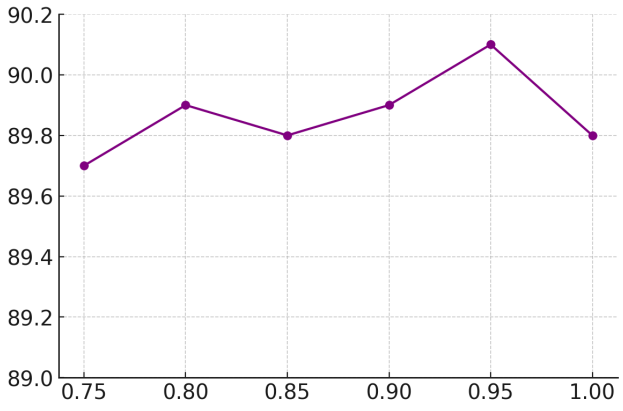
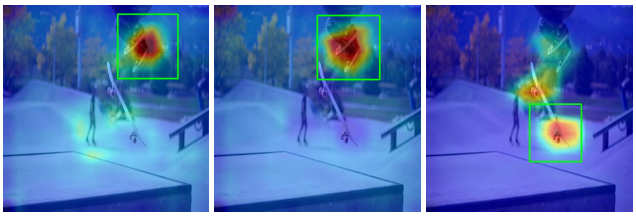
Classifier-guided CLIP Distillation for Unsupervised Multi-label Classification

Supplementary Material



(a) Ablation about CAM threshold

Figure 7. Ablation study on CAM threshold. The performance remains consistent around an mAP of 90%, with the peak near a threshold value of 0.95.



(a) skis (b) snowboard (c) skateboard

Figure 8. CAM corresponding to different classes. The similar class co-occur with human-feet, resulting similar CAM around the feet.

6. Detailed Analysis

6.1. Ablation of CAM threshold

Fig. 7 illustrates the changes in mAP relative to variations in the CAM threshold. The peak performance is observed at a threshold value of 0.95, with consistently high performance around this point. By using CAM-based local patch acquisition, we can sample patches near around regions of interest. As the threshold decreases, performance slightly declines due to the larger local patches, which include more and more regions. This transition shifts from sampling *around GT boxes* to *randomly sampled boxes* which visualized in the proof-of-concept study. This ablation result aligns with the proof-of-concept study, demonstrating that sampled boxes around GT reflect the key outcomes.

6.2. Per-class Performance

Tab. 4 shows the per-class average precision on the PASCAL VOC 2012 dataset. We investigate the impact of the CLIP debiasing and observe that the bias correction is effective for boosting the performance of biased classes. Furthermore, the performance of unbiased classes also improves alongside biased classes. This is attributed to correcting mispredicted biased classes, which would otherwise introduce noise when predicted as other classes. However, the amount of improved performance differs depend on the classes. Regarding the performance of *bottle*, *sofa*, and *tv*, the low performance for *bottle* can be attributed to its greater CLIP bias (relatively lower probabilities). In addition, the fact that *bottle* frequently appears but its detection rate remains low as illustrated in Fig. 10. On the other hand, *sofa* and *tv* classes captured as top-1 are similar to GT quantities.

Compared to CDUL, our CCD outperforms all classes except *bird* and *person*. This might be due to the gradient-alignment network training method from CDUL, which updates the labels during the whole training process with training loss. Since the person label presents the most frequently in the PASCAL VOC dataset (shown in Fig. 10), the training process is likely to update the person label to "Positive".

Tab. 5 shows the per-class average precision on the MS COCO dataset. Due to undisclosed hyperparameters, we cannot reproduce the CDUL. Thus, we only compare the results of CCD and CCD without debiasing. Our proposed CLIP debiasing successfully boosts the performance of biased classes.

