南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

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

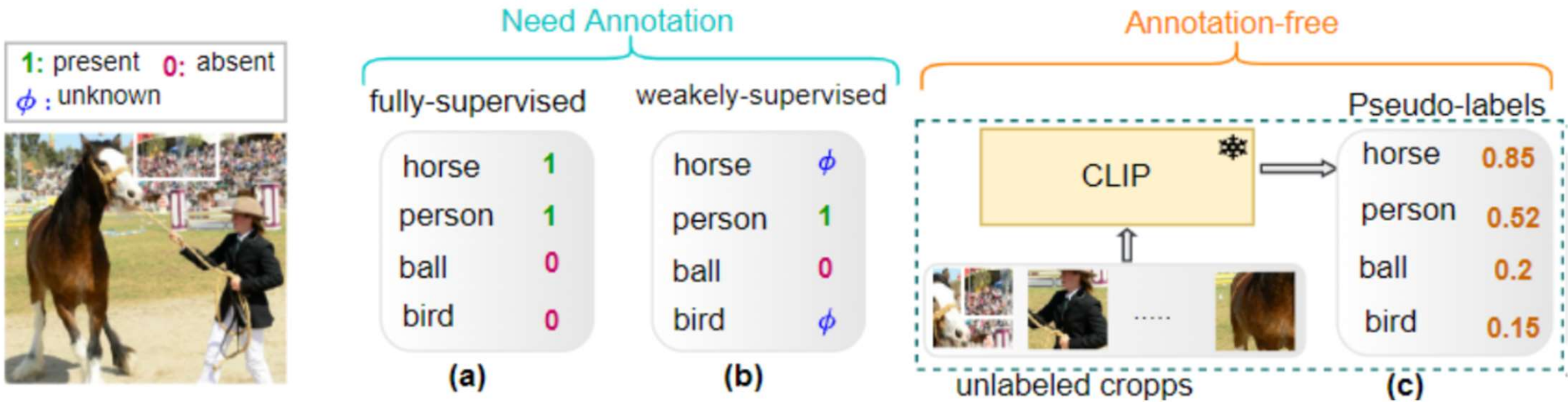
Rabab Abdelfattah[1], Qing Guo[2], Xiaoguang Li[3], Xiaofeng Wang[3], and Song Wang[3]

