CDUL: CLIP-Driven Unsupervised Learning for Multi-Label Image Classification

Rabab Abdelfattah[1], Qing Guo[2], Xiaoguang Li[3], Xiaofeng Wang[3], and Song Wang[3]

[1]University of Southern Mississippi, USA
rabab.abdelfattah@usm.edu
[2]IHPC and CFAR, Agency for Science, Technology and Research, Singapore
tsingqguo@ieee.org
[3]University of South Carolina, USA
xl22@email.sc.edu, {wangxi, songwang}@cec.sc.edu

Abstract

This paper presents a CLIP-based unsupervised learning method for annotation-free multi-label image classification, including three stages: initialization, training, and inference. At the initialization stage, we take full advantage of the powerful CLIP model and propose a novel approach to extend CLIP for multi-label predictions based on global-local image-text similarity aggregation. To be more specific, we split each image into snippets and leverage CLIP to generate the similarity vector for the whole image (global) as well as each snippet (local). Then a similarity aggregator is introduced to leverage the global and local similarity vectors. Using the aggregated similarity scores as the initial pseudo labels at the training stage, we propose an optimization framework to train the parameters of the classification network and refine pseudo labels for unobserved labels. During inference, only the classification network is used to predict the labels of the input image. Extensive experiments show that our method outperforms state-of-the-art unsupervised methods on MS-COCO, PASCAL VOC 2007, PASCAL VOC 2012, and NUS datasets and even achieves comparable results to weakly supervised classification methods.

1. Introduction

A multi-label classification task aims to predict all the objects within the input image, which is advantageous for various applications, including content-based image retrieval and recommendation systems, surveillance systems, and assistive robots, to name a few [9, 8, 6]. However, getting clean and complete multi-label annotations is very challenging and not scalable, especially for large-scale datasets, because an image usually contains multiple labels (Figure 1.a).

To alleviate the annotation burden, weakly supervised learning approaches have been studied [15, 20, 12, 1], in which only a limited number of objects are labeled on a subset of training images (Figure 1.b). Though less than the fully-labeled case, it still requires intensive manpower and time for annotations.

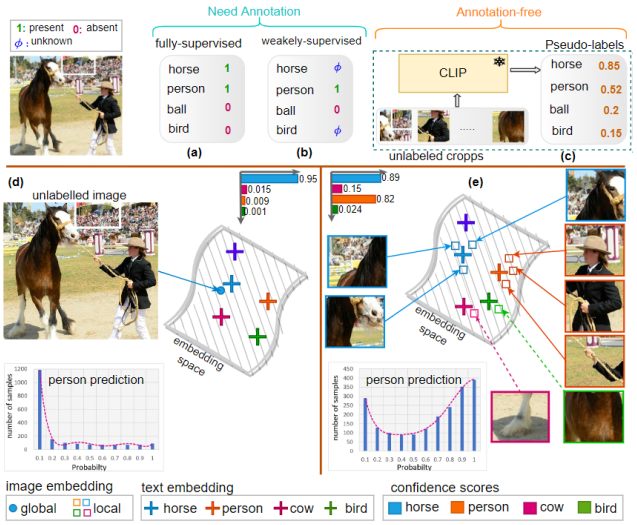

Figure 1. A comparison of our solution with fully and weakly-supervised multi-label classification. (a) The training dataset images for fully-supervised learning are fully labeled. (b) The training images used in weakly-supervised are partially labeled. (c) Our unsupervised multi-label classification method is annotation-free. (d) CLIP focuses on one class in the whole image, and the embedding is denoted by blue circle. Some classes are ignored such as "person". (e) In our approach, image snippets are mapped separately to the embedded space, where each snippet's embedding is denoted by squares. Local alignment allows to predict more labels.

To go one step further, we consider unsupervised multi-label image classification, leveraging the off-the-shelf vision-language models such as contrastive language-image pre-training (CLIP) [31]. CLIP is trained by matching each input

image to the most relevant text description over 400 million image-text pairs collected from the Internet. It has demonstrated remarkable zero-shot classification performance as a pre-trained model in image-text retrieval [31], video-text retrieval [27], and single-label image classification [31]. With CLIP, the encoded visual representations can be directly used for vocabulary categorization without additional training. However, CLIP is not suitable for multi-label classification, since it is trained only for recognizing a single object per image (Figure 1.d). Finding only one global embedding for the whole image may push CLIP to generate a high confidence score for the closest semantic text class, while neglecting other classes. In Figure 1.d, for instance, CLIP predicts class *"horse"* with a very high confidence score (0.98), but gives a very low weight to class *"person"*, given the fact that CLIP suffers from excessive polysemy [31].

To address these issues and make full use of CLIP in multi-label classification, this paper presents a CLIP-driven unsupervised learning method (CDUL) for multi-label image classification, which includes three stages: initialization, training, and inference. At the initialization stage, we use CLIP to generate global representation of the whole image and, more importantly, local representations of snippets of the image. A novel aggregation of global and local representations provides high confidence scores for objects on the image. As shown in Figure 1.e, the class "person" receives high confidence score in this case. At the training stage, the confidence scores will be used as the initial values of pseudo labels, with which a self-training procedure is proposed to optimize the parameters of the classification network as well as the pseudo labels. Finally, during inference, only the classification network is used to predict the labels of an image.

