Classifier-guided CLIP Distillation for Unsupervised Multi-label Classification

Dongseob Kim
Samsung Electronics, Republic of Korea
kou.kim@samsung.com

Hyunjung Shim
KAIST, Republic of Korea
kateshim@kaist.ac.kr

Abstract

Multi-label classification is crucial for comprehensive image understanding, yet acquiring accurate annotations is challenging and costly. To address this, a recent study suggests exploiting unsupervised multi-label classification leveraging CLIP, a powerful vision-language model. Despite CLIP's proficiency, it suffers from view-dependent predictions and inherent bias, limiting its effectiveness. We propose a novel method that addresses these issues by leveraging multiple views near target objects, guided by Class Activation Mapping (CAM) of the classifier, and debiasing pseudo-labels derived from CLIP predictions. Our Classifier-guided CLIP Distillation (CCD) enables selecting multiple local views without extra labels and debiasing predictions to enhance classification performance. Experimental results validate our method's superiority over existing techniques across diverse datasets. The code is available at https://github.com/k0u-id/CCD.

1. Introduction

Multi-label classification plays a pivotal role in identifying multiple classes within an image, serving as a foundational component in image understanding [25, 27, 29, 36, 40]. Its applications are widespread, including recommendation systems, surveillance, and image retrieval [43]. However, acquiring multi-label annotations poses challenges due to the difficulty and significant costs involved in meticulously annotating all relevant objects within an image.

To solve this problem, various label-efficient techniques have been proposed. A partial label setting [5, 10, 14, 20, 24, 30] annotates only a portion of the classes appearing in an image and considers the rest as unknown labels. In the extreme, single positive label settings [7, 19, 35, 42, 51] provide only one positive label per image, with all other labels assumed unknown. However, these methods still incur labeling costs. Besides, current studies simulate partial and single positive settings by randomly omitting labels from a fully labeled dataset, which may not accurately reflect the actual labeling practices. For example, humans tend to prioritize labeling trivial objects first.

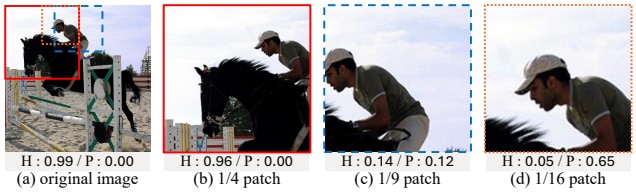


Figure 1. CLIP prediction probability of *horse* (H) and *person* (P) corresponding to different input patches from single image: (a) The original image and its CLIP prediction. (b) The $1/4$ sized patch and its CLIP prediction. (c) The $1/9$ sized patch and its CLIP prediction. (d) The $1/16$ sized patch and its CLIP prediction. The same colored box indicates each patch cropped from the corresponding box.

To overcome these limitations, we focus on an unsupervised multi-label classification problem leveraging a vision-language model. One such model is CLIP [31], trained on an extensive dataset of 400M {image, text} pairs from the web. As a result, this model excels at recognizing diverse visual attributes and object categories in a zero-shot manner. Most recently, CDUL [2] applied CLIP for unsupervised multi-label classification by using it as a pseudo-labeling function. CDUL generates pseudo-labels for both global and local views to help identify less prominent objects, training a multi-label classifier with each view-label pair. However, CDUL overlooks the limitations of CLIP, which hinders its prediction performance.

In this paper, we identified two reliability issues in CLIP, addressing them to fully harness CLIP's potential. Firstly, we demonstrate that CLIP's predictions are highly view-dependent, considerably impacting classification performance. In a toy experiment, we observe that changing the crop position and size of an image leads to notable changes in CLIP's predictions. For example, in Fig. 1, when the cropped area is slightly shifted or increased, the probabilities for the horse and person classes vary significantly. Based on these key observations, we use multiple local views near the target object to mitigate the uncertainty of CLIP's pseudo-labels caused by view changes. Specifically,

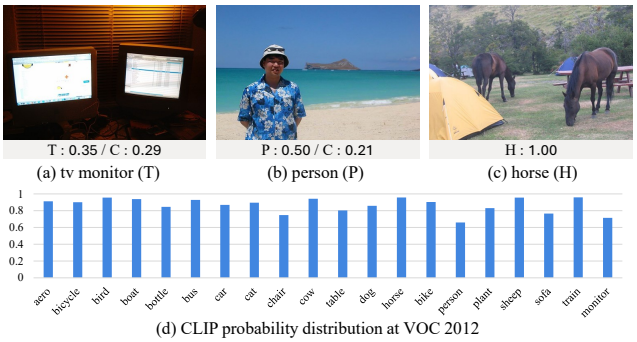


Figure 2. Sample images showcasing CLIP bias. C indicates the probability of *chair*. (a) Top 1 probability of "tv monitor" image is 35%. (b) Top 1 probability of a "person" image is 50%. (c) Top 1 probability of a "horse" image is 100%. (d) The mean class-wise probability of PASCAL VOC 2012. We can observe the class-wise prediction bias of CLIP from these results.

we acquire various local views extracted by the bounding box using the Class Activation Mapping (CAM) [49] method, where the CAM is computed from the trained classifier itself. Additionally, our method can draw more local views for hard samples whose prediction exceeds a certain threshold across multiple classes.

