Supplementary materials

We present empirical validation about the classifier weights biased on discriminative dimensions in Section A, the impact of different clustering and similarity methods in Section B, the analysis on the number of samples in clustering in Section C, ablation study on the MS COCO dataset in Section D supplementing for Table 1 (main paper), Sensitivity analysis on VOC in Section E supplementing for Figure 5 (main paper), and more qualitative results in Section F supplementing for Figure 4 (main paper).

A. Biased Classifier

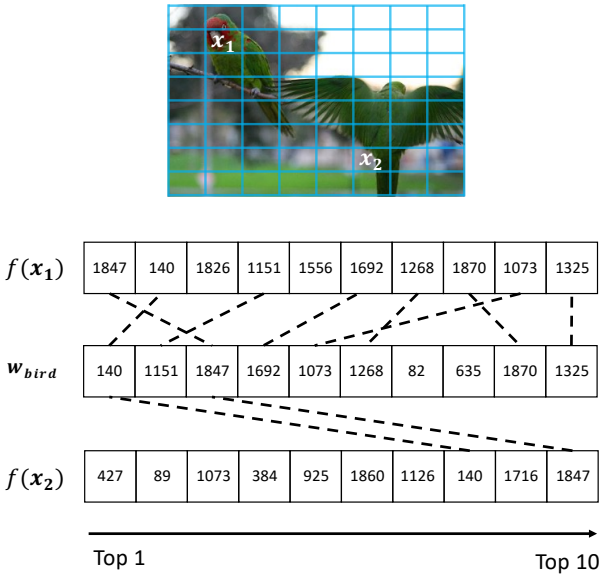

Figure S1. Empirical validation about the classifier weights biased on discriminative dimensions. We show the indices of the top 10 dimensions with the highest value for local region features ($f(\boldsymbol{x_1})$: "head" and $f(\boldsymbol{x_2})$: "tail") and the classifier weight of "bird" ($\mathbf{w}_{bird}$).

We empirically validate that the classifier weights biased on the discriminative dimensions. This is to supplement for Section 3.2 in the main paper.

In Figure S1 We show the indices of the top 10 dimensions with the highest value for local region features ($f(\boldsymbol{x_1})$: "head" and $f(\boldsymbol{x_2})$: "tail") and the classifier weight of "bird" ($\mathbf{w}_{bird}$). The number of overlap dimensions between $\mathbf{w}_{bird}$ and $f(\boldsymbol{x_1})$ is 8, but only 2 between $\mathbf{w}_{bird}$ and $f(\boldsymbol{x_2})$. This validate that the classifier weight of "bird" ($\mathbf{w}_{bird}$) biased on the dimensions of discriminative feature "head".

B. Impact of clustering and similarity methods

We study the impact of different clustering and similarity methods. 1) For clustering, we use K-Means and Hierarchical clustering. On VOC, the seed mask quality (mIoU) of Hierarchical clustering is 55.1% (slightly higher than the 54.9% of K-Means in Table 2), but the running time is around 5 times longer. 2) For similarity, we evaluate Euclidean and Cosine similarities in K-Means. The seed mask quality (mIoU) of using Euclidean on VOC dataset is 54.4%, which is close to that of Cosine (54.9% in Table 2).

C. Number of samples in clustering

As mentioned in Section 4.1 of the main paper: for k-Means clustering on MS COCO, we sample 100 images per class, rather than using the whole dataset of each class (to control the time costs for clustering). Here We study the impact of the number of images per class on the seed mask quality (mIoU) of LPCAM and show the results in Figure S2. There is little performance gain after 100. The possible reason is that for the common object on MS COCO, 100 samples can cover the variants of local features in the class.

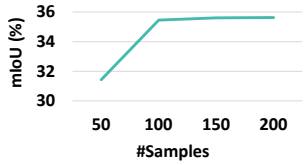

Figure S2. The seed mask quality (mIoU) of LPCAM regarding the number of images per class on MS COCO.

D. Ablation Study on MS COCO

	FP	FN	mIoU	Prec.	Recall
CAM	45.7	21.9	33.1	43.8	64.6
LPCAM-F	49.7+4.0	16.9-5.0	33.9+0.8	42.2-1.6	68.5+3.9
LPCAM	43.5-2.2	21.2-0.7	35.4+2.3	47.1+3.3	64.7+0.1

Table S1. An ablation study on MS COCO dataset. "-F" denotes only the "Foreground" term $\boldsymbol{FG}_n$ is used in Eq. 6 (main paper).