The contributions of this paper are listed as follows:

- We propose a novel method for unsupervised multi-label classification training. To the best of our knowledge, this is the first work that applies CLIP for unsupervised multi-label image classification. The aggregation of global and local alignments generated by CLIP can effectively reflect the multi-label nature of an image, which breaks the impression that CLIP can only be used in single-label classification.
- A gradient-alignment training method is presented, which recursively updates the network parameters and the pseudo labels. By this algorithm, the classifier can be trained to minimize the loss function.
- Extensive experiments show that our method not only outperforms the state-of-the-art unsupervised learning methods, but also achieves comparable performance to weakly supervised learning approaches on four different multi-label datasets.

2. Related Work

Weakly Supervised Multi-Label Classification. Due to high annotation costs, weakly supervised learning in multi-label classification becomes an interesting topic of research. Weakly supervised models are trained on a partial-label setting where some labels are annotated (called "observed labels"), and the rest are not annotated (called "unobserved or unknown labels"). Early work includes assuming the unobserved labels as negative [4, 34, 39],predicting the unobserved labels using label correlation modeling [14, 41, 43], and probabilistic modeling [36, 23]. However, these approaches rely on traditional optimization and cannot be scaled to train deep neural networks (DNNs). Recently, research effort has been made to train DNNs using partial labels [15, 20, 12, 30, 7, 33, 24]. In general, these approaches can be divided into two groups. The first group uses observed labels as the ground truth to build the label-to-label similarity graph [20], cross-images semantic correlation [7], encodes positive and negative contexts with class names [33], and blend category-specific representation across different images [30]. The second group starts with a subset of observed labels and soft pseudo labels for unobserved labels, and update pseudo labels during training, such as [35, 28, 12, 15]. Different from all these models, our method works without the need of annotations.

Vision-Language Pre-Training. Vision-language pre-training models achieve impressive performance on various tasks. Several techniques for learning visual representations from text representations have been presented using semantic supervision [29, 31, 40]. Among these models, the most effective one is CLIP [31], which exploits the large-scale image-text pairs collected from the Internet to achieve alignment of images and text representations in the embedding space. CLIP leverages contrastive learning, high-capacity language models, and visual feature encoders to efficiently capture interesting visual concepts. It shows remarkable performance in different tasks such as zero-shot inference and transfers learning in single image classification [22, 31]. However, CLIP is trained to focus on global representation, since the input image and text description both contain global semantic information. As a result, it only predicts the closest semantic text class, while neglecting other classes. Our method takes a different route by proposing a model to learn both global and local visual representations to enrich semantic concepts in multi-label classification. There are approaches using the CLIP model at the pre-training stage to help the models first develop a general understanding of the relationship between visual and textual concepts, such as RegionCLIP for object detection [46]. However, to fine-tune these pre-trained models, a large amount of labeled data is still needed, which does not belong to the category of weakly or unsupervised learning.

Unsupervised Feature Learning. Some methods for unsu-

Figure 2. Confidence scores from the off-the-shelf CLIP on sample images from COCO dataset

pervised multi-label classification in person re-identification [38, 45] focus on identity features, which are not relevant to our topic. Self-supervised learning approaches [5, 17, 19, 42, 47] use contrastive loss for instance-discriminative representations, but require ground-truth labels for fine-tuning, which does not suit our unsupervised multi-label classification problem [44]. Pseudo-label-based weakly supervised algorithms [35, 28, 12, 15] can be easily adapted for unsupervised multi-label classification by assigning pseudo labels to all objects without using observed labels. We will compare our solution to these methods in experiments.

3. Methodology

Notations. Let $\mathcal{X} = \{x_1, x_2, ...x_M\}$ denote the training set, where M is the number of images in $\mathcal{X}$ and x_m for $m = 1 \cdots, M$ is the mth image. In our formulation, $\mathcal{X}$ is totally unlabeled. Let C be the total number of classes in the dataset. Let $y_{u,m} \in \mathbb{R}^C$ denote the pseudo label vector of image x_m. Notice that each entry of $y_{u,m}$ belongs to the interval $[0, 1]$. The overall pseudo label set is denoted by $Y_u = [y_{u,1}, y_{u,2}, ..., y_{u,M}] \in \mathbb{R}^{C \times M}$. We also define the latent parameter vector of $\tilde{y}_{u,m}$ as $\tilde{y}_{u,m} = \sigma^{-1}(y_{u,m}) \in \mathbb{R}^C$ for image x_m where σ is the sigmoid function. The prediction set is $Y_p = [y_{p,1}, y_{p,2}, ..., y_{p,M}]$ where $y_{p,m} \in \mathbb{R}^C$ is the vector of the predicted labels for image x_m. In CLIP model, there are two encoders: the visual encoder and the text encoder, which are denoted by E_v and E_t, respectively. The visual encoder maps the input image x_m to the visual embedding vector $E_v(x_m) = f_m \in \mathbb{R}^K$ where K is the dimension length of the embedding. Similarly, the text encoder maps the input text (class i, $i = 1, \cdots, C$) to the text embedding vector $E_t(i) = w_i \in \mathbb{R}^K$. Here the input text is a predefined prompt, such as "a photo of a cat". Given a vector or a matrix Q, $Q^\top$ means the transpose of Q.

Overview. The proposed framework is shown in Figure 3, to address unsupervised multi-label image classification, which includes three stages: initialization, training, and inference. During initialization, the goal is to appropriately initialize the pseudo labels for the unobserved labels on each training image. Taking advantage of the off-the-shelf CLIP model, we propose a CLIP-driven approach to build the pseudo labels upon the aggregation of global and local semantic-visual alignments, which can significantly improve the quality of pseudo-labels. During training, the pseudo labels obtained in initialization will be used as the estimation of the unobserved labels to initialize training of the classification network. We propose an optimization method that minimizes the total loss by recursively updating the network parameters and the latent parameters of the pseudo-labels. During inference, only the classification network is used to predict the labels of the input image.