6.3. Bounding Box with Class Activation Mapping

Our method generates CAM bounding boxes for classes that exceed a certain threshold, allowing for multiple bounding boxes to be obtained around objects. Figure 11 shows sample bounding boxes extracted from an image in the PASCAL VOC 2012 dataset. The first row represents the best-case scenario, where multiple bounding boxes accurately surround the target classes. The second row illustrates a mixed scenario, with half of the boxes correctly identifying the target classes and the other half missing them. The third row represents a failure case, where only a single bounding box is close to the target object. These bounding boxes are then cropped and passed through CLIP to generate local labels.

Table 4. AP and mAP (in %) of unsupervised methods on PASCAL VOC 2012 dataset for all classes. The best score is in bold.

Methods	aero	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	bike	person	plant	sheep	sofa	train	tv	mAP
CDUL	99.0	92.7	**97.7**	91.8	72.5	95.4	84.7	98.6	76.4	91.9	73.2	97.1	92.0	94.1	**93.0**	67.5	94.2	74.2	97.7	89.0	88.6
w/o Debias	98.4	92.8	97.6	91.5	77.0	96.0	83.9	98.7	77.1	93.4	72.7	96.5	96.1	93.7	87.3	68.8	94.0	73.0	98.3	89.5	88.8
CCD (ours)	**99.1**	**93.6**	97.6	**92.4**	**77.6**	**96.0**	**86.1**	**99.0**	**79.0**	**95.2**	**74.8**	**97.7**	**96.0**	**95.4**	89.2	**72.1**	**95.3**	**76.6**	**98.1**	**92.0**	**90.1**

Table 5. AP and mAP (in %) of unsupervised multi-label classification on MS COCO 2014 dataset for all classes.

Class	w/o DB	Ours	Class	w/o DB	Ours	Class	w/o DB	Ours	Class	w/o DB	Ours	Class	w/o DB	Ours
person	81.7	82.0	horse	90.4	91.7	baseball bat	89.4	89.0	carrot	48.7	64.1	microwave	63.4	65.8
bicycle	62.9	69.8	sheep	94.7	94.3	baseball glove	83.0	86.9	hot dog	67.3	69.5	oven	69.3	67.6
car	59.2	61.3	cow	87.4	89.4	skateboard	95.8	96.2	pizza	92.1	93.6	toaster	1.9	5.7
motorcycle	86.3	88.8	elephant	97.4	98.0	surfboard	93.0	94.2	donut	76.1	79.2	sink	83.8	83.4
airplane	95.5	96.8	bear	95.8	94.8	tennis racket	98.6	98.6	cake	77.4	75.5	refrigerator	68.4	70.1
bus	81.3	81.7	zebra	97.8	99.0	bottle	40.1	45.0	chair	50.3	54.2	book	28.2	26.0
train	94.7	95.5	giraffe	97.9	98.9	wine glass	56.5	64.1	couch	68.3	71.7	clock	74.6	76.2
truck	55.9	54.8	backpack	20.7	32.9	cup	36.2	38.0	potted plant	37.5	45.3	vase	65.8	67.4
boat	81.9	85.9	umbrella	74.8	78.7	fork	42.1	53.3	bed	79.5	79.6	scissors	47.6	49.8
traffic light	74.2	76.9	handbag	20.6	25.7	knife	32.8	38.6	dining table	48.8	48.1	teddy bear	77.2	84.2
fire hydrant	78.1	79.2	tie	68.5	70.3	spoon	33.0	40.2	toilet	95.6	96.3	hair drier	27.7	22.2
stop sign	72.3	73.1	suitcase	62.8	65.6	bowl	34.7	38.5	tv	74.6	77.4	toothbrush	57.3	59.5
parking meter	60.8	63.7	frisbee	88.1	90.9	banana	74.5	77.6	laptop	80.5	84.2			
bench	52.0	53.6	skis	91.6	91.3	apple	45.9	46.9	mouse	48.1	45.9			
bird	69.4	75.5	snowboard	72.8	75.9	sandwich	70.0	73.9	remote	49.2	58.3			
cat	91.4	93.6	sports ball	33.9	30.1	orange	52.6	62.1	keyboard	75.3	77.1	mean	67.7	70.3
dog	76.8	81.9	kite	95.0	93.9	broccoli	89.7	90.3	cell phone	53.0	56.7			