Another notable issue with CLIP is its inherent prediction bias, which tends to reduce the predicted probability of specific classes. Due to polysemy [31], the words representing each class can be ambiguous in text embedding space. As illustrated in Fig. 2 (a) and (b), despite the presence of a single class (*tv monitor* and *person*, respectively), CLIP assigns considerably low top-1 probabilities of around 35% and 50%. Conversely, objects like horses exhibit notably higher probabilities, as depicted in Fig. 2 (c). Analyzing the average CLIP top-1 probabilities across classes, as shown in Fig. 2 (d), further highlights this bias, indicating a lower prediction for certain classes compared to others. The biased prediction is evident across various datasets such as MS COCO and NUSWIDE[1]. Since the proposed method relies on pseudo-labels generated by CLIP, its bias naturally transfers to the classifier. Consequently, the classifier shows a similar prediction bias and performance degradation. To mitigate this issue, we adjust all pseudo-labels by applying the inverse transform of the bias. While this adjustment can amplify unwanted noise in pseudo-labels, we apply consistency loss to handle this noise after debiasing. Our adjustment helps debiasing the classifier, thereby enhancing its generalization performance. As a result, our proposed Classifier-guided CLIP Distillation (CCD) achieves state-of-the-art performances on the PASCAL VOC 2012, PASCAL VOC 2007, COCO, and NUSWIDE benchmarks.

[1]CLIP probability distribution of other datasets are in the supplementary material.

2. Related Work

2.1. Label efficient multi-label classification.

Unlike multi-class classification, where a single class is identified in an image, multi-label classification requires accurately identifying all classes present in an image. Labeling every class in an image is significantly more challenging and resource-intensive than identifying just one class. To address this issue, a partial label setting has been proposed [5, 10, 14, 20, 24, 30], which utilizes only a subset of labels associated with each image.

Early approaches assumed any unannotated labels (*i.e.* "unknown labels") as negative labels [4, 37, 45]. These methods involved correcting these labels during the learning process. However, approaches that relied on probabilistic models [9, 16, 41, 46] for label correction were dependent on optimization techniques, resulting in limited scalability. Subsequently, alternative methods have been proposed, such as leveraging label-to-label similarity [14], cross-image semantic correlation [5], or category-specific representation [30] to refine labels. Some approaches suggest using soft-pseudo-labels instead of unknown labels [10, 26, 38].

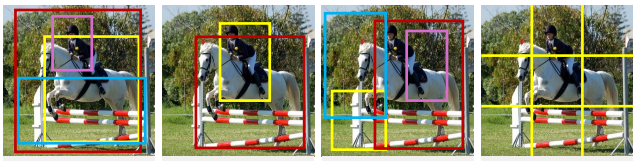
Research has extended to a more extreme scenario: single positive label setting [7, 19, 42, 51], where only a single positive label is provided per image. These methods build upon those proposed in the partial label setting and adapt them to function effectively in more challenging contexts.

More recently, CDUL [2] has addressed the unsupervised multi-label classification problem by utilizing CLIP. In a similar vein, our method tackles the unsupervised multi-label classification problem using CLIP while additionally providing local area information with CAM and addressing bias in CLIP predictions.

2.2. Vision-language pretraining.

Vision-language pretraining (VLP) has emerged as a pivotal approach for bridging the gap between visual and textual modalities [15, 28, 34]. Among the pioneering VLPs, CLIP [31] stands out as the most commonly utilized framework. CLIP was trained on large {image, text} pairs sourced from the Internet. Recently, several studies have adopted VLP to provide rich visual-semantic knowledge for downstream image tasks, such as detection [12, 48], segmentation [18, 32, 50], human-object interation [17, 22], and generation [8, 39, 44].

In multi-label classification, various methods have emerged to derive pseudo-labels from CLIP predictions [2, 13, 47]. Among these, CDUL [2] integrates local inferences with global inference, successfully expanding the utilization of the VLP to multi-label classification. However, CDUL fails to tackle two limitations inherent in CLIP: sensitivity to input view selection and inherent prediction bias.



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

Figure 3. The proof-of-concept study of local view selection method. We train the classifier with pseudo-label generated from four different local views: (a) Around GT boxes, (b) GT boxes, (c) Random boxes, (d) Uniform grid boxes. The numbers below each sample are the performance (mAP) of the classifier trained with the corresponding pseudo-label for the entire training set. The classifier trained with local view around GT boxes achieved 1.3%p higher performance compared to classifier trained with uniform-grid local view.

Consequently, the pseudo-labels generated by CDUL exhibit noise and bias. In contrast, our approach leverages a fine-tuned classifier to suggest local views and addresses CLIP bias through calibration.