3.1. Pseudo Label Initialization

3.1.1 Global Alignment Based on CLIP

CLIP is a powerful vision-language model that focuses on learning the global representation (the dominant concept) of an image (Figure 2). Therefore, we can directly use the CLIP model to generate the global alignment of an image without tuning the model parameters. Since the following discussion only focuses on an individual image x_m, we will drop the index m for notational simplicity.

Given an input image, the visual encoder of CLIP maps it to the embedding vector f. The relevant similarity score between f and the text embedding w_i is given by

$$p_i^{glob} = \frac{f^\top w_i}{||f|| \cdot ||w_i||}, \quad \forall 1 \leq i \leq C \tag{1}$$

$$s_i^{glob} = \frac{\exp(p_i^{glob}/\tau)}{\sum_{i=1}^{C} \exp(p_i^{glob}/\tau)}, \tag{2}$$

where p_i^{glob} denotes cosine similarity score between f and w_i for class i on the input image, s_i^{glob} is the normalized value for the similarity score using softmax, and τ is the temperature parameter learned by CLIP. Then, the soft global

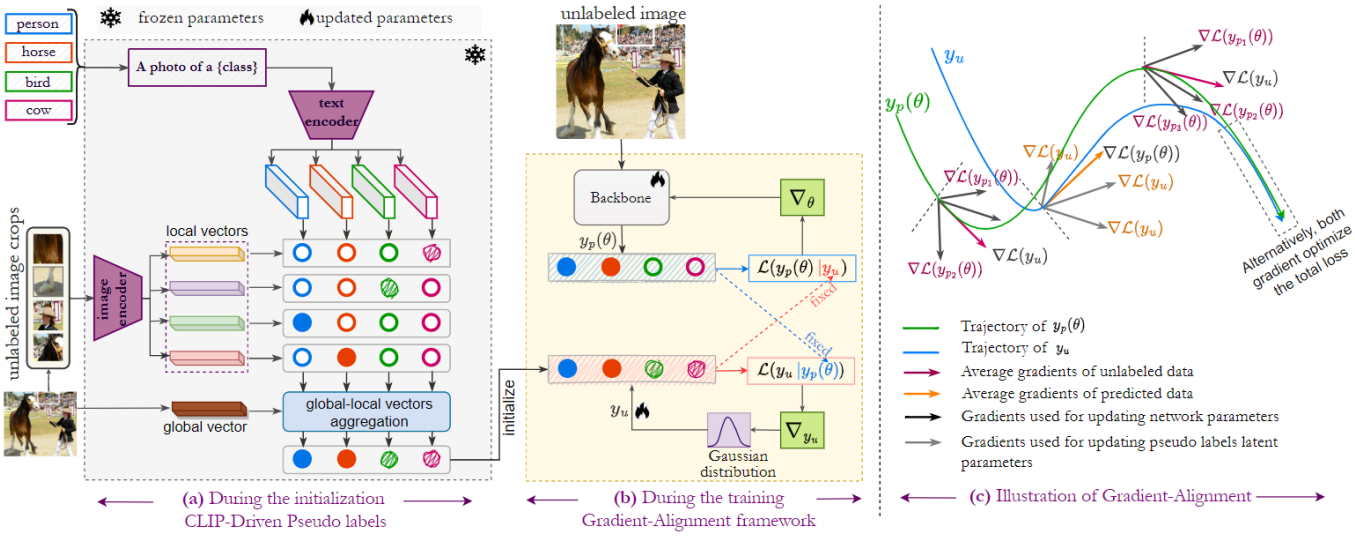

Figure 3. The overall framework for CDUL unsupervised multi-label image classification. **(a) During initialization**, we propose CLIP-driven global and local alignment and aggregation to generate pseudo labels. (i) Given an image, CLIP predicts the global similarity vector S^{global}; (ii) Given the snippets of this image, CLIP predicts local similarity vectors S_j^{local}; (iii) The global-local aggregator is used to generate the pseudo labels S^{final}. **(b) During training**, the pseudo labels generated from initialization are use to supervise the training of the classification network, using our proposed method *gradient-alignment method*. **(c)** The **gradient alignment illustration** shows that updating the network parameters and the pseudo labels by turns pushes both the pseudo label y_u and the predicted label y_p to the optimal solution to minimize the total loss function. **During inference**, we apply the whole image to the classification network to get the multi-label predictions.

vector of this image is defined as

$$S^{global} = \{s_1^{glob}, s_2^{glob}, ..., s_C^{glob}\},$$

which includes the similarity score for each class.

Notice that CLIP focuses on the most relevant class on an image and is limited to predict only one single label per image, while an image may contain multiple labels. In some cases, the highest confidence score from the CLIP prediction is not even correct due to the lack of appropriate prompt design (Figure 2.a). To alleviate this issue, we shift CLIP's attention from global to the local level, i.e., using CLIP to predict snippets of an image rather than the entire image, which will be discussed in the next subsection.

3.1.2 CLIP-Driven Local Alignment

To generate the local alignment, we split an input image to N snippets, denoted by $\{r_j\}_{j=1,...,N}$. Each snippet may contain multiple objects rather than just a single object. Accordingly, the visual embedding vector $g_j \in \mathbb{R}^K$ of snip r_j is extracted from the visual encoder, $E_v(r_j) = g_j$. Each image snippet is handled separately by finding the cosine similarity scores $p_{j,i}^{loc}$ between the snippet visual embedding g_j and the text embedding w_i for class i:

$$p_{j,i}^{loc} = \frac{g_j^\top w_i}{||g_j|| \cdot ||w_i||}, \ \forall\, 1 \le j \le N,\ 1 \le i \le C \quad (3)$$

The similarity scores will be forwarded to the Softmax function that normalizes these scores over all classes:

$$s_{j,i}^{loc} = \frac{\exp(p_{j,i}^{loc}/\tau)}{\sum_{i=1}^{C} \exp(p_{j,i}^{loc}/\tau)}. \quad (4)$$

So the local soft similarity vector S_j^{local} of snippet r_j is given by $S_j^{local} = \{s_1^{loc}, s_2^{loc}, ..., s_C^{loc}\}$.

