(Supplementary Material)
Contrastive Attention Maps for Self-supervised Co-localization

Minsong Ki[1] Youngjung Uh[2,3] Junsuk Choe[4*] Hyeran Byun[1,3]

[1]Department of Computer Science, Yonsei University
[2]Department of Applied Information Engineering, Yonsei University
[3]Department of Artificial Intelligence, Yonsei University
[4]Department of Computer Science and Engineering, Sogang University

This supplementary material contains four parts:

- Section A provides the hyperparameter of our pixel-wise top-k attention pooling (PTAP).
- Section B compares performance regarding employing our PTAP for inference.
- Section C shows performance in two ways of the top-k sampling, channel-wise and pixel-wise, for attention pooling.
- Section D indicates localization performances of combined transformations on four benchmark datasets.
- Section E shows more visual results of our method.

A. Hyperparameter of our PTAP

Figure 1 compares $\texttt{CorLoc}^{\texttt{Mean}}$ across different choice of the hyperparameter k for PTAP. $\texttt{CorLoc}^{\texttt{Mean}}$ indicates the average over the $\texttt{CorLoc}^{\texttt{IoUs}}$ at the three IoU thresholds: 0.3, 0.5, and 0.7. Ours with the max pooling uses only the value with the highest weight, and ours with the average pooling uses all of the input channels.

We also report the same comparison on the baseline plus PTAP (blue bar), without contrastive attention map loss. Our method shows the best performance when we selectively aggregate the feature map regarding their magnitudes across the top-70% of input channels.

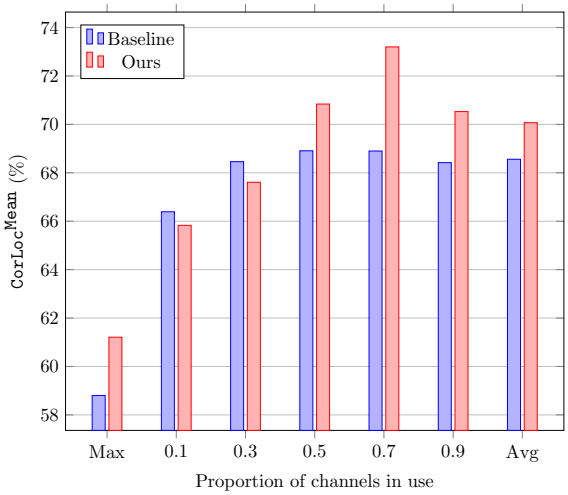

Figure 1: $\texttt{CorLoc}^{\texttt{Mean}}$ comparisons over varying ks for PTAP in the baseline and ours when contrastive attention map loss is computed using rotation (CUB dataset [4]). The horizontal axis denotes relative portion of the chosen channels, *e.g.*, $k = 1433$ for 0.7 (70%) when the feature map has 2048 channels. Max is identical to top-1 case where the highest value across the channel dimension is chosen for each pixel. Avg uses all input channels.

B. Our attention pooling for inference

Our PTAP is inserted into the last convolution layer, and it is activated in both the training and testing phase. Table 1 compares the co-localization performance with and without PTAP for testing (CUB dataset [4]). We observe that employing PTAP at both training and testing improves the co-localization performances by 2.1% when contrastive attention map loss is computed using rotation. It adds only 0.08% of forward time which is five seconds for the entire CUB testset and 12 minutes for the entire CUB trainset.

C. Pixel-wise vs. Channel-wise.

We extract the weighted feature map $\mathbf{F}_w$ and then apply two sampling methods to generate the final attention map. The pixel-wise approach performs average pooling on the

*Work done as a research scientist at NAVER AI Lab.

Table 1: Effectiveness of PTAT at test phase in CorLoc$^{\text{IoUs}}$. Bold texts denote the best performance in each column.

PTAP at train	PTAP at test	CorLoc$^{\text{IoUs}}$ 0.3	0.5	0.7	Mean
✔	✗	97.04	81.29	35.12	71.15
✔	✔	**97.30**	**83.65**	**38.64**	**73.20**

Table 2: CorLoc$^{\text{IoUs}}$ comparisons on our method according to the top-k sampling of PTAP.

Method: rotation	CorLoc$^{\text{IoUs}}$ 0.3	0.5	0.7	Mean
PTAP w/ channel-wise	96.63	80.66	35.27	70.86
PTAP w/ pixel-wise	**97.30**	**83.65**	**38.64**	**73.20**

$\mathbf{F}_w$ by selecting top-k% values for each pixel across channels. In the channel-wise approach, we first obtain the channel priority vector by applying GAP to $\mathbf{F}_w$. Then, we apply average pooling on the $\mathbf{F}_w$ with only the top-k% channels based on the priority vector. In Table 2, we observe that the pixel-wise approach has improved the co-localization performances more effectively than the channel-wise approach.