7. Dataset

7.1. Probability Distribution

Fig. 9 illustrates the CLIP probability distribution for MS COCO and NUSWIDE datasets. It is evident that the mean probability of NUSWIDE is comparatively lower than that of MS COCO. This discrepancy likely contributes to the lower classifier performance observed in NUSWIDE compared to MS COCO. Additionally, we note that the same "class name" exhibits similar low mean probability across both datasets (*e.g.* the probability of *person* is 0.63 for COCO, 0.67 for NUSWIDE). This observation underscores the presence of bias inherent to the CLIP model and text embedding, irrespective of the dataset.

7.2. Label Distribution

Fig. 10 shows the label distribution for PASCAL VOC 2012. We can observe that the label distribution of GT and Single Positive Label (SPL) is similar. However, the CLIP generated label shows does not reflect GT label's distribution. In particular, the number of *person* prediction is relatively lower. This result combine with the CLIP bias, making the prediction of biased class harder. This results support that CLIP shows biased prediction.

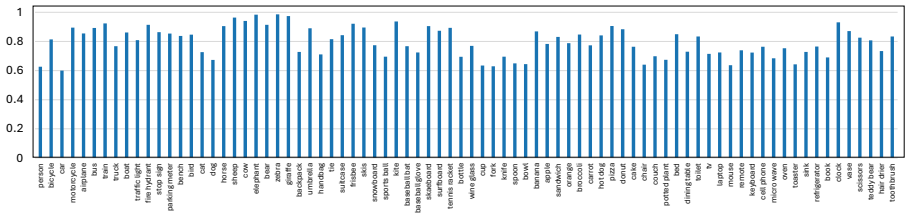
8. Limitation

A limitation of proposed method is that the generated local views often struggle to distinguish between co-occurring objects, an issue inherited from CAM. In Fig. 8, the activation map of *skies* and *snowboard* focus on the human feet. This phenomenon makes them potential candidates for local views. This redundant information results in noisy pseudo-labels, reducing overall performance. Moreover, in Fig. 11, the second row shows the *bottle* (the 1st image) or *dinning table* (the 4th image) generates false local patches. Developing an improved local view proposal method to better handle noisy inferences and focusing on the target objects will be an important direction for future work.

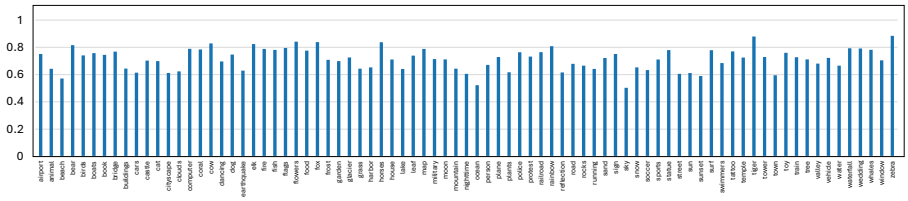
In addition, it is evident from Tab.1 in main text that all unsupervised multi-label classification methods exhibit relatively degraded performance on MS COCO and NUSWIDE datasets. Both datasets share the characteristic of having a larger number of classes (80 and 81, respectively) compared to the PASCAL VOC datasets (20 classes). This performance degradation can be attributed to the simplistic prompt, "a photo of the [class name]," which might not accurately represent the target classes. For instance, the performance of initial labels of the training sets for PASCAL VOC 2012, MS COCO, and NUSWIDE is 85.3%, 65.4%, and 41.2%, respectively.

This suggests that CLIP predictions already show varying performance across datasets with different class counts (Although MS COCO and NUSWIDE have similar class counts, the inclusion of "none images" in NUSWIDE exacerbates the degradation of CLIP performance). To address this limitation, manipulating the text embeddings of each dataset may be necessary.

For "none images," the pseudo-labels should be set to zero. To address this, we suggest filtering out the probability of "none images," as a possible solution. Specifically, we can identify irregular and noisy probability patterns from those images and then apply thresholding to filter them out.