Notice that different snippets may contain different objects or different attributes for the same object. Therefore, a specific class, which cannot obtain the highest similarity score from CLIP when focusing on the entire image, may now get the highest score in several snippets. Such a local alignment can enhance semantic transfer per snippet. Figure 4 shows the comparison of the confidence score distributions of using CLIP to predict three classes ("bottle", "chair", and "tvmonitor") in PASCAL VOC 2012 dataset, using global images and local snippets, respectively. It can be observed that when focusing on global images, CLIP may neglect some classes due to the "domain gap" between the pre-training datasets used to train CLIP and the target multi-label dataset. For instance, in Figure 4.b, the "chair" class get very low scores in most images, which means that very few "chair" labels will be predicted. This will affect the training performance at the training stage. When snippets are considered, they can enhance the prediction distribution toward higher confidence scores. It is worth mentioning that, as a cropping method, CLIP-Driven Local Alignment (CDLA) has advantages over class-agnostic object detection (COD) [21]. Our CDLA does not need the ground truth to get the snippets, while COD needs the ground truth to train the model to extract the snippets containing the objects.

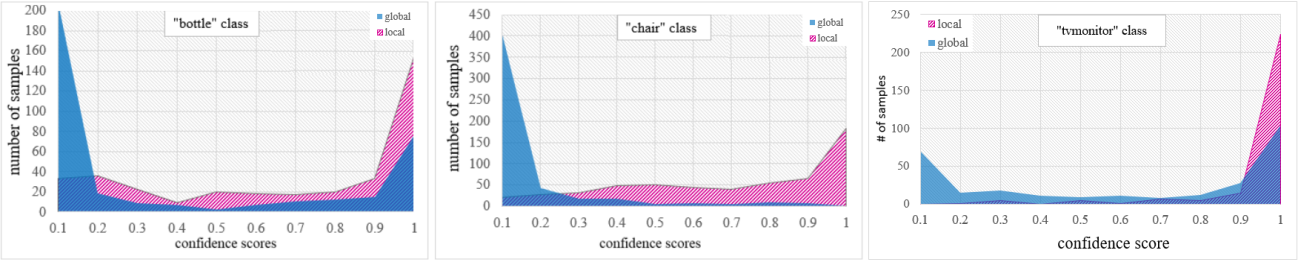

Figure 4. The distributions of the predicted labels across the confidence scores using off-the-shelf CLIP on the whole image (global) and snappets (local).

Thus, our CDLA is less expensive than COD in computation. Moreover, our CDLA technique is more robust than COD in the situations where the object of interest is occluded or partially visible. In those cases, object detection methods may not be able to detect the object, while cutting images may still be valid to obtain a partial view of the object.

3.1.3 Global-Local Image-Text Similarity Aggregator

Note that each input image is associated with one global image-text similarity vector S^{global} and N local similarity vectors S_j^{local}. We propose an aggregation strategy that complements mutual information from S_j^{local} and generates a unified local similarity vector $S^{aggregate}$ for each image, using a min-max method. Let

$$\alpha_i = \max_{j=1,\cdots,N} s_{j,i}^{loc},$$
$$\beta_i = \min_{j=1,\cdots,N} s_{j,i}^{loc}, \quad \forall\, 1 \le i \le C$$

and

$$\gamma_i = \begin{cases} 1 & \alpha_i \ge \zeta \\ 0 & \alpha_i < \zeta \end{cases} \tag{5}$$

where ζ is the threshold parameter. The aggregation score for class i is given by

$$s_i^{ag} = \gamma_i \alpha_i + (1-\gamma_i)\beta_i. \tag{6}$$

This strategy basically means that if the highest similarity score that class i obtains among all snippets, α_i, is greater than ζ, we will consider that this class likely exists on the image with α_i assigned to s_i^{ag}. On the contrary, if the similarity scores of class i in all snippets are less than ζ, the likelihood of class i existing on this image is small. Therefore, the strategy assigns the minimum score β_i to s_i^{ag}. With the aggregation scores, we define the soft aggregation vector of all classes for each input image as follows:

$$S^{aggregate} = \{s_1^{ag}, s_2^{ag}, ..., s_C^{ag}\}.$$

Now we can leverage the global similarity, which adds more comprehensive and complementary semantics, to local similarity by calculating the average:

$$S^{final} = \frac{1}{2}\left(S^{global} + S^{aggregate}\right),$$

which will be used as the initial pseudo labels for unobserved labels at the training stage. The high quality of S^{final} will significantly enhance the training performance, which is discussed in Subsection 4.3.