3. Method

3.1. Generating initial pseudo-labels

CLIP, a powerful vision-language model, has been trained to align images with corresponding textual descriptions. Leveraging the powerful zero-shot capability of CLIP, we initially generate pseudo-labels similar to CDUL. For each image in the dataset, we extract its image embedding using the CLIP image encoder. Simultaneously, we obtain the text embedding by inputting the class name associated with each dataset and a fixed prompt (*e.g.* "a photo of the [class name]") into the text encoder. Next, we compute the similarity between the image embedding and the class-specific text embedding. This similarity score is then translated to compute softmax probabilities. These probabilities serve as the initial pseudo-labels. Mathematically, the similarity score for each class can be expressed as:

$$s_c = \frac{f^\top w_c}{\|f\| \cdot \|w_c\|}, \quad 1 \leq c \leq C \tag{1}$$

where f is a CLIP image embedding, w_c is a class-specific CLIP text embedding and c indicates a class index. With class-wise similarity, we can calculate a class-wise softmax probability p_c of each image as follows:

$$p_c = \frac{\exp(s_c/\tau)}{\sum_{c=1}^{C} \exp(s_c/\tau)}, \tag{2}$$

where τ is a CLIP temperature parameter. We concatenate each class-wise softmax probability to generate initial pseudo-labels. Finally, the initial pseudo-label for each image is defined as:

$$l^{initial} = \left\{ p_1^{global}, p_2^{global}, \cdots, p_c^{global} \right\}. \tag{3}$$

Following the acquisition of softmax probabilities, the initial pseudo-labels are used for the warm-up training of the classifier. Since the softmax calculation inherently normalizes the probabilities to ensure a sum of 1, it does not fit the multi-label classification problem where multiple classes have the same high probability (*e.g.* 1). Nevertheless, after the classifier undergoes warm-up training, it can predict each class independently, leading to recognizing multiple classes present in an image, albeit with limited accuracy. To address this performance limitation, in the subsequent section, we introduce a method to construct {training data, pseudo-label} pairs tailored for the multi-label classification problem, leveraging a warm-up trained classifier.

3.2. Updating classifier-guided pseudo-label

Prior research [2] has shown that CLIP's inference results on the entire images may be suboptimal for multi-label classification. CLIP tends to assign high probabilities primarily to prominent objects in the input image. As a result, it does not effectively distribute probabilities across all existing objects. CDUL addresses this limitation by dividing the image into uniform patches in a grid fashion. Despite its simplicity, employing a uniform patch-based approach for local view acquisition may lead to the inference of background batch, where the patch does not contain any class object. More importantly, treating both simple images (*i.e.* those containing a single object) and complex images (*i.e.* those containing multiple objects) equally with the same number of local views is clearly inefficient in optimizing the resource (*i.e.* CLIP inference).

Proof-of-concept study. We perform a proof-of-concept study to evaluate the impact of using object-specific locations in local view proposals. Our hypothesis is that incorporating class-related object locations enhances performance. We extract local views in four different scenarios, using ground truth bounding boxes (*i.e.*, the bounding box that tightly covers the target object): (1) random cropping around the bounding box, (2) using only the bounding box itself[2], (3) random cropping without guidance from the bounding box, and (4) uniform grid-based cropping. Each scenario produces nine patches per image. Subsequently, we infer pseudo-labels for these patches and train with each {patch, pseudo-label} pair. As shown in Fig. 3, our results indicate that the first scenario–(a) random cropping around the bounding box–achieves the best performance.

Our proof-of-concept study reveals two critical insights: (1) the local view generation policy significantly influences

[2]For images with fewer than nine boxes, we additionally infer their resized versions.