(a) CLIP probability distribution at MS COCO



(b) CLIP probability distribution at NUSWIDE

Figure 9. CLIP probability distribution for other datasets: (a) The mean class-wise probability of MS COCO. (b) The mean class-wise probability of NUSWIDE. The probability distribution showcases the presence of CLIP bias in other datasets.

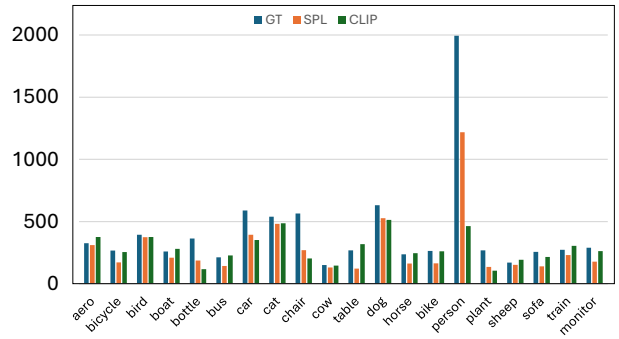


Figure 10. The label distribution of GT, Single Positive Label, CLIP generated at PASCAL VOC 2012. For the CLIP generated label, we count top-1 as predicted. Single Positive setting reflects the label distribution of GT labels, which cannot reflect the real behavior of labelers.

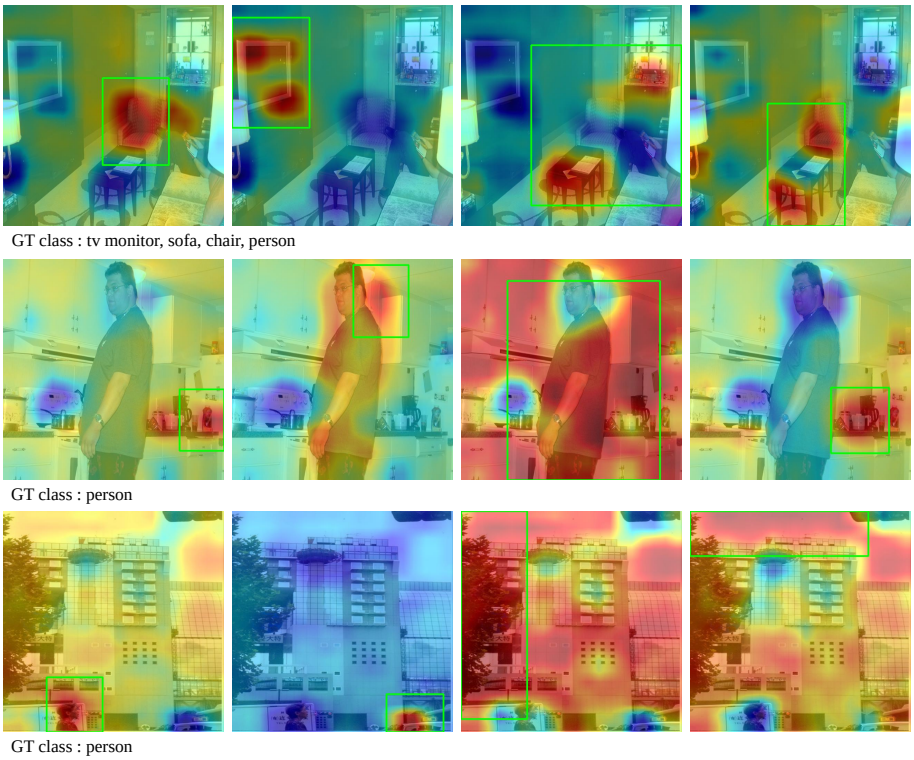


Figure 11. The sample bounding box with Class Activation Mapping. The first row shows optimal case, where every boxes capture gt-related objects. The second row illustrates mixed case, where half of the boxes captures *person* class. The third row depict failure case, where only one box contains *person* class. These multiple local inferences around objects help make better pseudo-labels.