3.2. Gradient-Alignment Network Training

This subsection proposes the gradient-alignment method to leverage unsupervised consistency regularization, which updates the network parameters and the pseudo labels by turns. To be more specific, one can first train the network parameters according to the Kullback-Leibler (KL) loss function between the predicted labels and the initial pseudo labels obtained from the initialization stage, treating the pseudo labels as constants. After that, we fix the predicted labels and employ the gradient of the loss function with respect to the pseudo labels to update the latent parameters of pseudo labels. Once the pseudo labels are updated, we can fix them again and re-update the network parameters. This optimization procedure will continue until convergence occurs or the maximum number of epochs is reached. This idea is inspired by the previous work [44, 10, 2], which shows that during the training process the previously generated pseudo labels can provide valuable information to supervise the network. The detailed procedure can be described as follows. At the beginning of training, the pseudo label vector is initialized by S^{final} from the global-local aggregation module, i.e., $y_u = S^{final}$. Then y_u will be fixed and used to supervise the train of the network with the Kullback-Leibler (KL) loss function $\mathcal{L}(Y_p|Y_u, \mathcal{X})$. When the training is done, we fix the predicted labels Y_p and update the latent parameters of pseudo labels $\tilde{y}_u$:

$$\tilde{y}_u = \tilde{y}_u - \psi(y_u) \circ \nabla_{y_u} \mathcal{L}(Y_u|Y_p, \mathcal{X}) \tag{7}$$

where $\circ$ means the element-wise multiplication, $y_u = \sigma(\tilde{y}_u)$, and $\psi(y_u)$ is a Gaussian distribution with the mean at 0.5 which is given by:

$$\psi([y_u]_i) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{[y_u]_i - 0.5}{\sigma}\right)^2}. \tag{8}$$

Here $[y_u]_i$ is the ith entry of the vector y_u. Since the whole dataset is unlabelled, we need to use $\psi(y_u)$ to increase the rate of change for the unconfident pseudo labels and reduce the rate for the confident pseudo labels. The Gaussian distribution can perform as such a function for pseudo labels. The confidence of the pseudo label is evaluated based on $2|[y_u]_i - 1|$. For instance, if $[y_u]_i$ is 0.5, the Gaussian distribution will achieve its maximal value, which means that our module is not confident about the pseudo label and the rate of

Table 1. Mean average precision mAP in (%) for different multi-label classification methods under different supervision levels: Fully supervised, Weakly supervised and unsupervised, in addition to compare to zero-shot CLIP for four different datasets. Blue color represents the best results.

Supervision level	Annotation	Method	VOC2012	VOC2007	COCO	NUS
Fully Supervised	Fully labeled	BCE-LS [12]	91.6	92.6	79.4	51.7
		BCE	90.1	91.3	78.5	50.7
Weakly Supervised	10% labeled	SARB *et al.*[30]	-	85.7	72.5	-
		ASL *et al.*[3]	-	82.9	69.7	-
		Chen *et al.*[7]	-	81.5	68.1	-
	one observed labeled	LL-R [24]	89.7	90.6	72.6	47.4
		G^2NetPL [1]	89.5	89.9	72.5	48.5
Unsupervised	Annotation-free	Naive AN [25]	85.5	86.5	65.1	40.8
		Szegedy *et al.* [35]	86.8	87.9	65.5	41.3
		Aodha *et al.*[28]	84.2	86.2	63.9	40.1
		Durand *et al.*[15]	81.3	83.1	63.2	39.4
		ROLE [12]	82.6	84.6	67.1	43.2
		CDUL (ours)	88.6	89.0	69.2	44.0

change should contribute more in the iteration of the pseudo label to push it away from 0.5. Otherwise, if $[y_{u,t}]_i = 0$ or $[y_{u,t}]_i = 1$, the Gaussian distribution value reaches its minimum, which indicates high confidence on the current pseudo label and the rate of change should contribute less so that the value of the pseudo label can be maximally kept.

Once Y_u is updated, we switch back to the network training with a fixed Y_u again. This procedure will continue until convergence takes place or the maximum number of epochs is reached. As shown in Figure 3.c, this training process pushes both Y_u and Y_p (the predictions is a function of the network parameters) to a non-trivial optimal point to minimizing $\mathcal{L}(Y_p, Y_u|\mathcal{X})$.

3.3. Inference

We simply feed the whole image, without splitting, to the classification network to get the prediction. It is worth noting that we use the whole image without cropping during the training and testing process to reduce the computational cost.

4. Experiments

4.1. Setups

Datasets. We evaluate our model on four different multi-label image classification datasets. PASCAL VOC 2012 [16] has 5,717 training images and 5,823 images in the official validation set for testing, while PASCAL VOC 2007 contains a training set of 5,011 images and a test set of 4,952 images. MS-COCO [26] consists of 80 classes, with 82,081 training images and 40,137 testing images. NUSWIDE [11] has nearly 150K color images with various resolutions for training and 60.2K for testing, associated with 81 classes. The validation set is used for testing, whereas the training set is used to extract pseudo labels. During the training, we used these datasets without any ground truth.

Implementation Details. For initialization, to generate the pseudo labels based on our strategy, we use CLIP with ResNet-50×64 as image encoder and keep the same CLIP Transformer [37, 32] as the text encoder. During the training, for fair comparisons, all the models are trained using the same classification network architecture ResNet−101 [18], which is pre-trained on ImageNet [13] dataset. End-to-end training is used to update the parameters of the backbone and the classifier for 20 epochs. We train all four datasets using 10^{-5} learning rate. The batch size is chosen as 8 for both VOC datasets, and 16 for the COCO and NUS datasets.

Pre-Training Setting. As previously mentioned, our classification network is trained on unlabeled images without the use of any manual annotations during training; they are solely reserved for evaluation purposes. Therefore, we adapt CLIP using our global-local aggregation strategy to generate the pseudo labels for unlabeled data. We do not change or fine-tune the CLIP encoders or the prompt parameters, in which one fixed prompt is used, "a photo of the [class]", for all datasets. To get the local similarity vectors, we split the input images into 3x3 snippet images to generate image embedding for each snippet, in addition to generating an embedding for the whole image to enhance the quality of the generated pseudo labels. All the unsupervised models are initialized and trained using our generated pseudo labels as initials for the unlabeled data. Additionally, CLIP is not exploited during the training or inference processes.

