Classifier-guided CLIP Distillation for Unsupervised Multi-label Classification


Dongseob Kim[*1]

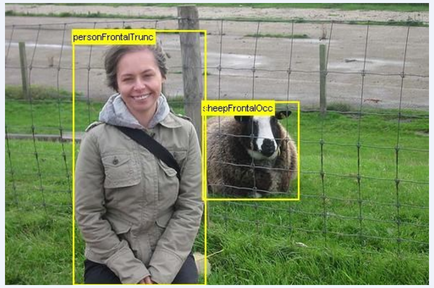

Hyunjung Shim[†2]

[1]*Samsung Electronics,*
[2]*Korea Advanced Institute of Science & Technology*

[†] indicates a corresponding author.

Unsupervised Multi-label Classification

Fully Supervised Multi-label Classification

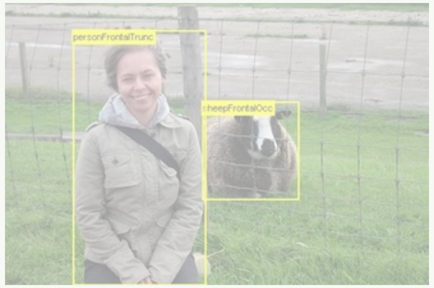


Image

Person
Sheep

Image-level label

Unsupervised Multi-label Classification



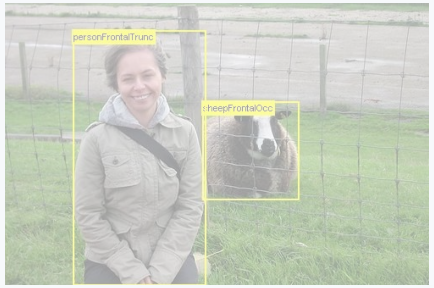
Image

no label

Our research introduces a novel method for improving the performance of CLIP-based unsupervised multi-label classification.

Unsupervised Multi-label Classification

Fully Supervised Multi-label Classification



Person
Sheep

Image | **Image-level label**

Unsupervised Multi-label Classification

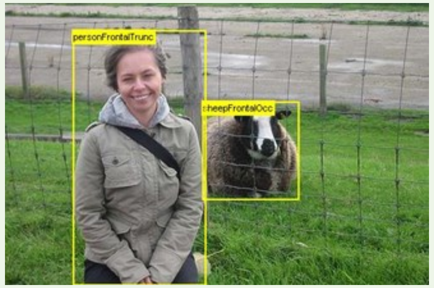
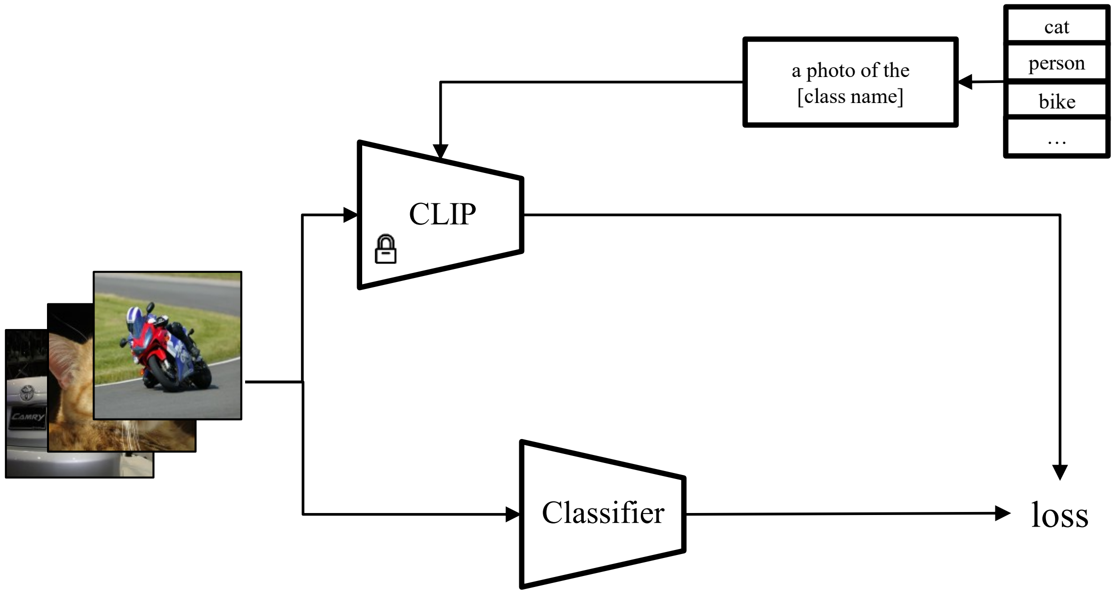


Image | **No label**

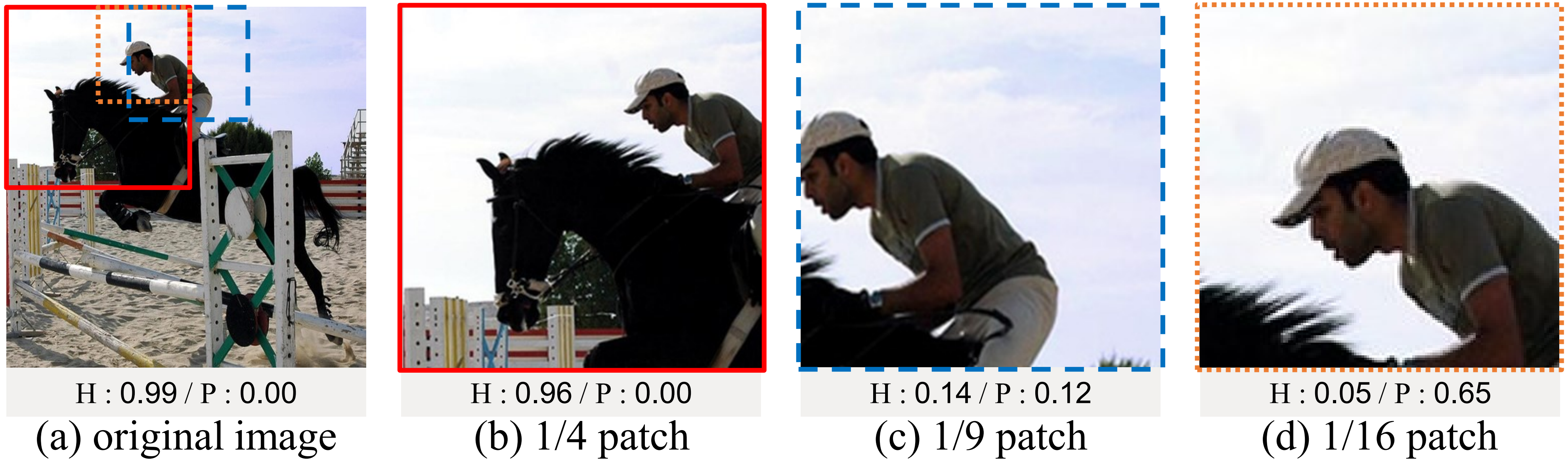
Multi-label classification is essential but presents significant challenges due to the high cost and complexity of precise labeling.