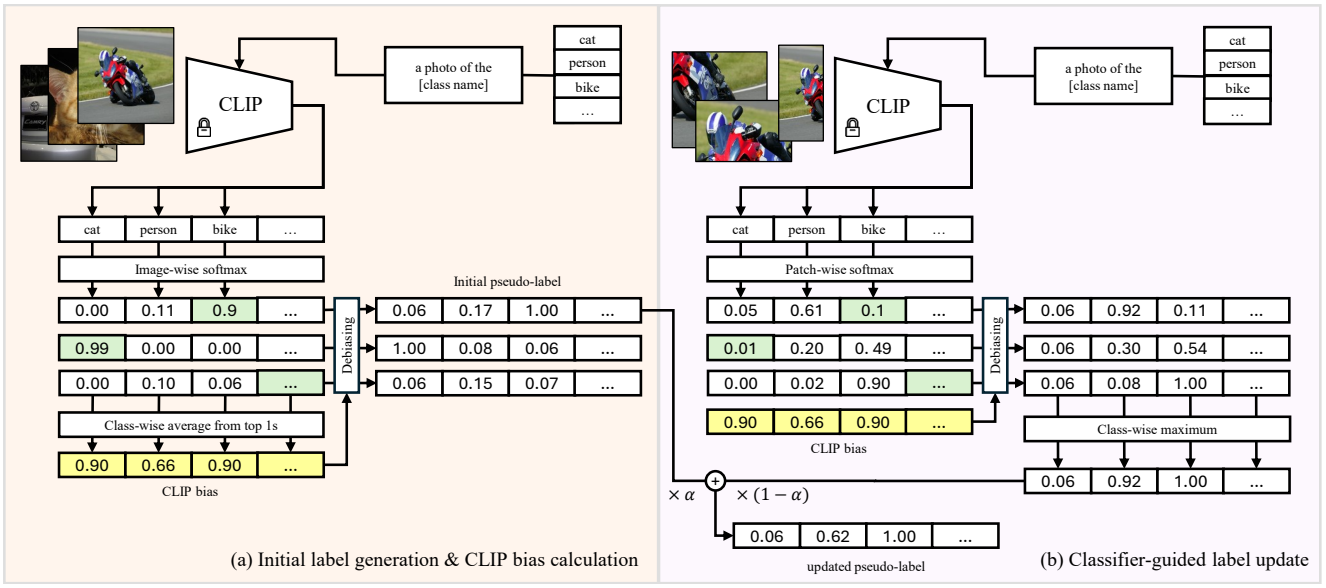


Figure 4. The overview of label preparation. (a) We calculate the cosine similarity between text embeddings and image embeddings. The softmax probability of these similarities is the CLIP prediction of each image. The top 1 probabilities are highlighted in *green boxes*. By class-wise averaging the top-1 probabilities, the CLIP bias is derived (*yellow box*). Pseudo-labels are then generated by debiasing CLIP predictions. (b) For label updating, classes above a threshold are selected from the classifier output, and local views corresponding to these classes are extracted. The process for acquiring local labels for each patch mirrors the initial label acquisition. The final pseudo-label is obtained through a weighted sum of initial pseudo-labels and local labels.

CLIP's predictions, and (2) framing the view around the target object greatly improves classification accuracy.

CAM-based view-selection. Motivated by the above two findings, we decided to approximate the first scenario in our approach. We apply Class Activation Mapping (CAM) with the trained classifier to identify vital regions within the image. This is called a CAM-based view selection method. This method generates {training image, pseudo-label} pairs by focusing on the object periphery, without relying on any ground-truth annotations.

By employing the trained classifier, CAM visualizes activations that most significantly influence class prediction. CAM calculation is specifically based on the weights assigned to each channel of the feature map, which influence the classification score. This can be represented by the following formula:

$$G_c(x) = \sum_{i=1}^{Q} \mathbf{w}_i^{c\top} \cdot g_i(x), \tag{4}$$

where x is input image, g is CNN network, $g(x)$ is the feature map of the CNN, $\mathbf{w}^c$ is class-wise weight of classifier, and Q indicates the number of channels in the feature map.

We first normalize the activation map to a range between 0 and 1, then apply a threshold to identify highly activated regions. Next, we generate candidate bounding boxes with slight perturbation offsets. These bounding boxes define local views of the image. The regions highlighted by CAM act as reasonable approximations of the object's location, allowing us to extract local views without additional labeling costs. These localized view images are fed into CLIP to obtain softmax probabilities for the patches. The probabilities of local views are computed using the same method as for the global view, and then generating local labels that are focused on each local region.

The number of local views. Additionally, we adjust the number of local views based on image complexity. For simpler images, the classifier typically provides confident, high-probability predictions. This translates to high logits for well-trained classes and near-zero for others. Conversely, complex images often produce uncertain predictions, characterized by higher logits distributed broadly across multiple classes. To account for this, we conduct additional local view sampling when the classifier output of a certain class surpasses a predefined threshold. Specifically, we crop the bounding box of any class exceeding this threshold value and obtain its local probability. Through our experiments, we empirically determined a fixed threshold that ensures appropriate average number of classes exceeding this threshold. Details on this ablation study can be found in the experiments section. This approach allocates fewer local views to simple images and more local views to

complex ones, optimizing the local view assignment.

To obtain a single local pseudo-label for each input image, we select the maximum value of class-wise probability across all local patches from the corresponding image. Denoting the probability of $i-$th patch for class c as $p_{i,c}^{patch}$, we conduct patch-wise aggregation using the maximum value as follows:

$$p_c^{local} = \max_{i=1,\cdots,N} p_{i,c}^{patch}, \quad 1 \leq c \leq C \tag{5}$$

where N represents the number of local patches corresponding to each image, dependent on the input image itself. We then aggregate the local class-wise probability for generating a local pseudo-label as follows:

$$l^{local} = \left\{p_1^{local}, p_2^{local}, \cdots, p_c^{local}\right\}. \tag{6}$$

We update the pseudo-label by the weighted sum of initial and local pseudo-labels. The final pseudo-label is calculated as follows:

$$l^{final} = \alpha l^{initial} + (1-\alpha) l^{local}, \tag{7}$$

where α is a hyperparameter and we used 0.4. Through this label update, the pseudo-label efficiently incorporates information about less salient objects in multi-label classification. The updated pseudo-label is utilized throughout the remainder of the training process.

3.3. Debiasing CLIP pseudo-label

Earlier, we aimed to identify highly activated areas within images through Class Activation Mapping and to infer additional label information focused on these areas. However, a challenge arises when using pseudo-labels generated by CLIP due to its inherent prediction bias. Consequently, classifiers trained using these pseudo-labels may inadvertently inherit this bias. Previous work [3] reveals that while classifiers are robust at overcoming random noise, they cannot rectify consistent biases associated with specific classes. Therefore, we need to overcome the inherent prediction bias of CLIP to enhance the performance.

To understand this bias, given CLIP's focus on salient objects, we treat global CLIP prediction as a single-label classification task. In a single-label classification, the predicted probability distribution for each class aligns with the class-wise average of the highest probabilities from each prediction. Following this principle, we compute the average prediction probability for each class, defining it as CLIP's prediction bias. Specifically, we compute CLIP bias as follows: (1) Obtain initial pseudo-labels by running each image in the "target training set" through CLIP; (2) Record top-1 probabilities for samples below the entropy threshold to constrain the task to single-label classification; (3) Count each class occurrences only if it achieves the top 1 probability; (4) Normalize each class's softmax probability by its occurrence count, defining this normalized probability as CLIP bias. This process does not require any ground-truth annotation, solely relying on the result of CLIP prediction.