Evaluation Metrics. For a fair comparison, we follow current works [12, 24, 1] that adopt the mean average precision (mAP), across the entire classes, as a metric for evaluation. We also measure the average precision (AP) per class to evaluate the class-wise improvement.

4.2. Comparison with State-of-the-art Methods

Mean Average Precision mAP Results. We report mAP (%) results compared to the state-of-the-art models under different supervision levels in Table 1 on four different multi-label datasets. We compare our model to three different supervision levels. *Fully supervised level* is used as a reference to the performance of the models using the fully labeled data [12, 24], upper part in Table 1. At *weakly supervised*

Table 2. AP and mAP (in %) of unsupervised methods on PASCAL VOC 2012 dataset for all classes. ALL methods trained by our proposed pseudo labels. Blue color represents the best results.

Methods	aeroplane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	diningtable	dog	horse	motorbike	person	pottedplant	sheep	sofa	train	tvmonitor	mAP
Naive AN [25]	98.9	92.3	96.2	86.6	70.0	94.7	83.6	97.6	73.3	80.4	70.0	94.8	85.5	92.3	91.6	60.7	90.5	70.8	97.3	83.7	85.5
Szegedy *et al.* [35]	99.2	92.6	97.2	91.6	71.1	93.6	81.8	98.4	75.7	89.7	69.4	96.2	86.9	93.9	92.5	65.6	93.2	72.3	98.2	77.3	86.8
Aodha *et al.*[28]	99.1	92.3	96.3	87.4	71.2	94.3	82.9	97.4	73.9	70.5	69.4	92.1	81.4	90.9	92.3	59.0	85.3	67.3	97.3	81.8	84.1
ROLE [12]	96.3	86.1	92.5	85.5	64.7	93.7	80.4	95.6	72.3	79.8	57.9	91.5	83.2	89.6	92.6	58.0	88.4	66.1	95.3	82.3	82.6
CDUL (ours)	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

level, we compare our model with [24, 1] methods that are trained using a single annotated label per each image following to [24]. We also compare our model with another group of weakly supervised models that used a partial number of annotation labels (10% per each image) for training such as [30, 7, 3]. Finally, in the third group report in Table 1, the *unsupervised level*, all the methods [25, 35, 28, 15, 12], including our method, are trained without any label annotation. We can observe that: ❶ Compared to fully supervised models, we drastically removed the manual labeling costs without sacrificing performance compared to the fully supervised scenario. ❷ Compared to weakly supervised models, our model can perform considerably better (mAP) comparable to those models without leveraging manually labeled data for the training set, which can be interpreted as meaning that our generated pseudo label that includes the high-quality fine-grained semantics based on our aggregator can help the classification network for training and get predictions that are competitive to those models that depend on partially annotated labels per image. Additionally, our model achieves better performance on the COCO and VOC 2007 datasets compared to the [7] method, which uses 10% annotated label. ❸ Compared to unsupervised models, our method outperforms the whole unsupervised models by a good margin on all datasets. The main reason is that our gradient-alignment optimization method can help the classification network to minimize the total loss based on the alternative updating methodology for the network parameters and the pseudo-label latent parameters. Our model can achieve +6.0%, +4.4%,and +2.1%, compared to Role [12] on VOC2012, VOC2007, and COCO, respectively. Our method cannot be simply classified as weakly supervised models due to distinct input characteristics. Our approach utilizes CLIP to generate pseudo labels for all images, which often contain numerous unknown and incorrect labels (e.g., mAP using original CLIP is 65.3% in COCO dataset). In contrast, weakly supervised models assumes that all provided partial labels are correct and can be trusted for training. This distinction is significant since our method tackles the joint training of the multi-label classification model and the refinement of incorrect pseudo labels, which is a key contribution of our work. Our method successfully increases the accuracy of CLIP-generated pseudo labels to 69.2% mAP in COCO as reported in Table 1.

Class-Wise AP Improvement. Table 2 reports the class-wise AP improvement for the unsupervised multi-label classification models on test sets of Pascal VOC 2012. All the methods start training based on our proposed global-local pseudo labels. Our method can improve performance in most classes in VOC 2012 dataset. We observe that although all the methods start training based on our proposed global-local pseudo labels, our model outperforms in most of the classes, especially the classes that have a small size, such as *"potted plant"*, *"book"*, *"cup"*, and *"wine glass"* which can be interpreted that gradient-alignment can help our model to capture more information during the training. We also demonstrate the Class Activation Maps (CAM) for some images in the COCO dataset, as shown in Figure 5. Our model can classify the *"bottle"* and *"wine glass"* in the first row, *cup* in the second row, and *"remote"* in the third row.

4.3. Ablation Study

Quality of Initial Pseudo Labels. To study the quality of the pseudo-labels, we measure the mAP for the CLIP-pseudo labels based on the global and local alignments using our proposed aggregation strategy (ours) as reported in Table 3. We also report different aggregation strategies such as ❶ average (avg): by getting the average between all the local and global similarity vectors, ❷ Maximum (max): to get the maximum similarity score per each class among all the local and global similarity vectors. As reported in Table 3, we observe that the quality of pseudo label for the global alignment achieves mAP less than any strategy depending on local alignment. We also observe that our aggregation strategy generates the highest quality pseudo labels compared to

Figure 5. Class activation maps for several examples corresponding to highest confidences for three labels on COCO dataset. The highlighted area indicates where the model focused to classify the image. Best viewed in color.