D. Additional results of combined transformations

In Table 3, we provide the localization performances with combined transformations which are best and second-best on CUB [4]: rotation + scale, scale + translation on four benchmark datasets [1, 2, 3, 4]. Our method shows consistent improvement over the baseline on all datasets. Especially, there are large performance gains in scale and translation tasks where undefined regions occur after transformation. We suppose predicting rotation is the most helpful pretext task for the baseline, compared to other transformations.

E. Additional qualitative results

In Figure 2, we illustrate more examples from our model on four benchmarks: CUB-200-2011, Stanford Cars, FGVC-Aircraft, and Stanford Dogs. Qualitative evaluation results show that our method localizes the full extent of the object and ignores the background.

References

[1] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2, 2011. 2

[2] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 2

[3] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 2

[4] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1, 2

Table 3: $\texttt{CorLoc}^{\texttt{IoU=0.5}}$ comparisons with combined transformations that shows the best and second-best on CUB dataset. R: rotation, S: scale, T: translation.

Task	CUB-200-2011		Stanford Cars		FGVC-Aircraft		Stanford Dogs	
	Baseline	Ours	Baseline	Ours	Baseline	Ours	Baseline	Ours
R + S	79.98	85.88 (+5.90)	91.19	97.26 (+6.07)	91.11	96.60 (+5.49)	76.84	82.82 (+5.98)
S + T	35.98	84.15 (+48.17)	42.86	97.50 (+54.64)	68.22	96.72 (+28.50)	34.86	83.62 (+48.76)

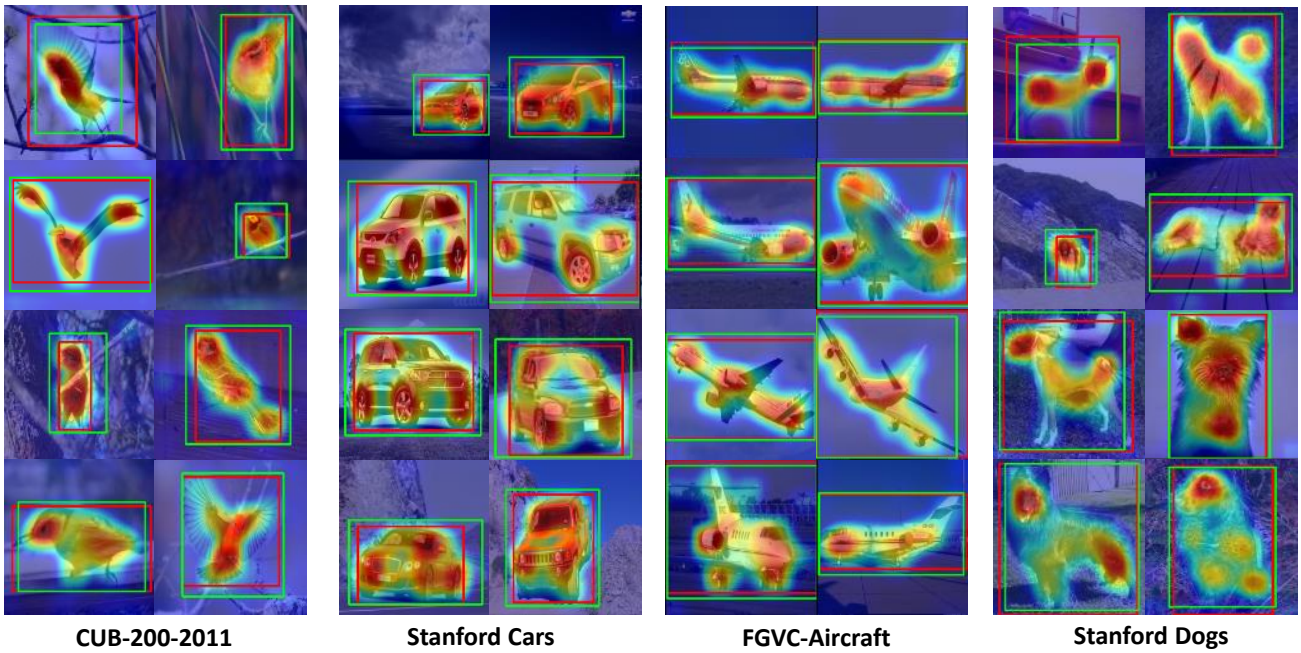

Figure 2: Qualitative examples of activation map and localization produced by our model on the four benchmarks. These maps output with colors ranging from red (higher importance) to blue (lower importance like a background). The red boxes are the ground-truth, and the green boxes are the predicted ones.