bus**, which is present in the background and marginally occluded by the person standing in front of it.

Figure III illustrates a few of the qualitative results on the CUB-200-2011 dataset. We observe that our method also focuses on less-discriminative body parts of birds like the wings in case of rows (1), (2), and (6), and tail of the bird in case of row (4), along with their most-discriminative part like the head of the bird. Overall, our model generally learns to efficiently localize the objects present in an image by looking at complementary image regions. Also, notice that the attention maps generated by our proposed model are quite precise.

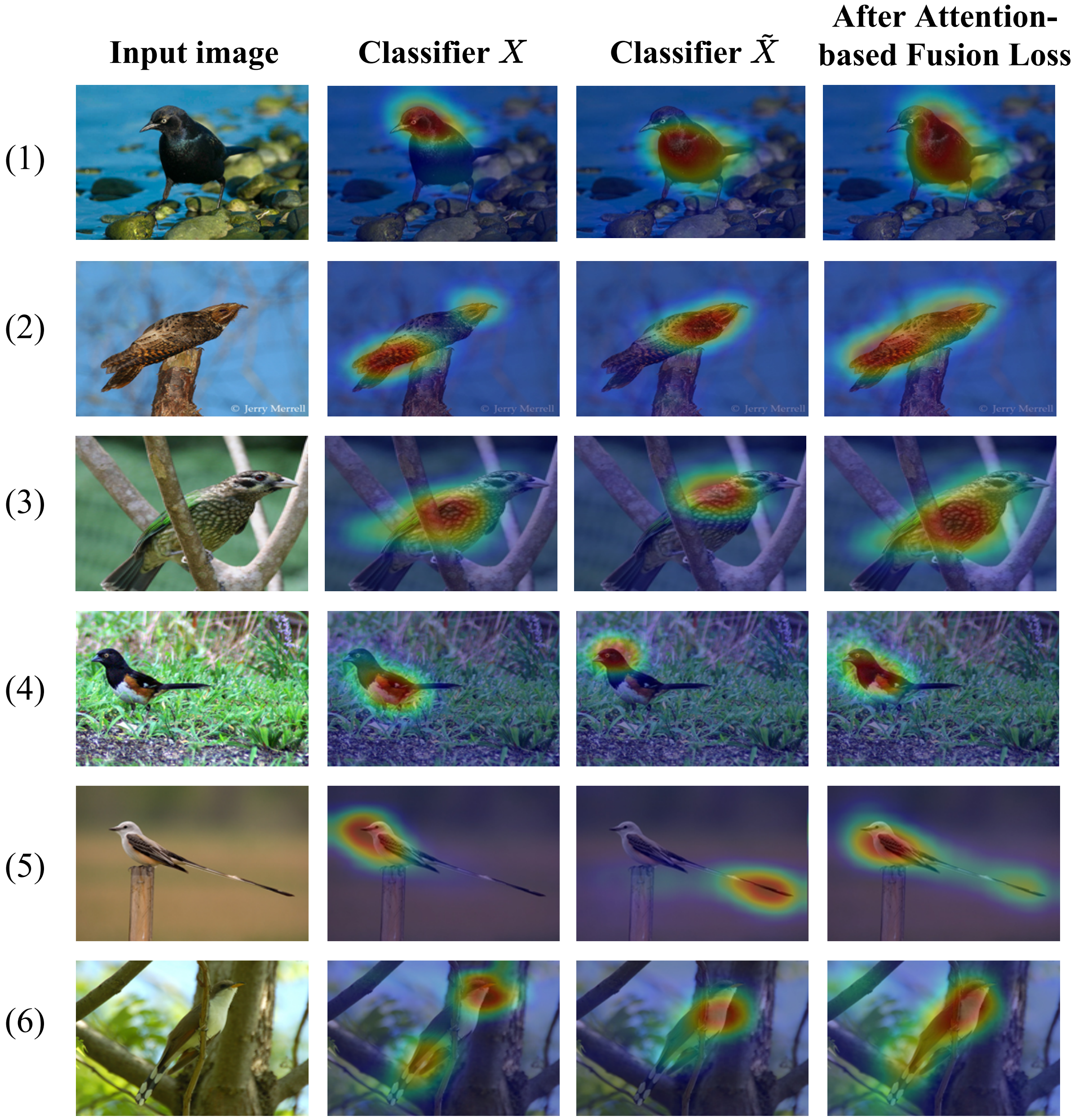

Figure I: **Visualizing the effect of the proposed Attention-based Fusion Loss.** During training, we visualize the effect of our proposed loss function. The left column denotes the input image, the second and third columns denote the localization maps of our two classifiers and the right column denotes the localization map after fusion of the outputs of classifier X and classifier $\tilde{X}$.