[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

ICCV 2023

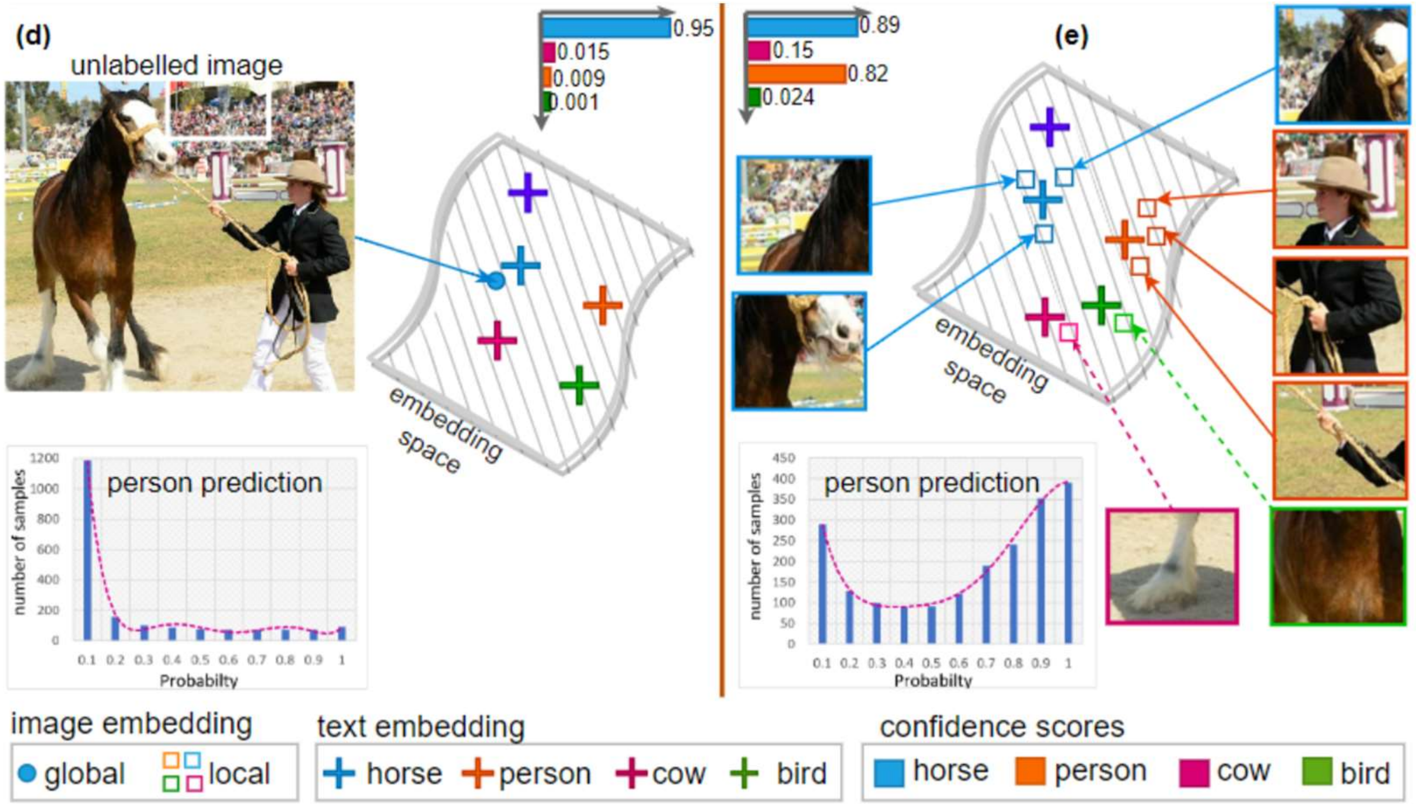
1 Overview

- A multi-label classification task aims to predict all the objects within the input image.
- 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.
- To alleviate the annotation burden, weakly supervised learning approaches have been studied which only a limited number of objects are labeled on a subset of training images which it still requires intensive manpower and time for annotations.
- 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 pretraining (CLIP).