CLIP based Unsupervised Multi-label Classification



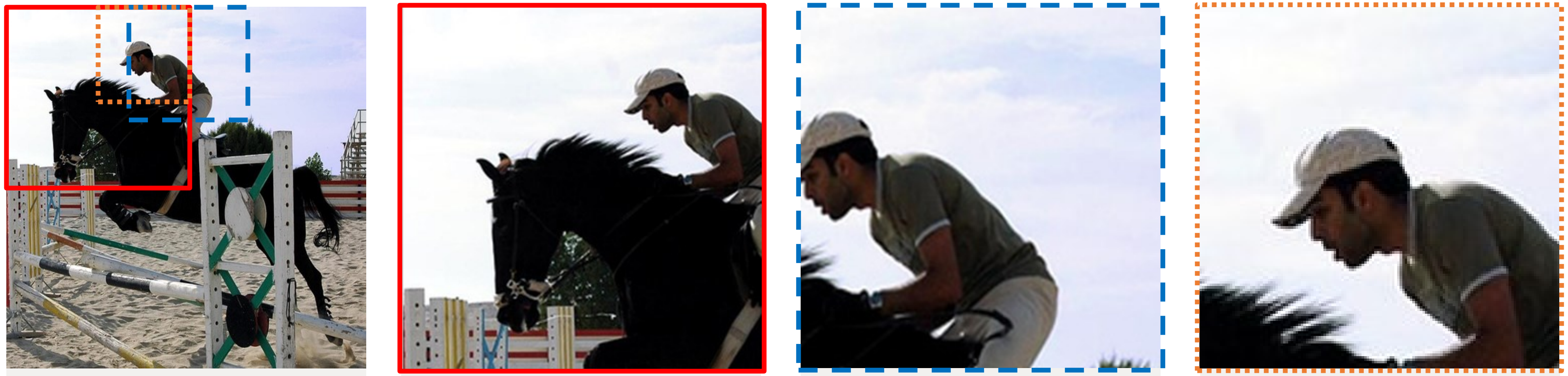
The fundamental methodology involves generating pseudo-labels via CLIP and utilizing these labels to train a classifier.

Critical Limitations of Existing Methods



We identified two critical issues with existing approaches. The first is significant variability in CLIP predictions depending on input views.

Critical Limitations of Existing Methods



H : 0.99 / P : 0.00 (a) original image

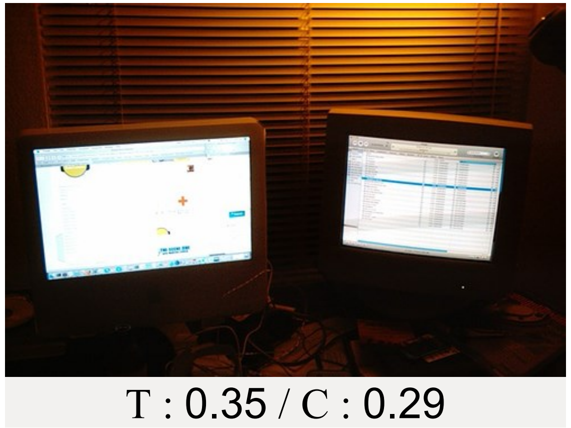
H : 0.96 / P : 0.00 (b) 1/4 patch

H : 0.14 / P : 0.12 (c) 1/9 patch

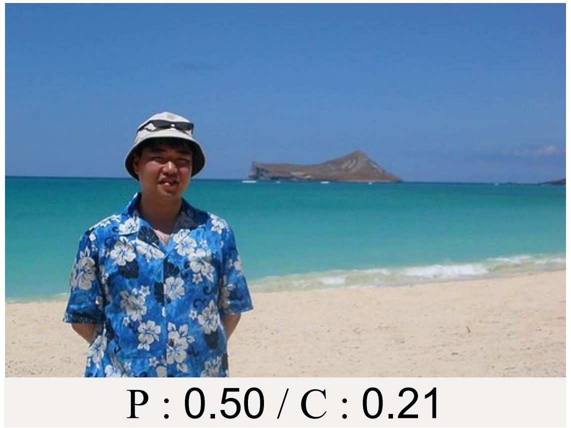
H : 0.05 / P : 0.65 (d) 1/16 patch

Particularly in cases (c) and (d), performance improves notably when a person is centrally positioned and prominently displayed.

Critical Limitations of Existing Methods

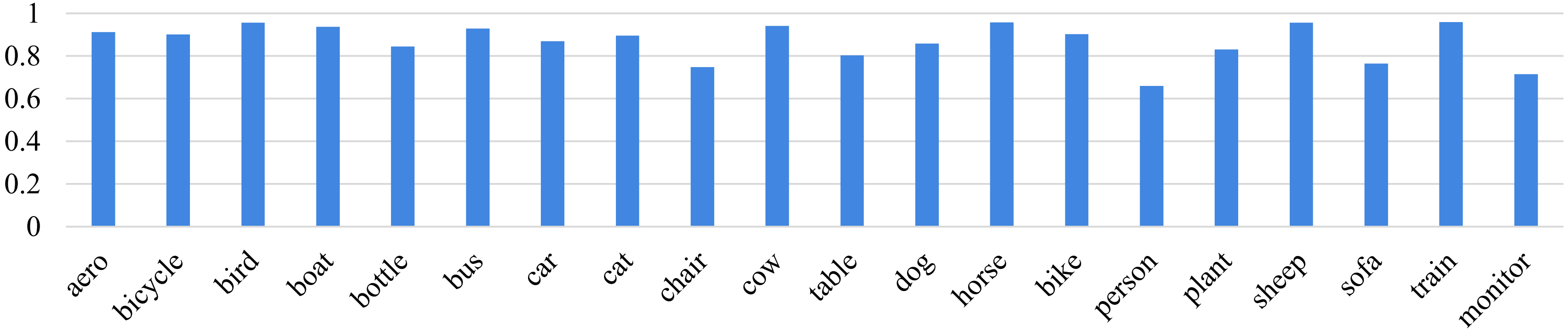


(a) tv monitor (T)



(b) person (P)

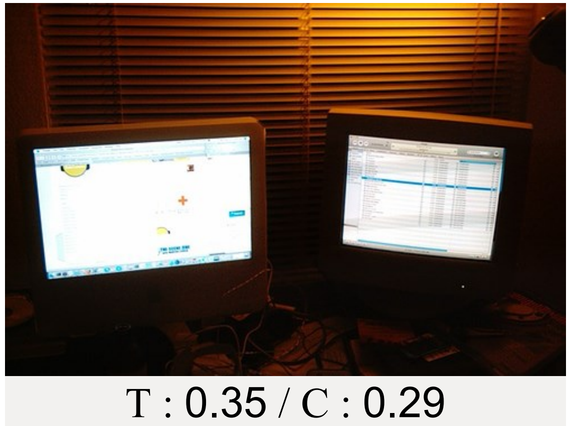
(c) horse (H)



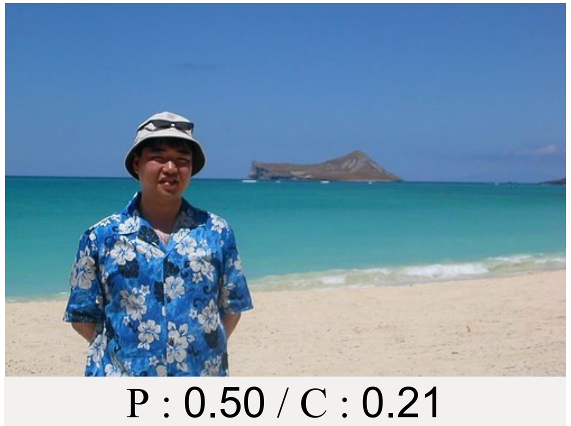
(d) CLIP probability distribution at VOC 2012

Additionally, due to polysemy, semantic embedding spaces are not adequately formed for certain classes, such as person, man, woman, and traveler.