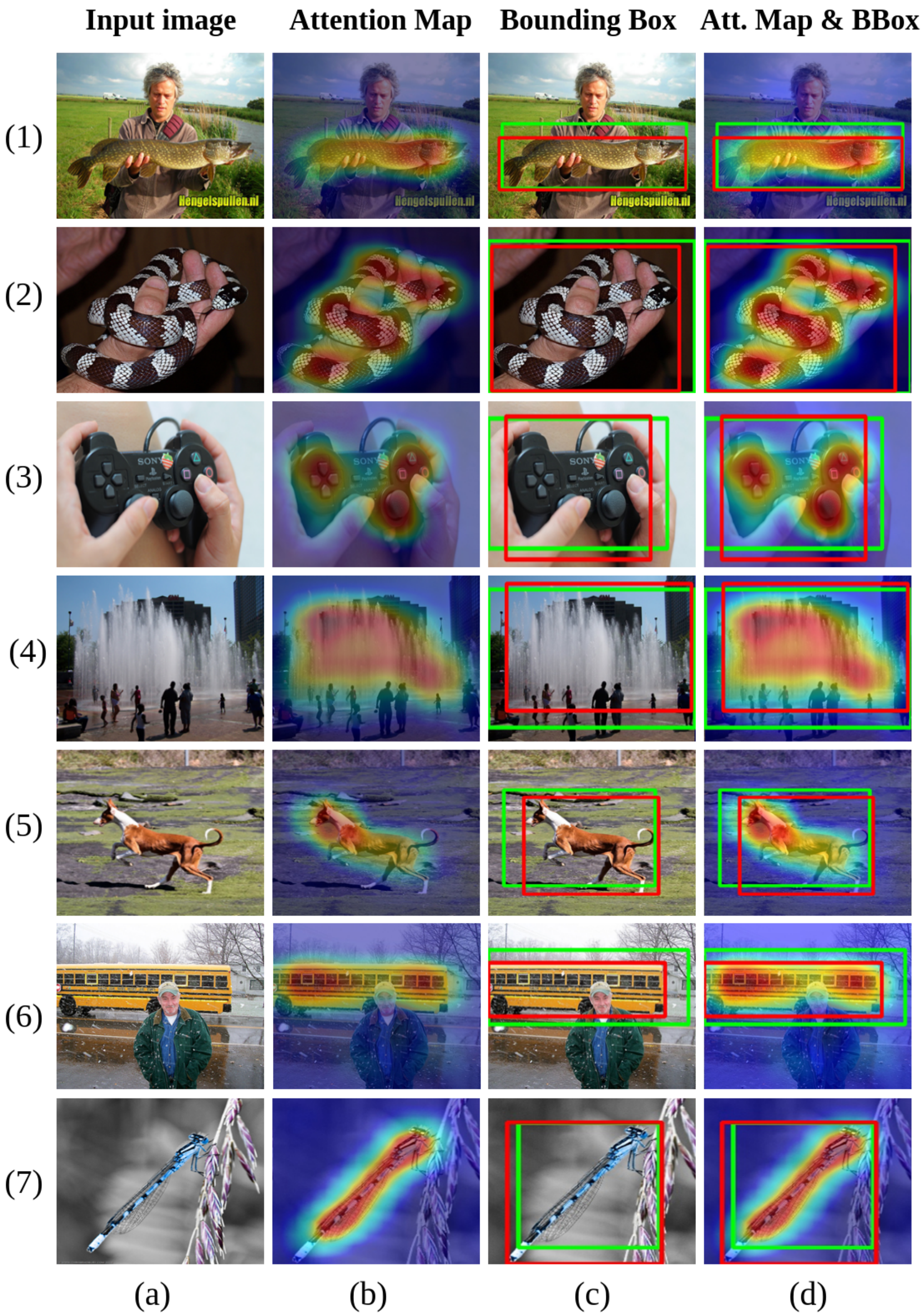

Figure II: **Qualitative Results on the ILSVRC 2016 dataset.** The right-most column gives the results (bounding boxes in green) of our proposed model, which are close to that of the ground-truth (in red). Attention maps resemble the entire object being targetted.