Table 3. Ablation study of quality of pseudo labels on the training set in three different datasets using ResNet-50 × 64.

Datasets	global alignment	global-local alignment		
		Aggregator		
		avg	max	ours
VOC 2012	85.3	88.5	89.5	90.3
COCO	65.4	70.0	71.6	72.8
NUS	41.2	41.8	42.3	43.1

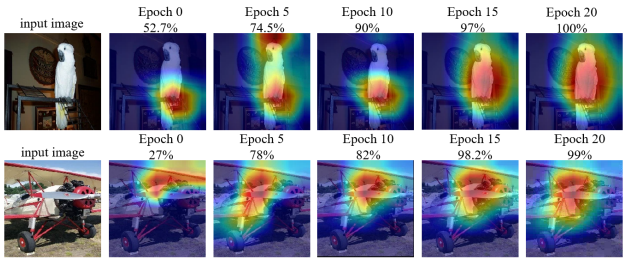

Figure 6. CAM visualization for the classification task on Pascal2012 dataset. CAM shows that the improvement of the classification during the epoch. Best viewed in color.

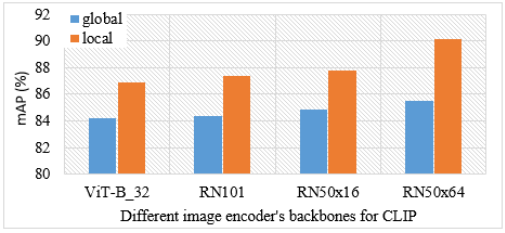

Figure 7. Quality of pseudo labels using different backbones for CLIP's image encoder

its global counterparts +5%, 7.4%, and 1.9% on Pascal VOC 2012, COCO, and NUS datasets, respectively. Consequently, our aggregation strategy can retain the most fine-grained semantics in the input image. We also prove that the good quality of the pseudo label helps the classification network to learn during the epochs and the classification performance is improved by demonstrating the CAM visualization for samples of images from Pascal VOC 2012. As shown in Figure 6, during the epochs, the classification network can learn the correct prediction and mAP is increased.

Different Backbones. In our study, we evaluate the quality of pseudo labels generated using different depths of image encoders in CLIP, namely ViT-B-32, ResNet-101, ResNet-50 × 16, and ResNet-50 × 64 following the EfficientNet-style model scaling [31]. We employ CLIP with different backbones during the initialization time to generate the pseudo labels on the training set of Pascal VoC 2007 dataset. The results in Figure 7 show that the quality of the generated pseudo labels consistently improves with different backbones. Furthermore, we observed that the quality of pseudo labels improves significantly with the use of local alignment, achieving up to a 2.7% improvement for ViT-B-32, 3% for ResNet-101, 2.9% for ResNet-50 × 16, and 4.6% for ResNet-50 × 64, compared to their global alignment counterparts. Since we use this backbone only once at the initialization

Table 4. Ablation study when initialized with various pseudo labels based on different CLIP's image encoder backbones. mAP results on Pascal 2007 dataset

ViT-B-32	ResNet-101	ResNet-50x16	ResNet-50x64
86.7	86.8	86.9	89.0

Table 5. Ablation study to evaluate CLIP-GLA's performance.

Models	Pascal2012	COCO	NUS	# param. (M)
CLIP-GLA	84.7	63.6	38.9	102
CDUL (ours)	86.4	67.1	41.9	25

to generate the pseudo labels, no computational cost will be used at the testing time. Additionally, the generated pseudo labels are used as initials for all unsupervised multi-label models. As reported in Table 4, the performance of our model is improved with different backbones, achieving up to (+2%) increase using ResNet-50 × 64 compared to its counterparts. The improvement is reasonably related to the quality of pseudo labels used during the training Figure 7.

CLIP Performance on Test Set. Table 5 presents a comparison of our ResNet-50-based model's performance with on-the-shelf CLIP with ResNet-50 combined with our global-local alignment strategy (CLIP-GLA). Our model surpasses CLIP-GLA in both mAP. Moreover, our model uses a smaller number of parameters during inference. This is due to the fact that in our framework, CLIP is used exclusively to initialize the pseudo labels during the initialization phase, while a smaller backbone is utilized during the training and testing phases, as illustrated in Fig. 3. Thus, our approach is more cost-effective during both the training and inference phases.

Effectiveness of Other Parameters. We also study the impact of removing the Gaussian distribution module in the gradient-alignment training, the performance is dropped by 0.5%, and 0.4% on VoC 2012, and COCO datasets, respectively, as compared to our model (Table 1). Additionally, we study applying the hard pseudo labels instead of soft pseudo labels; the performance is reduced by 0.9%, and 1.2% on VoC 2012, and COCO datasets, respectively, as compared to our model (Table 1).

5. Conclusions

In this paper, we propose a new method for unsupervised multi-label image classification tasks without using human annotation. Our key innovation is to modify the vision-language pre-train model to generate the soft pseudo labels, which can help training the classification network. To inject the fine-grained semantics in the generated pseudo labels, we proposed a new aggregator that combines the local(global) similarity vectors between image snippets(whole image) and text embedding. Finally, we use the generated pseudo label to train the network classification based on the gradient-alignment to get multi-label classification prediction without any annotation. Extensive experiments show that our method outperforms state-of-the-art unsupervised methods.