Critical Limitations of Existing Methods

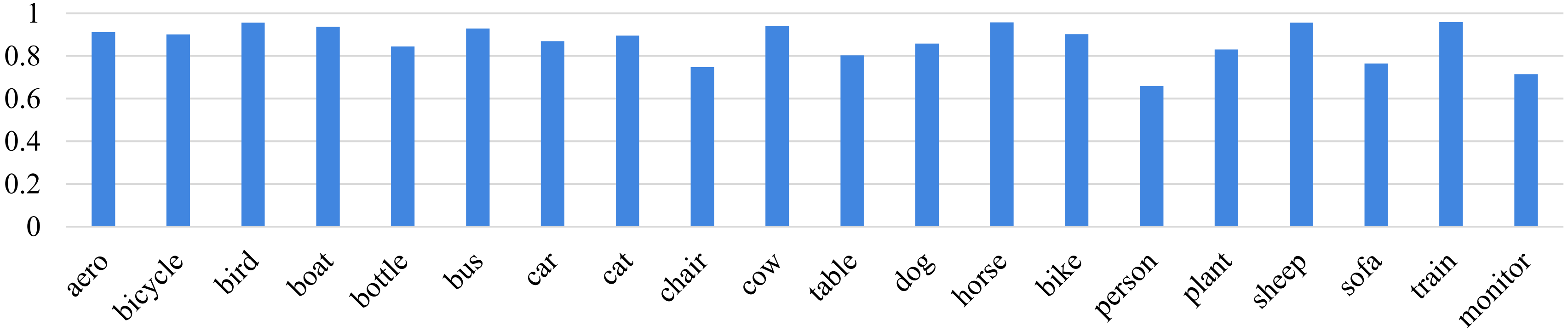


(a) tv monitor (T)



(b) person (P)

(c) horse (H)

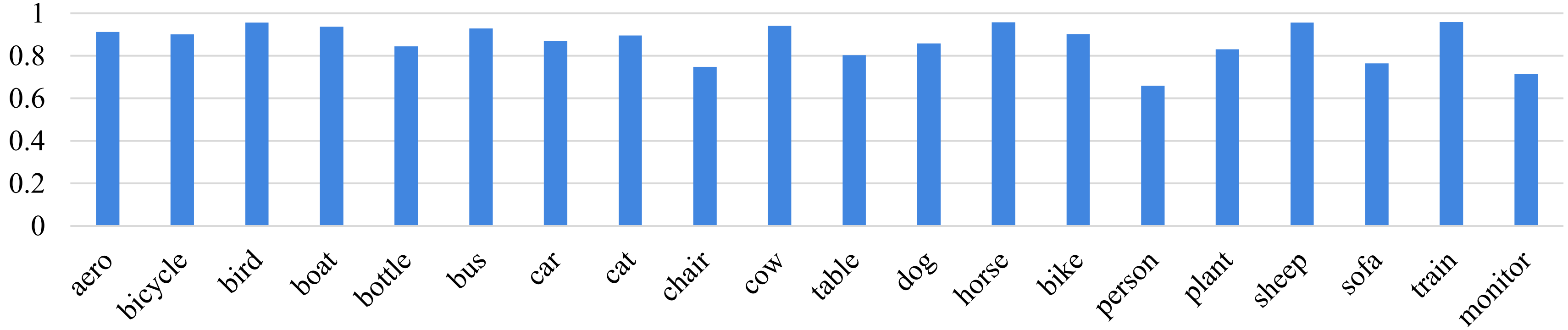


(d) CLIP probability distribution at VOC 2012

Consequently, image (b), which contains only a person, may receive a relatively lower similarity score, whereas image (c), featuring a horse, obtains higher similarity due to fewer synonyms.

Critical Limitations of Existing Methods

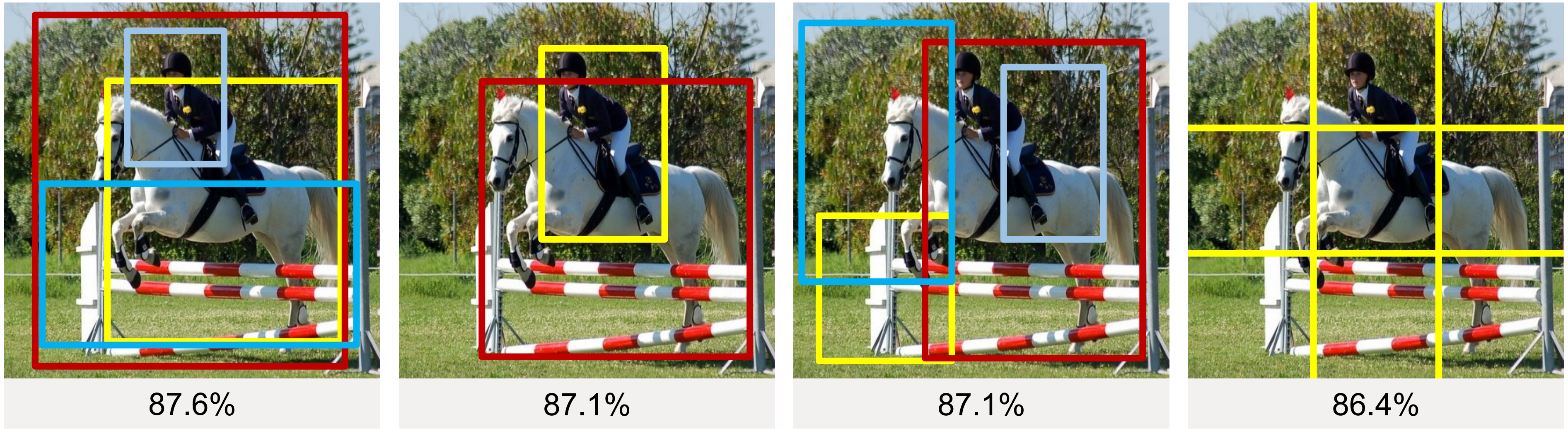
(a) tv monitor (T) (b) person (P) (c) horse (H)



(d) CLIP probability distribution at VOC 2012

Analyzing class-wise probabilities across the entire dataset reveals inherent biases as (d), leading to performance degradation in the final classification results.