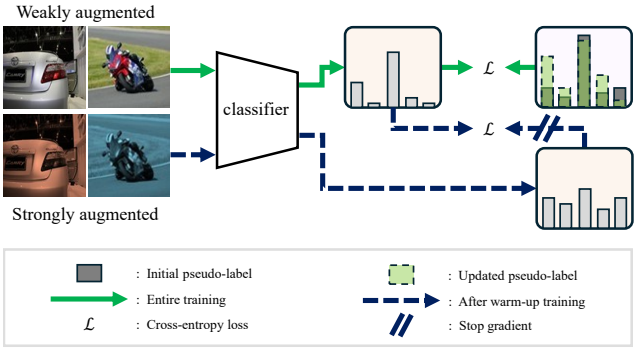


Figure 5. The training process of our method. We train the classifier using cross-entropy targeting initial pseudo-label during the warm-up phase which is illustrated as the *green line*. After the classifier-guided label update, we train the classifier using the cross-entropy loss targeting updated pseudo-label and the cross-entropy loss between the logits of differently augmented inputs which are illustrated as the *green line* and the *blue dashed line*.

The bias obtained in this manner serves as a basis for calibrating CLIP-derived probabilities. Specifically, we calibrate by multiplying each class's softmax probability from CLIP by the inverse of the CLIP bias. This calibration process enables the classifier to learn a rectified probability distribution. Our debiasing technique leads to a noticeable improvement in the performance of classes previously challenged due to the inherent CLIP bias (see Tab. 3). Furthermore, we adjust excessively small probabilities to prevent the minimum probability value from approaching zero, aiding in learning the corrected distribution. In doing so, we guide the classifier as a whole to learn from this inverse distribution. This calibration is applied not only to the global probability but also to local inference, minimizing the influence of the CLIP bias on the classifier as much as possible.

3.4. Overall training

Consistency loss. The proposed method tackles noisy data by utilizing pseudo-labels as target labels for all {training data, pseudo-label} pairs. Since pseudo-labels inherently carry noise and our debiasing process can potentially amplify this noise, we employ consistency loss between the data and its augmented data as additional critics. This approach is inspired by the common practice in a semi-supervised regime, improving feature extractors for handling noisy data. In our framework, we introduce consistency loss alongside the conventional binary cross-entropy loss during training. Consistency loss involves applying different augmentation intensities to the same image. Also, it passes each augmented version through the classifier to en-

sure consistent predictions. In Fig. 5, we employ weak augmentations, such as resizing and flipping, and strong augmentations, which comprise weak augmentation and additional color jittering.

Training process. Initially, the entire training dataset undergoes inference through CLIP, followed by calculating the CLIP bias. Subsequently, we derive a global label by debiasing the global probability, utilizing it as an initial pseudo-label in Fig. 4 (a). The classifier is then warmed up with this pseudo-label as the target (the green line in Fig. 5). Following warm-up training, we update the pseudo-label with classifier-guided local inference. During this phase, the obtained local probabilities undergoes debiasing. Then they are combined to derive a local label, which then updates the existing pseudo-label as depicted in Fig. 4 (b). The classifier is subsequently trained with the updated pseudo-label for the remaining epochs (the green line and the blue dashed line in Fig. 5). For whole trainig, our loss function combines two terms:

$$\mathcal{L}_{total} = L_{ce} + \beta L_{consist}, \tag{8}$$

where $\mathcal{L}_{ce}$ is cross-entropy loss and $\mathcal{L}_{consist}$ is consistency loss. β is a hyperparameter and we fix it as 0 during warm-up phase and 1 after.

4. Experiments

4.1. Experimental setup

Dataset & evaluation Metric. We conducted evaluations on four distinct datasets. The PASCAL VOC 2012 [11] dataset comprises 20 classes, with 5,717 images in the training set and 5,823 in the validation set. Similarly, the PASCAL VOC 2007 dataset also encompasses 20 classes, with a combined total of 5,011 training and validation images and 4,952 test images. MS COCO [23] consists of 80 classes, encompassing 82,081 training images and 40,137 validation images. Additionally, the NUSWIDE [6] dataset comprises 81 classes, with a training set of 150,000 images and a test set of 60,200 images. In cases where a separate test set was not available, the validation set was utilized for testing purposes. Mean Average Precision (mAP) was employed as the performance metric, with ground truth labels utilized solely for performance evaluation and not incorporated during the model training phase.