We conduct an ablation study on the MS COCO dataset to evaluate the two terms of LPCAM in Eq. 6 (main paper): foreground term $\boldsymbol{FG}_n$ and background term $\boldsymbol{BG}_n$ that accord to class and context prototypes, respectively. This is to supplement for Table 1 in the main paper. In Table S1, we show the mIoU results (of seed masks), false positive (FP), false negative (FN), precision, and recall. We can see that our method of using class prototypes (LPCAM-F) greatly improve the recalls—3.9% higher than CAM, and thus reduces the rates of FN a lot. This validates the ability of our methods to capture non-discriminative regions of the image. We also notice that LPCAM-F increases the rate of FP

over CAM. The reason is that confusing context features (e.g., "railroad" for "train") may be wrongly taken as class features. Fortunately, when we explicitly resolve this issue by applying the negative context term $-\boldsymbol{BG}_n$ in LPCAM, this rate can be reduced (by 6.2% for MS COCO), and the overall performance (mIoU) can be improved (by 1.5% for MS COCO).

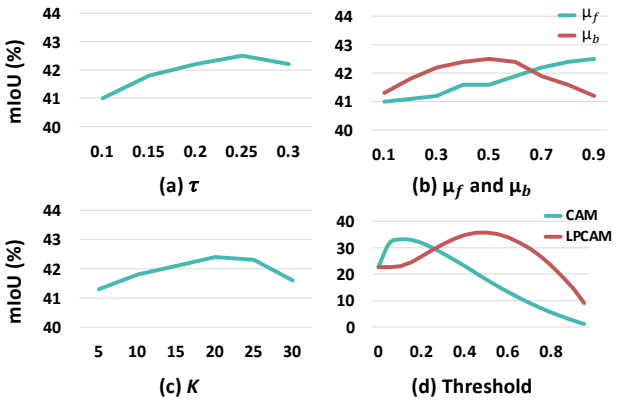

Figure S3. Sensitivity analysis on MS COCO, in terms of (a) τ for dividing foreground and background local features, (b) μ_f for selecting class prototypes and μ_b for selecting context prototypes, (c) the number of clusters K in k-Means, and (d) the threshold used to generate 0-1 seed masks from heatmaps.

E. Sensitivity Analysis on MS COCO

In Figure S3, we show the quality (mIoU) of generated seed masks when plugging LPCAM in AMN on VOC dataset. We perform hyperparameter sensitivity analyses by changing the values of (a) the threshold τ for dividing foreground and background local features, (b) the threshold μ_f for selecting class prototypes and the threshold μ_b for selecting context prototypes, (c) the number of clusters K in K-Means, and (d) the threshold used to generate 0-1 seed mask (a common hyperparameter in all CAM-based methods). Figure S3(a) shows that the optimal value of τ is 0.25. Adding a small change does not make any significant effect on the results, e.g., the drop is less than 1% if decreasing τ to 0.15. We use a higher τ on MS COCO because the quality of CAM on MS COCO is poorer (than VOC) and a higher value can filter out noisy activation. Figure S3(b) shows that the optimal values of μ_f and μ_b are 0.9 and 0.5, respectively. The gentle curves show that LPCAM is little sensitive to μ_f and μ_b. This is because classification models (trained in the first step of WSSS) often produce overconfident (sharp) predictions, i.e., output probabilities are often close to 0 or 1. It is easy to set thresholds (μ_f and μ_b) on such sharp values. In Figure S3(c), the best mIoU of seed mask is 42.5% when K=20, and it drops by only 0.8 percentage points when K goes up to 30. In Figure S3(d), LPCAM shows much gentler slopes than CAM around their respective optimal points, indicating its lower sensitivity to the changes of this threshold.

F. Qualitative Results on VOC

Figure S4 shows qualitative examples where LPCAM leverages both discriminative and non-discriminative local features to generate heatmaps and 0-1 masks on VOC dataset. In both single-object images ("cow", "boat", "car", and "bird") and multi-objects images ("horse" and "cat"), CAM focuses on only discriminative features e.g., the "head" regions of "cow", while our LPCAM has better coverage on the non-discriminative feature, e.g., the "body" and "leg" regions. In the "car" example, the context prototype term $-\boldsymbol{BG}_n$ in Eq. 6 (main paper) helps to remove the context "plants". In the last two examples, we show two failure cases: LPCAM succeeds in capturing more object parts of "train" and "TV Monitor" but unnecessarily covers more on the context "railroad" and "keyboard". We think the reason is the strong co-occurrence of "train" and "railroad" in the images of "train" ("TV Monitor" and "keyboard" in the image of "TV Monitor").