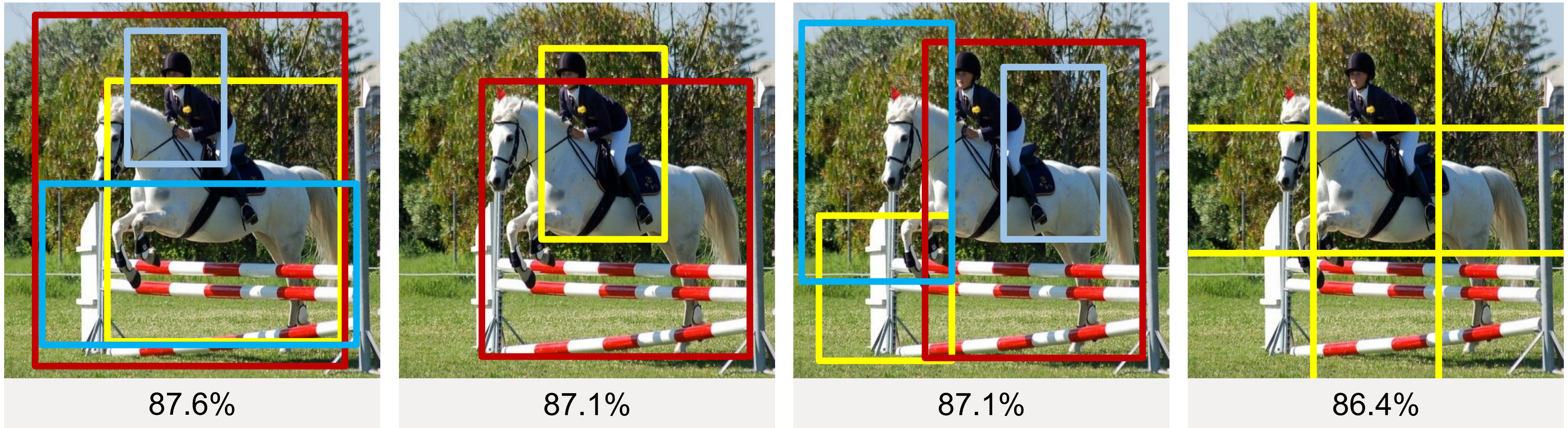
Key observation



(a) Around GT box (b) GT box (c) Random box (d) Uniform grid box

To address the significant variability in prediction probabilities due to view differences, we conducted a toy experiment.

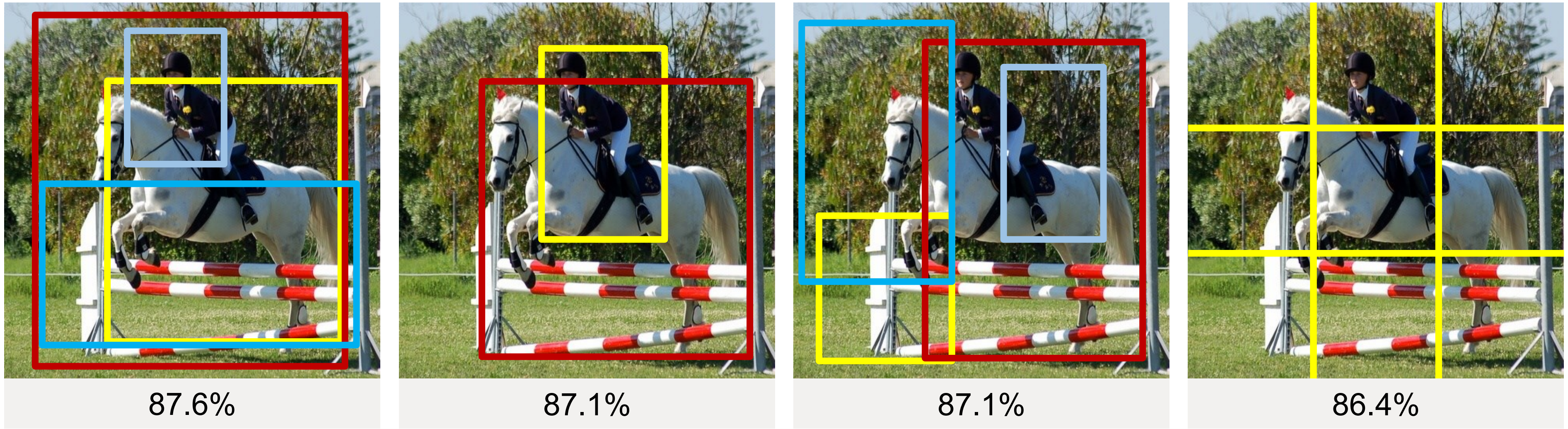
Key observation



(a) Around GT box (b) GT box (c) Random box (d) Uniform grid box

The uniform grid approach showed the poorest performance,
whereas the results from ground truth boxes (b) and completely random boxes (c) were similar.

Key observation



(a) Around GT box (b) GT box (c) Random box (d) Uniform grid box

The best results were obtained by randomly cropping near GT boxes,
leading us to consider methods to randomly sample around these GT boxes.