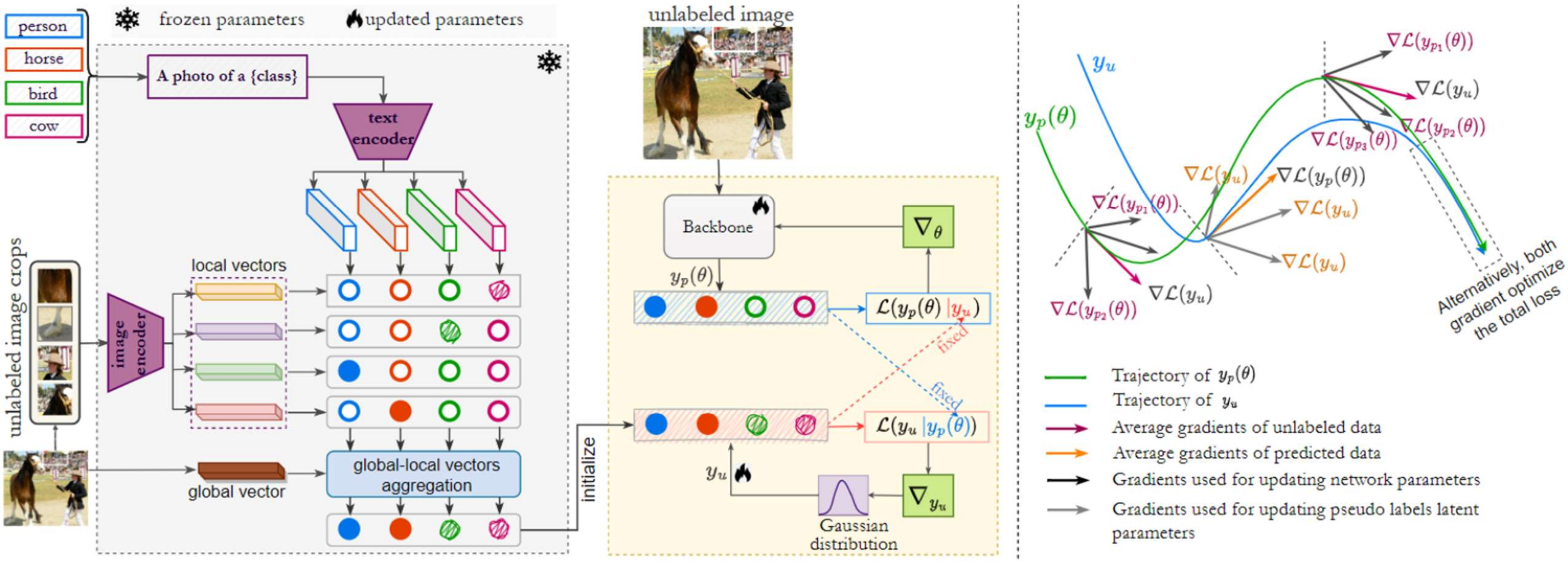


2 Motivation

CLIP is not suitable for multi-label classification, since it is trained only for recognizing a single object per image.



3 Proposed Model

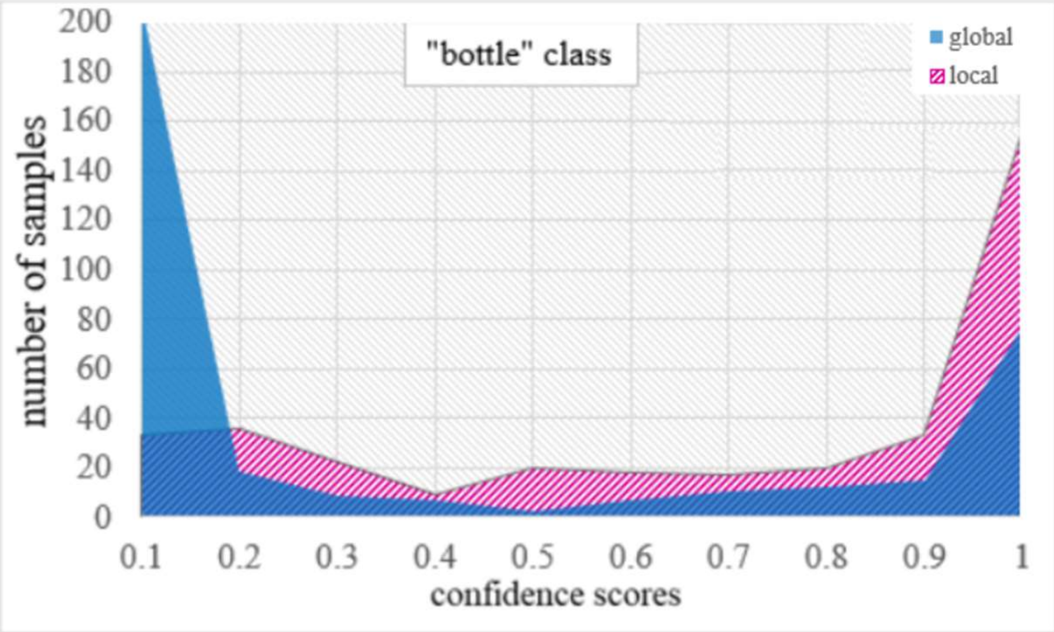


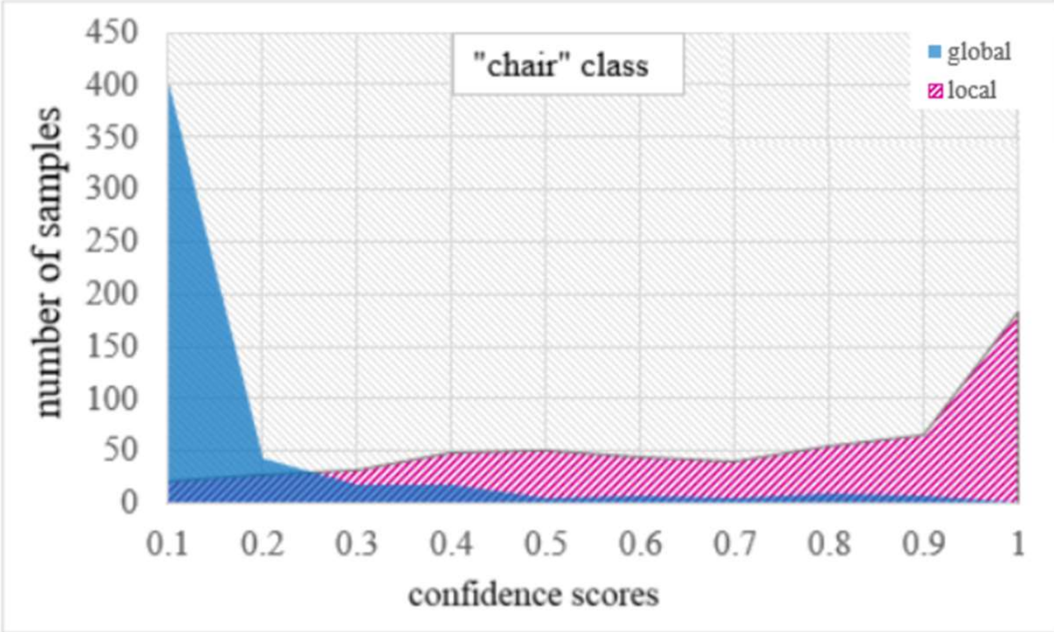
(a) During the initialization CLIP-Driven Pseudo labels