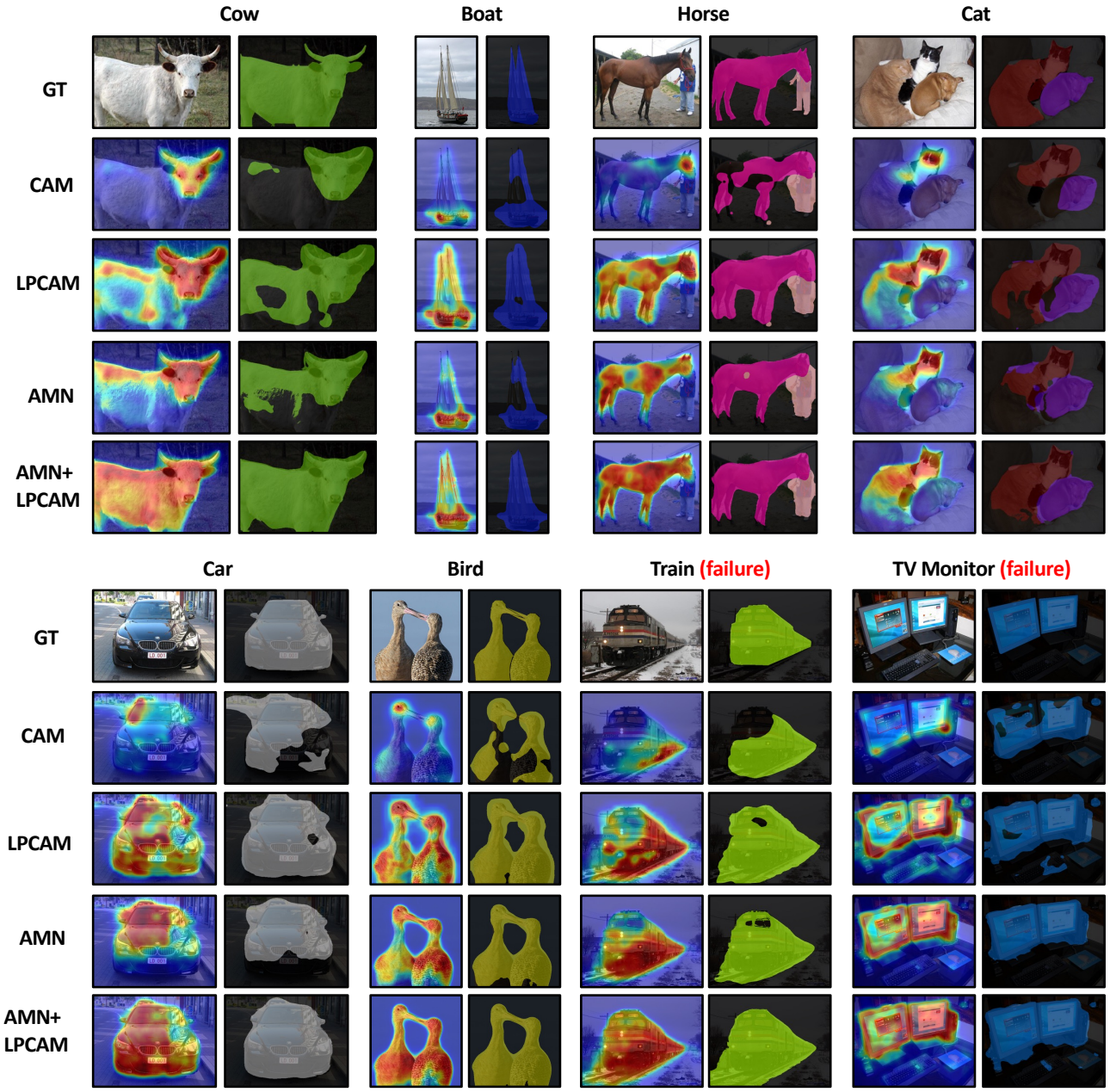

Figure S4. Qualitative results on VOC. In each example pair, the left is heatmap and the right is seed mask.