Key observation

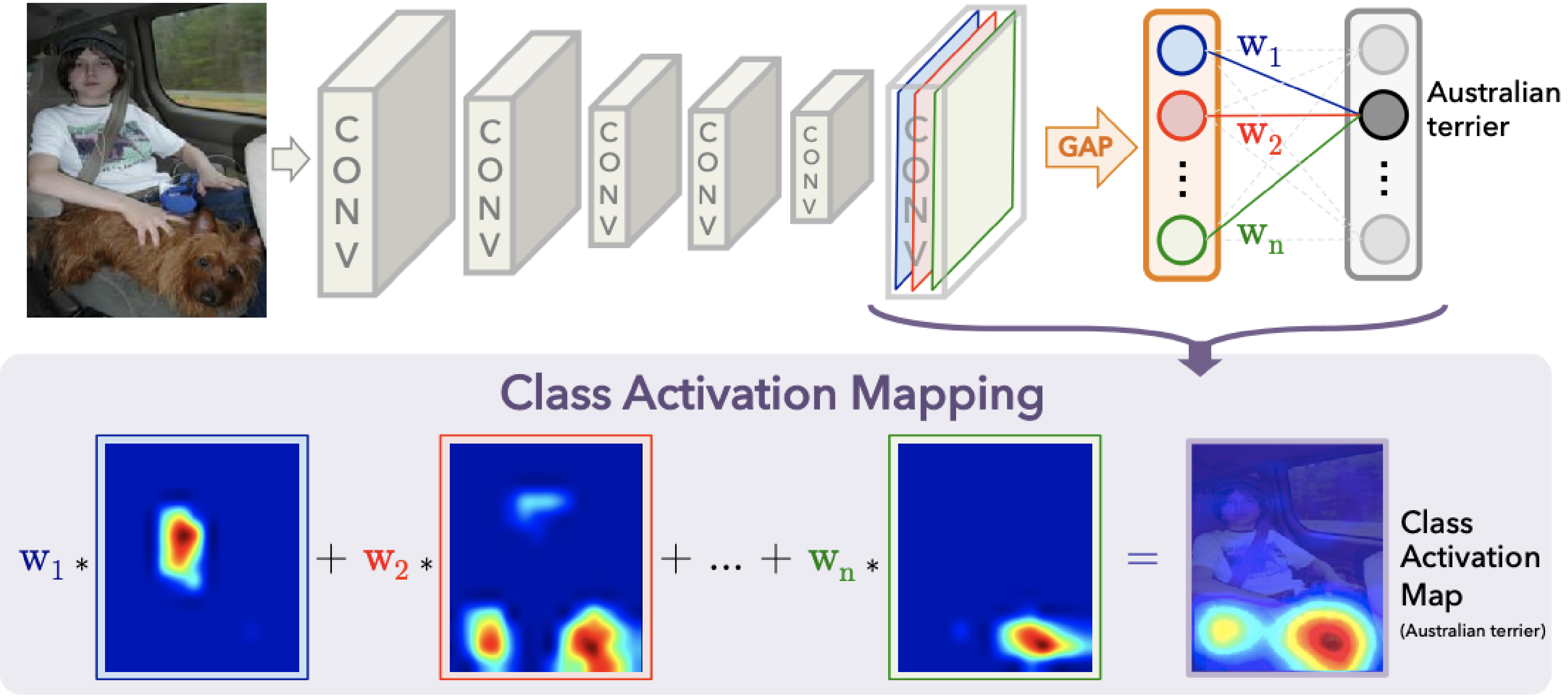
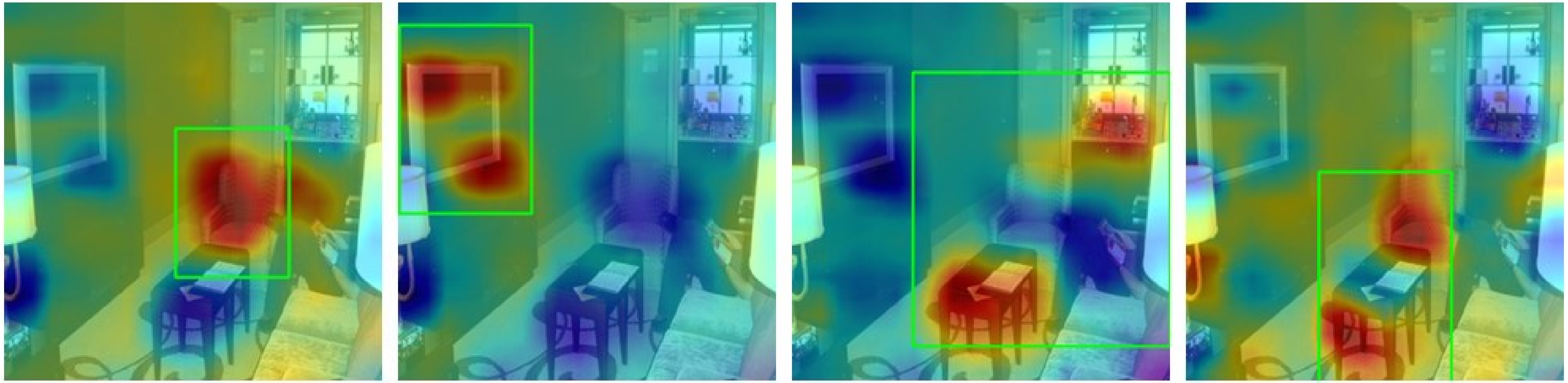


Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Image source : Zhou, Bolei, et al. "Learning deep features for discriminative localization." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

This led us to leverage Class Activation Mapping (CAM) from the classifier being trained.