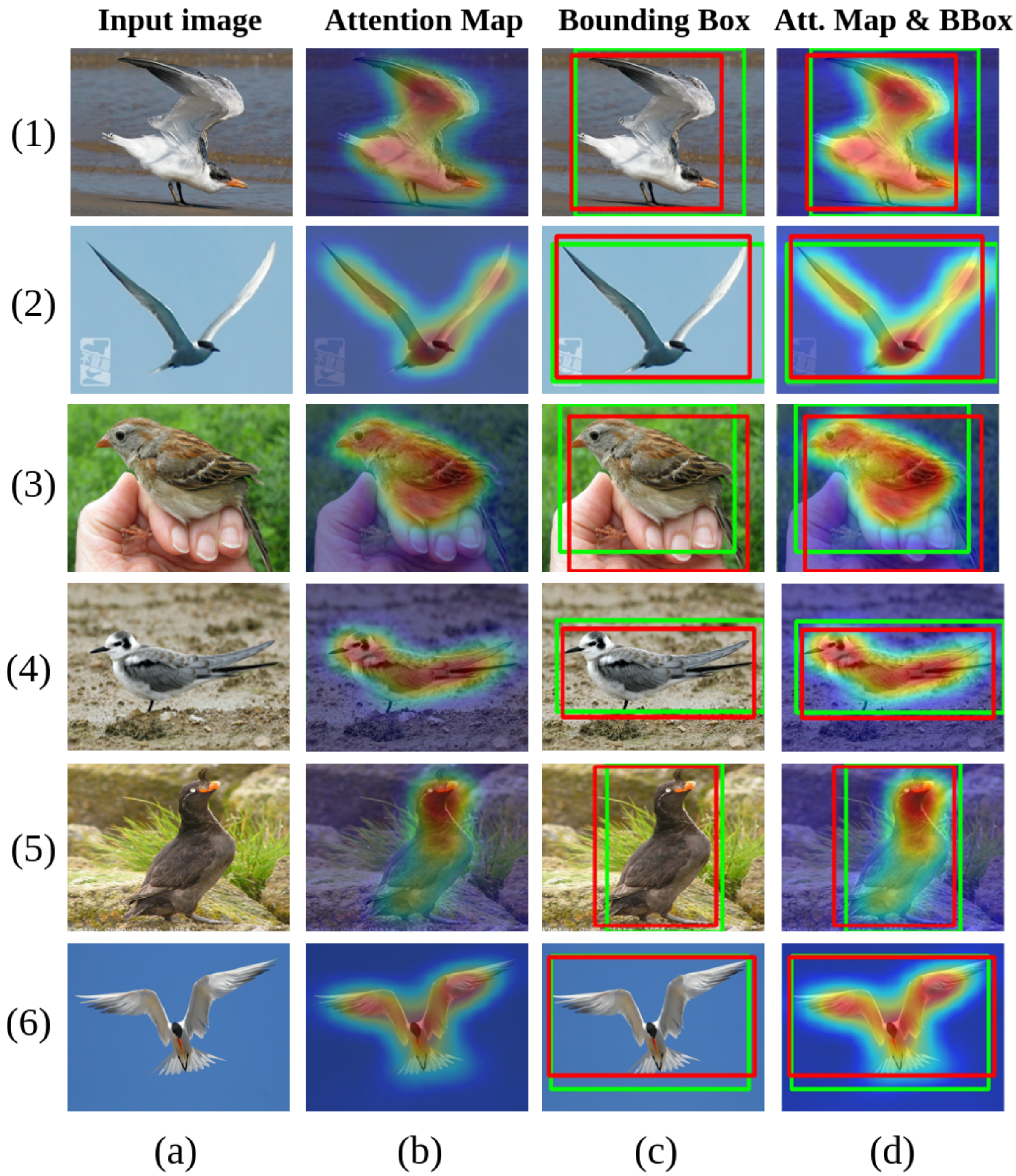

Figure III: **Qualitative Results on the CUB-200-2011 dataset.** The right-most column gives the results (bounding boxes in green) of our proposed model, which are close to that of the ground-truth (in red). Attention maps resemble the entire object being targetted.

References

[1] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020. 2

[2] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019. 2

[3] Krishna Kumar Singh and Yong Jae Lee. Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3544–3553. IEEE, 2017. 1, 2

[4] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019. 2

[5] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018. 2

[6] Xiaolin Zhang, Yunchao Wei, Guoliang Kang, Yi Yang, and Thomas Huang. Self-produced guidance for weakly-supervised object localization. In *Proceedings of the European Conference on Computer Vision*, pages 597–613, 2018. 2

[7] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 2