Implementation detail. For a fair comparison, we endeavored to set the hyperparameters for CLIP inference as closely as possible to those of the primary competitive method, CDUL. We acquired labels using the frozen CLIP ResNet50×64, and employed the ImageNet pre-trained ResNet101 as the backbone network for our classifier. The entropy threshold for calculating CLIP bias and the classifier threshold for local view inference is both 0.5. It Our warm-up classifier is trained for two epochs. While the entire training process can proceed up to 10 epochs, we obtain our final model via early stopping to prevent overfitting. Specifically, we terminate training based on the gradient of mean Average Precision (mAP) of the *training labels* throughout the learning process. Training was ceased when the gradient of mAP reached the first local minimum. We utilized a learning rate of 10^{-5} across all datasets and employed a batch size of 16. Undisclosed hyperparameters of CDUL were manually reproduced. For generating the initial pseudo-labels of global views, we maintained a fixed image resolution of 640×640. The CLIP prompts consistently used the phrase "a photo of the [class name]" except for the NUSWIDE dataset. (For a fair comparison with the existing method, CDUL, we employ the phrase "a photo of a [class name]" to match the pseudo-label performance.) For acquiring the initial pseudo-labels of local views, we introduced an offset value of 80 pixels for each vertexes when generating bounding boxes (*i.e*. a thresholded CAM-bounding box ± 80 as our local views). Additionally, cropped images were resized to have a longer side of 640 pixels while preserving their aspect ratios to obtain CLIP probabilities.

4.2. Comparison with state-of-the-art methods

Competitive methods. Multi-label classification can be partitioned depending on the type of supervision: fully supervised, weakly supervised, and unsupervised approaches. Firstly, in a fully supervised setting, we choose (1) the Naíve method using only binary cross-entropy (BCE) loss and (2) the Naíve method with additional label smoothing (LS) [7]. Secondly, we compared our method with recent weakly supervised methods, including partial labeled methods utilizing only 10% of total labels such as (3) SARB [30], (4) ASL [33], (5) Chen et al. [5]. For single positive label methods, we choose (6) LL-R [19], (7) G^2NetPL [1]. Finally, in an unsupervised scenario, we obtained initial labels using CLIP and evaluated the performance of (8) Naíve AN [21], (9) Szegedy et al. [38], (10) Aodha et al. [26], (11) Durand et al. [10], (12) ROLE [7], which leveraging CLIP generated initial labels. We conducted a comparative analysis with (13) CDUL [2], the current state-of-the-art in an unsupervised multi-label classification task.

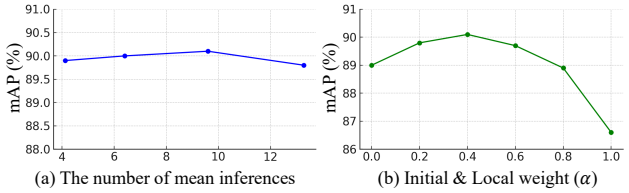
Comparison with fully & weakly methods. Compared to fully supervised methods, our Classifier Guided CLIP Distillation (CCD) demonstrates comparable performance on the PASCAL VOC 2012 and PASCAL VOC 2007 datasets. In particular, achieving similar performance without any human annotation is impressive. CCD outperforms various weakly supervised methods on the PASCAL VOC 2012 and PASCAL VOC 2007 datasets, although our performances on MS COCO and NUSWIDE are relatively lower. This

Table 1. Mean average precision mAP in (%) for different multi-label classification methods under different supervision levels. Bold represents the best results among unsupervised methods. The proposed achieved state-of-the-art performance.

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**

Table 2. Ablation study of the proposed moduels. Mean average precision mAP in (%) for different configuration uon the PASCAL VOC 2012 dataset. The best score is in bold.

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



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

Figure 6. Classification results (mAP) on PASCAL VOC 2012 depending on (a) the change of the number of mean inferences (b) the change of ratio of pseudo-label updating.

suggests that CCD is more effective on datasets with a smaller number of classes, such as PASCAL VOC 2012 and 2007, which consist of 20 classes. On the other hand, the performance gain is relatively small for complex datasets with many classes. It can be attributed to the simplistic prompt, "a photo of the [class name]," which might not accurately represent the target classes and co-occurrence issue, which bring out noisy local labels. 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. CCD consistently shows about a 3 4%p performance gain over all datasets.

Comparison with unsupervised methods. Compared to unsupervised methods, CCD exhibits superior performance across all datasets. Particularly noteworthy are the significant improvements of 1.5%p, 2%p, and 1.1%p on PASCAL VOC 2012, PASCAL VOC 2007, and MS COCO, respectively, compared to the existing state-of-the-art methods. These results highlight the effectiveness of our method. However, in the case of the NUSWIDE dataset, the performance improvement is not as substantial. This can be attributed to the unique characteristics of the NUSWIDE dataset, where 20% of the train/test images do not contain any class-related objects (*i.e.* "none images"). As a result, the initial pseudo-label acquisition process fails to form a high probability for specific classes in these images. Consequently, during local inference, a significant portion of inferences is assigned to "none images," thereby reducing the quality of pseudo-labels of CCD.

4.3. Ablation

Effects of each module. We evaluate the effectiveness of each component of our method in Tab. 2. The performance of the baseline model, the classifier trained with CLIP-generated global pseudo-labels, is 86.4%. When we update the pseudo-label with classifier-guided local inference, it shows a significant improvement of 2.3%p. This shows that our CAM-based local view selection method and adaptively allocating the number of CLIP inferences are effective. It already exceeds the performance of previous state-of-the-art methods. Debiasing CLIP also contributed 0.7%p, indicating that our proposed debiasing method effectively relieves the CLIP bias. The consistency loss alone adds only 0.1%p performance gain. However, when CLIP debiasing and consistency loss are combined, a substantial improvement of

Table 3. Per-class AP in (%) for without and with debiasing. We focus on the seven classes from the lowest mean probability. The result showcases that our debiasing is effective for enhancing the performance of biased classes.