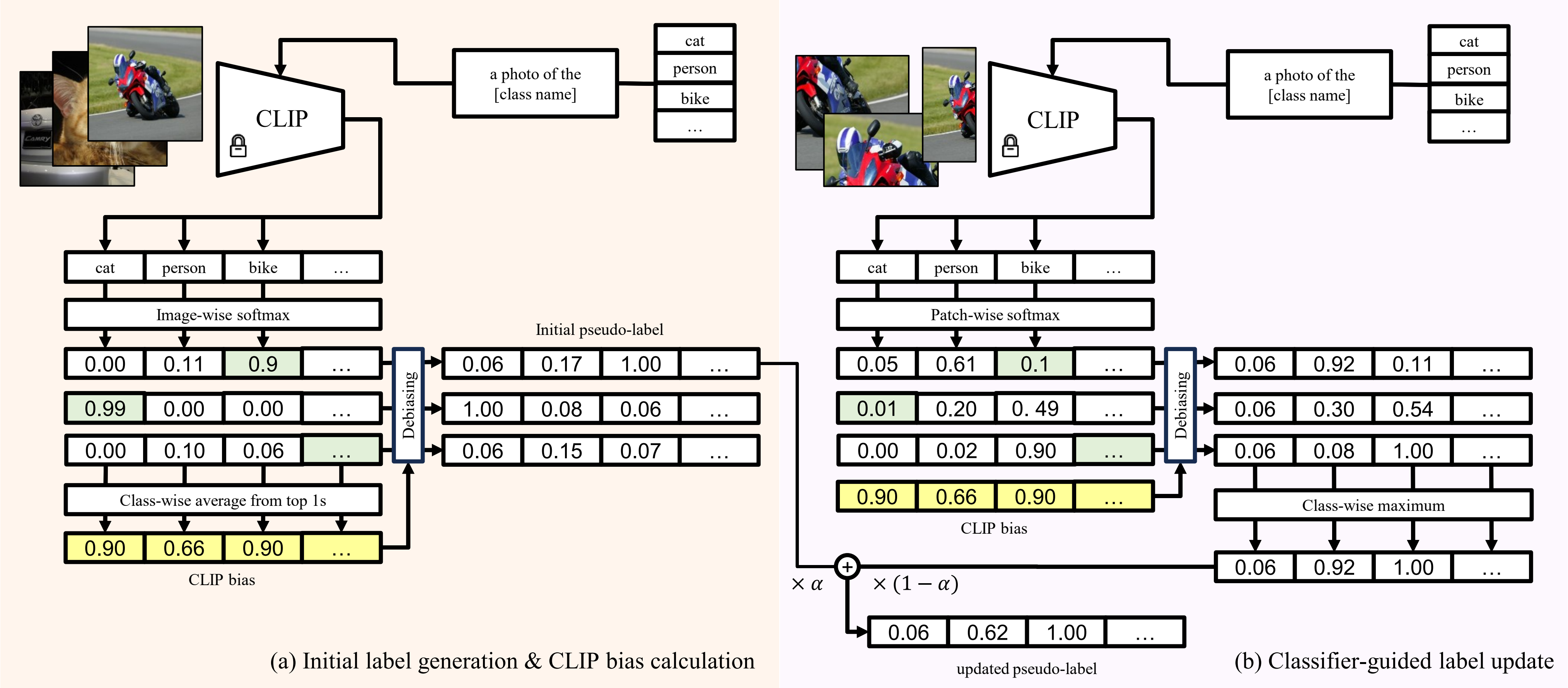
Key observation



GT class : tv monitor, sofa, chair, person

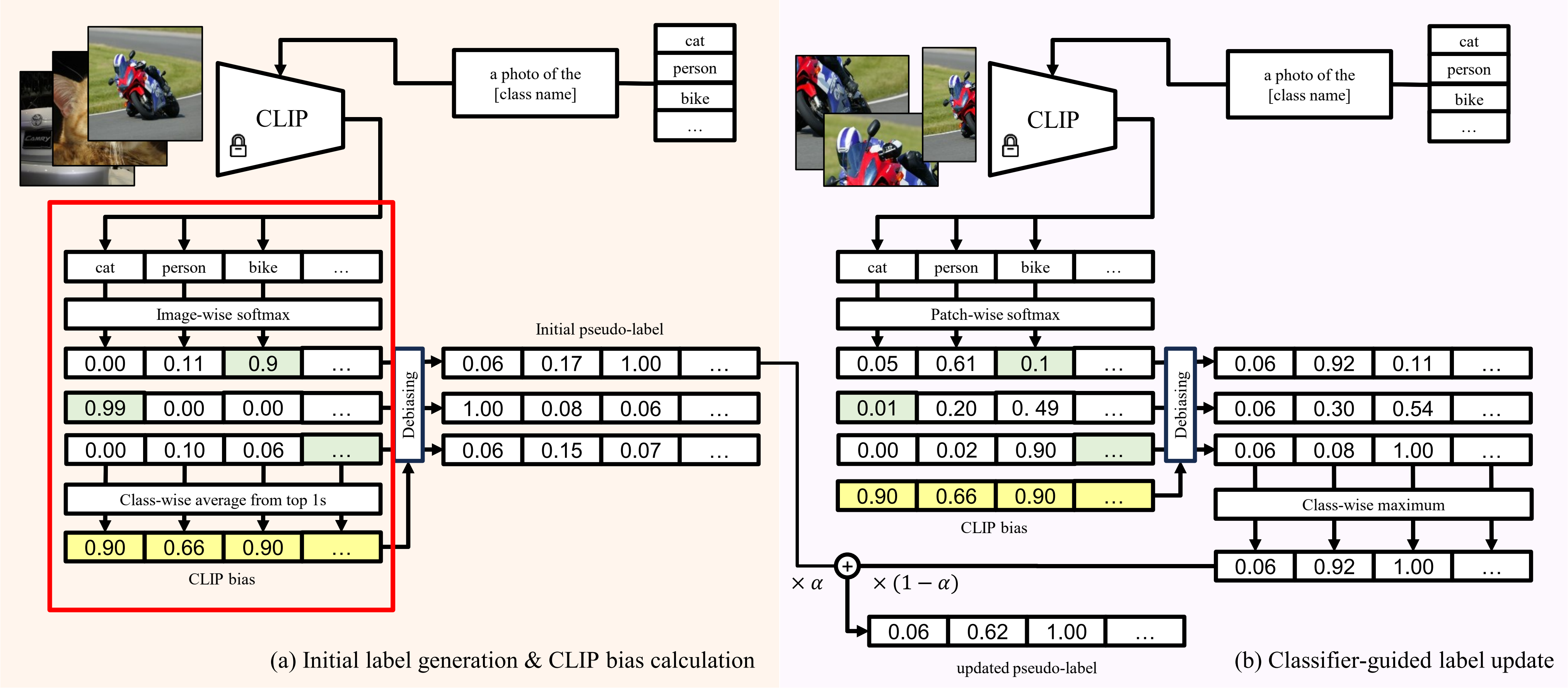
We obtained CAM from a classifier warmed up using initial CLIP-derived pseudo-labels,
subsequently cropping and re-feeding these patches into CLIP to update pseudo-labels.

Method



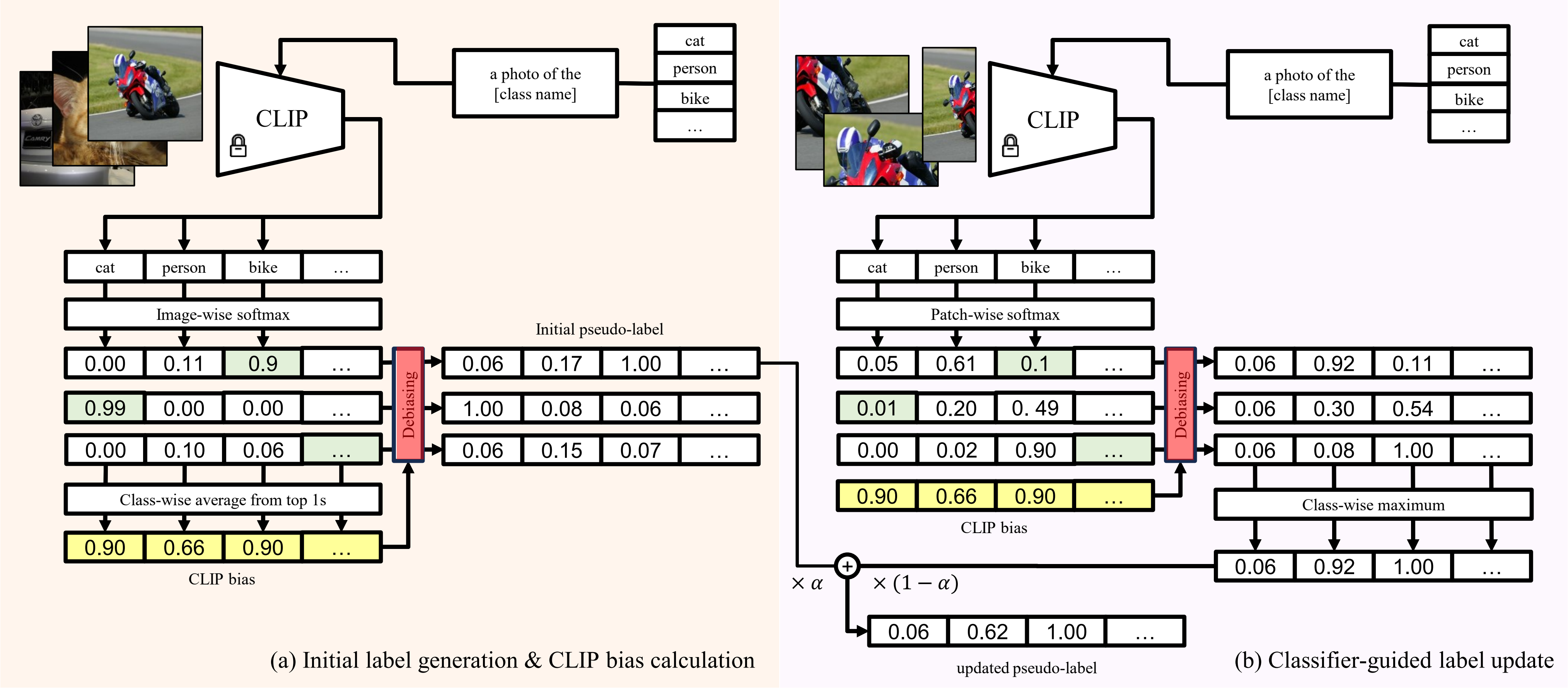
The complete pseudo-label updating process starts by acquiring initial pseudo-labels, obtaining classifier-guided patches, and finally deriving updated pseudo-labels from these.