References

[1] Rabab Abdelfattah, Xin Zhang, Mostafa M Fouda, Xiaofeng Wang, and Song Wang. G2netpl: Generic game-theoretic network for partial-label image classification. In *Proceedings of the British Machine Vision Conference*, 2022. 1, 6, 7

[2] Rabab Abdelfattah, Xin Zhang, Zhenyao Wu, Xinyi Wu, Xiaofeng Wang, and Song Wang. Plmcl: Partial-label momentum curriculum learning for multi-label image classification. In *Proceedings of the European Conference on Computer Vision workshop*, 2022. 5

[3] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020. 6, 7

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

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[6] Tianshui Chen, Liang Lin, Xiaolu Hui, Riquan Chen, and Hefeng Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[7] 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*, volume 36, pages 339–346, 2022. 2, 6, 7

[8] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 522–531, 2019. 1

[9] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019. 1

[10] Yoonki Cho, Woo Jae Kim, Seunghoon Hong, and Sung-Eui Yoon. Part-based pseudo label refinement for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7308–7318, 2022. 5

[11] 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

[12] 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, 3, 6, 7, 11

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[14] 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

[15] 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, 3, 6, 7

[16] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007:1–45, 2012. 6

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[19] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3

[20] 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, 11

[21] Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and Premkumar Natarajan. Class-agnostic object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 919–928, 2021. 4

[22] 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

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

[24] 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. 2, 6, 7

[25] 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

[26] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[27] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 2

[28] 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, 3, 6, 7

[29] 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

[30] Tao Pu, Tianshui Chen, Hefeng Wu, and Liang Lin. Semantic-aware representation blending for multi-label image recognition with partial labels. 2022. 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, 8

[32] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 6

[33] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *arXiv preprint arXiv:2206.09541*, 2022. 2

[34] Yu-Yin Sun, Yin Zhang, and Zhi-Hua Zhou. Multi-label learning with weak label. In *Twenty-fourth AAAI conference on artificial intelligence*, 2010. 2

[35] 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, 3, 6, 7

[36] 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

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 6

[38] Dongkai Wang and Shiliang Zhang. Unsupervised person re-identification via multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10981–10990, 2020. 3

[39] Qifan Wang, Bin Shen, Shumiao Wang, Liang Li, and Luo Si. Binary codes embedding for fast image tagging with incomplete labels. In *European Conference on Computer Vision*, pages 425–439. Springer, 2014. 2

[40] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[41] 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

[42] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3733–3742, 2018. 3

[43] Miao Xu, Rong Jin, and Zhi-Hua Zhou. Speedup matrix completion with side information: Application to multi-label learning. *Advances in neural information processing systems*, 26, 2013. 2

[44] Xiao Zhang, Yixiao Ge, Yu Qiao, and Hongsheng Li. Refining pseudo labels with clustering consensus over generations for unsupervised object re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2021. 3, 5

[45] Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7369–7378, 2022. 3

[46] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022. 2

[47] Chengxu Zhuang, Alex Lin Zhai, and Daniel Yamins. Local aggregation for unsupervised learning of visual embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6002–6012, 2019. 3

6. Supplementary Material

6.1. Initialization:

Global-Local Aggregator:

We used the off-shelf CLIP to get similarity scores for both the entire input image (referred to as "global") and and each individual snippet within the image (referred to as "local"). Subsequently, we employed two different aggregation approaches to get S^{final}: (I) aggregation based on maximum ($\lambda = 1$): by getting the maximum similarity score per each class among global and max-min local similarity vectors, and (II) aggregation based on average ($\lambda = 0$): by averaging between the global similarity scores and max-min of local similarity scores for each class. S^{final} is described as follow;

$$S^{final} = \frac{1}{2}\left(S^{global} + S^{aggregate} + \lambda|S^{global} - S^{aggregate}|\right),$$

where λ is the smoothing hyper-parameter changes between the aggregation based on maximum to aggregation based on average across the global and min-max local similarity scores.

6.2. During the training:

We trained the network using Kullback-Leibler (KL) loss function. Then fix the predicted labels to update the latent parameters of pseudo labels using equation (7), and Gaussian distribution with the mean at 0.5, which is given by:

$$\psi([y_u]_i) = \frac{c1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{[y_u]_i - 0.5}{\sigma}\right)^2} + c2$$

where $c1$ and σ are the hyperparameters and $c2$ is a constant that ensures the Gaussian function $\psi([y_u]_i)$ close to zero when $[y_u]_i$ has high confidence score with values at 0 and 1. At epoch 0, the latent parameters are initialized with pseudo labels obtained during the initialization phase via a local-global aggregator. We used warm-up until epoch 3 without updating the pseudo labels. Starting from epoch 4, the network parameters and the latent parameters of pseudo labels are updated alternatively and reported the results at epoch 20. We initialized the latent parameters with pseudo labels aggregated in different cases, as discussed in section 6.1. For example, the pseudo labels are initialized with aggregated scores at $\lambda = 1$ in Table 1. The ζ values range from 0 to 0.4, where the global-local aggregated pseudo labels can achieve mAPs higher than the global pseudo labels.

6.3. Testing Phase:

During the test phase, we only used the network to test the input image, where the network takes an entire image as input rather than snippets.

7. Evaluation Metrics

This section introduces the metrics used to evaluate the performance of the network for multi-label image classification. We assume that each image is assigned with the estimated label vector y_o, whose entries are soft pseudo labels from the global-local aggregator. During testing, each image is associated with the fully labeled ground truth y_g, whose entries can be 1 or 0, representing observed positive or observed negative labels, respectively.

The mean average precision (mAP) is applied to evaluate the performance of different approaches for multi-label classification in our paper, similar to [12, 20]. We measure the average precision (AP) for each class to calculate the mAP across all L classes as following:

$$mAP = \frac{1}{L}\sum_{\ell=1}^{L} AP_\ell. \tag{9}$$