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

1.4%p is achieved. This result implies that pseudo-labels generated by compensating for CLIP bias may amplify both accurate information and noise. Since the consistency loss is effective for dealing with noisy data, it creates a synergy with CLIP debiasing.

The number of inference. We examine how model performance varies with different numbers of CLIP local inferences. We incrementally increase the average number of inferences to approximately 4, 6, 10, and 14. In Fig. 6 (a), the model achieves its best performance when the average number of inferences is around 10. Notably, increasing the number of inferences does not consistently enhance performance. Fewer inferences limit the allocation of local views to challenging samples, which hinders the ability to extract sufficient information from complex images, thereby negatively impacting negatively. In contrast, an excessive number of inferences results in the overallocation of local views, even for easy samples. It potentially introduces redundant inference for easy samples, making noisy pseudo-labels. In particular, when the average number of inferences is set to 10 or 14, the maximum number of local inferences remains capped at 20 (*i.e*. the total number of classes). It suggests that the inferences allocated to complex images have already reached saturation.

Initial & local weight. We investigate the impact of adjusting the ratio α between initial and local labels when updating the pseudo-labels (Eq. 7). In Fig. 6 (b), the results indicate that utilizing both global and local labels together leads to performance enhancement. This trend arises because two different label sources have complementary properties. The local label contains more information through multiple CLIP inferences of different local patches. However, since the reliability of CAM is not 100%, it may draw an invalid patch. Conversely, the initial label contains less information, but it is always generated from valid images. By blending these two sources, we reinforce consistent predictions between local and initial labels, thereby strengthening the signal of reliable labels. Additionally, we highlight that the use of our local labels alone (*i.e*. at weight 0) can outperforms CDUL. This confirms that our choice of local views and their labels extracts more informative learning signals than the mixture of the global view and uniformly sampled local views.

Per-class mAP of biased classes. We assess the effectiveness of our proposed method by examining the improvement in performance for specific target classes before and after CLIP debiasing. In Tab. 3, we focus on seven classes from the biased class, including *bottle*, *chair*, *table*, *person*, *potted plant*, *sofa*, and *tv monitor* as confirmed in the Fig. 2. Our results demonstrate consistent improvements in the performance of all these classes, ranging from 0.6%p (*bottle*) to 3.6%p (*sofa*). Moreover, we observe comparable or slightly enhanced performance for the remaining classes not shown in the table. The result affirms the effectiveness of our method across a broader spectrum. Detailed class-wise performance results for the entire dataset are provided in the supplementary material.

4.4. 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. This challenge is especially evident in complex datasets like COCO, where similar classes–such as *skies*, *snowboards*, and *skateboards*–frequently appear in scene with human activity. These classes often co-occur with human feet, leading classifiers to associate high activations near the feet with these classes. This phenomenon makes them potential candidates for local views. This redundant information results in noisy pseudo-labels, reducing overall performance. Examples of such failure cases are provided in the supplementary materials. 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.

5. Conclusion

In this paper, we propose a novel unsupervised multi-label classification method, Classifier-guided CLIP Distillation (CCD). We address two critical issues in CLIP: inconsistent predictions due to subtle input changes and inherent prediction bias. To tackle these issues, we devised a classifier-guided label update and CLIP debiasing. Firstly, we identified the effectiveness of acquiring local views near class-related objects and proposed a classifier-guided label update. It selects local views utilizing the Class Activation Mapping (CAM) and acquires corresponding labels from CLIP. Our approach results in allocating more local views for complex images. Secondly, we address the CLIP's bias. We investigated the biased prediction distribution of CLIP and rectified the pseudo-labels. This step notably boosts performance in certain classes and reduces the impact of bias on CLIP. Through experiments on four benchmarks, CCD newly achieved state-of-the-art performances in unsupervised multi-label classification. Remarkably, CCD achieved comparable performance to fully supervised methods on PASCAL VOC 2012 and PASCAL VOC 2007 datasets without relying on annotations.

Acknowledgements. This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (RS-2025-00520207, RS-2023-00219019), KEIT grant funded by the Korean government (MOTIE) (No. 2022-0-00680, No. 2022-0-01045), the IITP grant funded by the Korean government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project, No. 2021-0-02068 Artificial Intelligence Innovation Hub, RS-2019-II190075 Artificial Intelligence Graduate School Program (KAIST)), Samsung Electronics Co., Ltd (IO230508-06190-01) and SAMSUNG Research, Samsung Electronics Co., Ltd.