Method



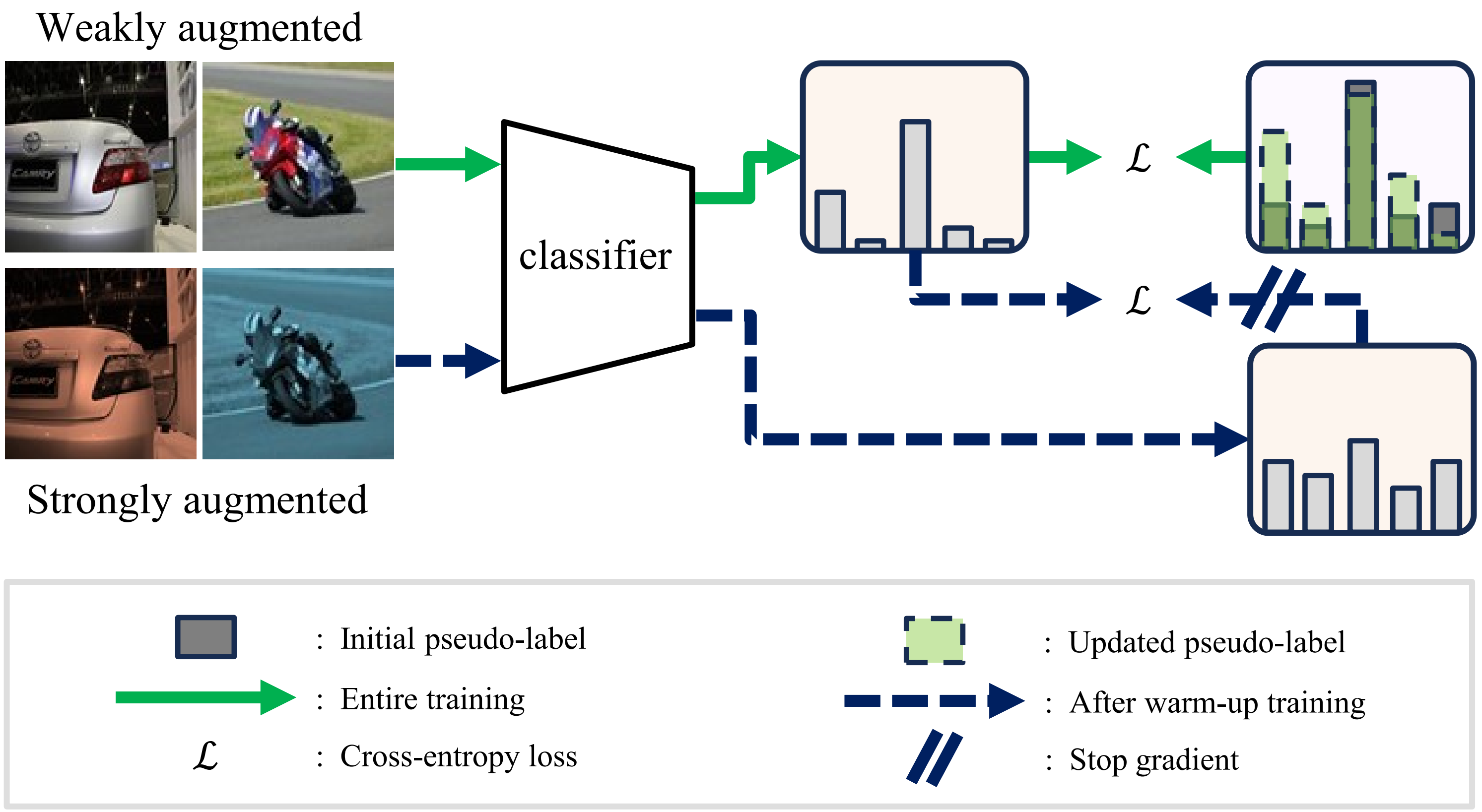
During this process, we gathered pseudo-labels for the entire training dataset, extracting top-1 probabilities and aggregating them class-wise to define a CLIP bias.

Method



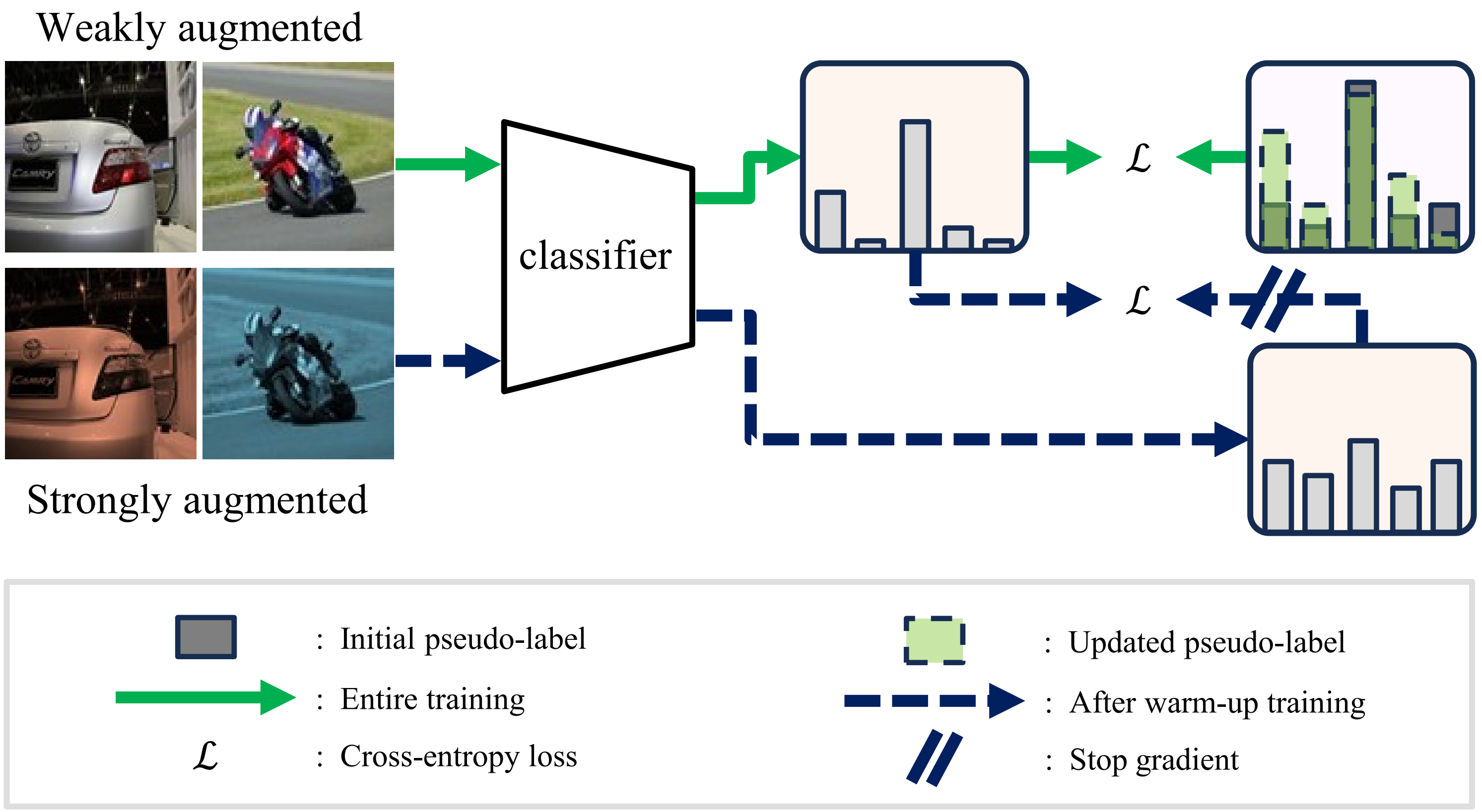
We then debiased by applying the inverse of this bias to all the probabilities we had previously obtained.