(b) During the training Gradient-Alignment framework

(c) Illustration of Gradient-Alignment

3 Proposed Model



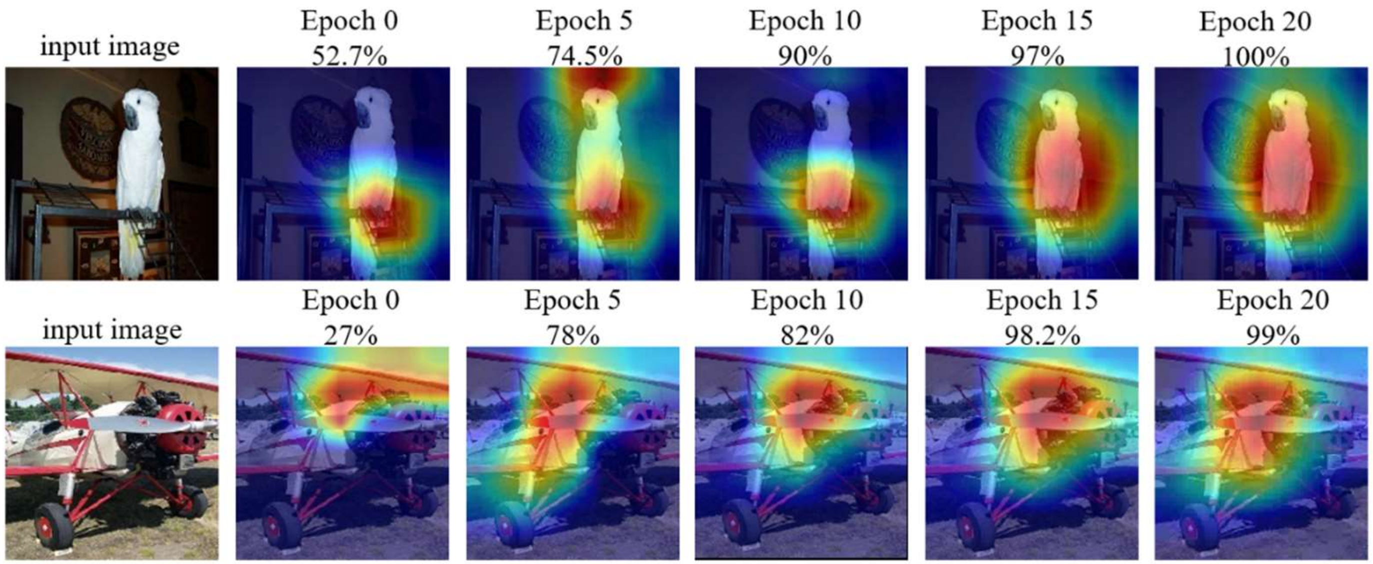


4 Experiments

Quality of pseudo labels on the training set in three different datasets using ResNet-50 x 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

CAM visualization for the classification task on VOC2012 dataset. CAM shows that the improvement of the classification during the epoch.



4 Experiments

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.

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

5 Conclusion

- To the best of our knowledge, this is the first work that applies CLIP for unsupervised multi-label image classification.
- Our key innovation is to modify the vision-language pre-train model to the soft pseudo labels, which can help training the classification network.
- 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.

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

模式分析与机器智能
工业和信息化部重点实验室
MIIT Key Laboratory of
Pattern Analysis & Machine Intelligence

THANKS