References

[1] Rabab Abdelfattah, Xin Zhang, Mostafa M Fouda, Xiaofeng Wang, and Song Wang. G2netpl: Generic game-theoretic network for partial-label image classification. *arXiv preprint arXiv:2210.11469*, 2022. 6, 7

[2] Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. Cdul: Clip-driven unsupervised learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1348–1357, 2023. 1, 2, 3, 6, 7

[3] Görkem Algan and İlkay Ulusoy. Label noise types and their effects on deep learning, 2020. 5

[4] Serhat Selcuk Bucak, Rong Jin, and Anil K. Jain. Multi-label learning with incomplete class assignments. In *CVPR 2011*, pages 2801–2808, 2011. 2

[5] Tianshui Chen, Tao Pu, Hefeng Wu, Yuan Xie, and Liang Lin. Structured semantic transfer for multi-label recognition with partial labels. In *Proceedings of the AAAI conference on artificial intelligence*, pages 339–346, 2022. 1, 2, 6, 7

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 6

[7] Elijah Cole, Oisin Mac Aodha, Titouan Lorieul, Pietro Perona, Dan Morris, and Nebojsa Jojic. Multi-label learning from single positive labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 933–942, 2021. 1, 2, 6, 7

[8] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *European Conference on Computer Vision*, pages 88–105. Springer, 2022. 2

[9] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S Bernstein, Alex Berg, and Li Fei-Fei. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3099–3102, 2014. 2

[10] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 647–657, 2019. 1, 2, 6, 7

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, 2010. 6

[12] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[13] Ping Hu, Ximeng Sun, Stan Sclaroff, and Kate Saenko. Dualcoop++: Fast and effective adaptation to multi-label recognition with limited annotations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[14] Dat Huynh and Ehsan Elhamifar. Interactive multi-label cnn learning with partial labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9423–9432, 2020. 1, 2

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[16] Ashish Kapoor, Raajay Viswanathan, and Prateek Jain. Multilabel classification using bayesian compressed sensing. *Advances in neural information processing systems*, 25, 2012. 2

[17] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14829–14838, 2022. 2

[18] Dongseob Kim, Seungho Lee, Junsuk Choe, and Hyunjung Shim. Weakly supervised semantic segmentation for driving scenes. *arXiv preprint arXiv:2312.13646*, 2023. 2

[19] Youngwook Kim, Jae Myung Kim, Zeynep Akata, and Jungwoo Lee. Large loss matters in weakly supervised multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14156–14165, 2022. 1, 2, 6, 7

[20] Youngwook Kim, Jae Myung Kim, Jieun Jeong, Cordelia Schmid, Zeynep Akata, and Jungwoo Lee. Bridging the gap between model explanations in partially annotated multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3408–3417, 2023. 1, 2

[21] Kaustav Kundu and Joseph Tighe. Exploiting weakly supervised visual patterns to learn from partial annotations. *Advances in Neural Information Processing Systems*, 33:561–572, 2020. 6, 7

[22] Yue Liao, Aixi Zhang, Miao Lu, Yongliang Wang, Xiaobo Li, and Si Liu. Gen-vlkt: Simplify association and enhance interaction understanding for hoi detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20123–20132, 2022. 2

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[24] Bing-Qing Liu, Bin-Bin Jia, and Min-Ling Zhang. Towards enabling binary decomposition for partial multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13203–13217, 2023. 1, 2

[25] Weiwei Liu and Ivor Tsang. On the optimality of classifier chain for multi-label classification. *Advances in Neural Information Processing Systems*, 28, 2015. 1

[26] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9596–9606, 2019. 2, 6, 7

[27] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018. 1

[28] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[29] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2939, 2016. 1

[30] Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. Semantic-aware representation blending for multi-label image recognition with partial labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2091–2098, 2022. 1, 2, 6, 7

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[32] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 2

[33] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021. 6, 7

[34] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022. 2

[35] GyeongRyeol Song, Noo-ri Kim, Jin-Seop Lee, and Jee-Hyong Lee. Ignore: Information gap-based false negative loss rejection for single positive multi-label learning. In *European Conference on Computer Vision*, pages 472–488. Springer, 2025. 1

[36] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017. 1

[37] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):593–598, 2010. 2

[38] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2, 6, 7

[39] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *European Conference on Computer Vision*, pages 358–374. Springer, 2022. 2

[40] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007. 1

[41] Deepak Vasisht, Andreas Damianou, Manik Varma, and Ashish Kapoor. Active learning for sparse bayesian multilabel classification. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 472–481, 2014. 2

[42] Thomas Verelst, Paul K. Rubenstein, Marcin Eichner, Tinne Tuytelaars, and Maxim Berman. Spatial consistency loss for training multi-label classifiers from single-label annotations. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3879–3889, 2023. 1, 2

[43] Changhu Wang, Shuicheng Yan, Lei Zhang, and Hong-Jiang Zhang. Multi-label sparse coding for automatic image annotation. In *2009 IEEE conference on computer vision and pattern recognition*, pages 1643–1650. IEEE, 2009. 1

[44] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022. 2

[45] Qifan Wang, Bin Shen, Shumiao Wang, Liang Li, and Luo Si. Binary codes embedding for fast image tagging with incomplete labels. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 425–439. Springer, 2014. 2

[46] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. Ml-mg: Multi-label learning with missing labels using a mixed

graph. In *Proceedings of the IEEE international conference on computer vision*, pages 4157–4165, 2015. 2

[47] Xin Xing, Zhexiao Xiong, Abby Stylianou, Srikumar Sastry, Liyu Gong, and Nathan Jacobs. Vision-language pseudo-labels for single-positive multi-label learning. *arXiv preprint arXiv:2310.15985*, 2023. 2

[48] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[49] 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

[50] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 2

[51] Donghao Zhou, Pengfei Chen, Qiong Wang, Guangyong Chen, and Pheng-Ann Heng. Acknowledging the unknown for multi-label learning with single positive labels. In *European Conference on Computer Vision*, pages 423–440. Springer, 2022. 1, 2