Method



This figure illustrates the overall training method: solid lines represent processes maintained throughout training, while dashed lines indicate operations conducted only after the warm-up phase.

Method



In addition to label updates and debiasing, consistency loss is utilized to robustly train the classifier against noisy signals in pseudo-labels.

Results

Supervision level	Annotation	Method	VOC12	VOC07	COCO	NUS
Fully supervised	Fully labeled	BCE	90.1	91.3	78.5	50.7
		BCE-LS [7]	91.6	92.6	79.4	51.7
Weakly supervised	Partial labeled (10%)	SARB [30]	-	85.7	72.5	-
		ASL [33]	-	82.9	69.7	-
		Chen *et al.* [5]	-	81.5	68.1	-
	Single positive labeled	LL-R [19]	89.7	90.6	72.6	47.4
		G^2NetPL [1]	89.5	89.9	72.5	48.5
Unsupervised	Annotation free	Naive AN [21]	85.5	86.5	65.1	40.8
		Szegedy *et al.* [38]	86.8	87.9	65.5	41.3
		Aodha *et al.* [26]	84.2	86.2	63.9	40.1
		Durand *et al.* [10]	81.3	83.1	63.2	39.4
		ROLE [7]	82.6	84.6	67.1	43.2
		CDUL [2]	88.6	89.0	69.2	44.0
		CCD (Ours)	**90.1**	**91.0**	**70.3**	**44.5**

Our method significantly outperforms the previous state-of-the-art (CDUL),
achieving comparable performance to fully labeled methods, particularly on VOC datasets.

Results

Label update	CLIP debias	Consistency	mAP (VOC12)
✘	✘	✘	86.4
✔	✘	✘	88.7
✔	✔	✘	89.4
✔	✘	✔	88.8
✔	✔	✔	**90.1**

Ablation of the proposed method.

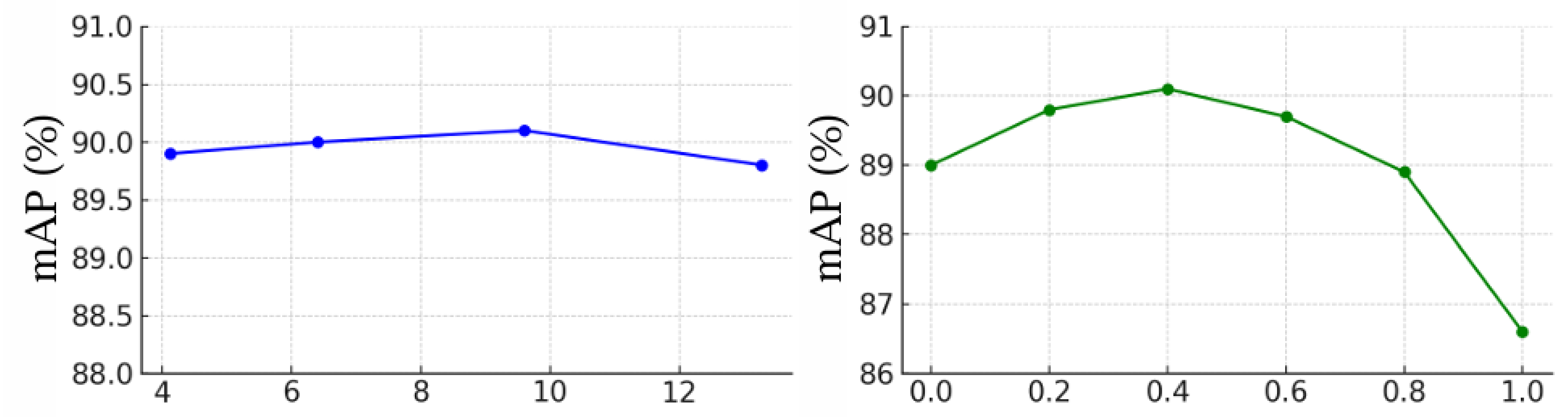
Our ablation study confirms that each proposed component contributes meaningfully to performance enhancement, with classifier-guided label updating having the greatest impact at 2.3 percentage points.

Results

Method	bottle	chair	table	person	plant	sofa	tv
w/o debias	77.0	77.1	72.7	87.3	68.8	73.0	89.5
W debias	77.6	79.0	74.8	89.2	72.1	76.6	92.0

Debiasing also markedly improved performance on classes previously hindered by inherent biases.

Results



(a) The number of mean inferences (b) Initial & Local weight (α)

We further performed ablation studies on the average inference count per image and label update ratio, identifying optimal parameters at around 10 inferences and a 40% initial label ratio.

Conclusion

- We proposed a novel CLIP based unsupervised multi-label classification method with classifier guidance.
- Proposed CCD effectively addresses CLIP's limitations: view-dependent prediction and inherent bias.
- Enhances unsupervised multi-label classification significantly in benchmark datasets (e.g., VOC07, VOC12, COCO, NUSWIDE)

We propose a novel CLIP-based unsupervised multi-label classification method employing classifier guidance.

Conclusion

- We proposed a novel CLIP based unsupervised multi-label classification method with classifier guidance.
- Proposed CCD effectively addresses CLIP's limitations: view-dependent prediction and inherent bias.
- Enhances unsupervised multi-label classification significantly in benchmark datasets (e.g., VOC07, VOC12, COCO, NUSWIDE)

The proposed method successfully tackles fundamental CLIP challenges,
including view-dependent predictions and inherent biases, achieving meaningful performance gains.

Conclusion

- We proposed a novel CLIP based unsupervised multi-label classification method with classifier guidance.
- Proposed CCD effectively addresses CLIP's limitations: view-dependent prediction and inherent bias.
- Enhances unsupervised multi-label classification significantly in benchmark datasets (e.g., VOC07, VOC12, COCO, NUSWIDE)

Our method achieves state-of-the-art performance across several benchmark datasets.

Thank you

Thank you for listening.