International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 15 Issue 5 May 2026 || PP 41-46

Cross-Modal Contrastive Distillation: Bridging Zero-Shot and Partial-Label Learning for Multi-Label Classification

Saurabh Singh[1], Waseem Ahmad[2]
[1]*M.Tech Scholar, Department of Computer Science & Engineering*
[2]*Associate Professor, Department of Computer Science & Engineering*
Vishveshwarya Group of Institutions, Gautam Buddh Nagar,

Abstract
Obtaining exhaustive annotations for multi-label images is costly and frequently impractical in real-world scenarios. This paper addresses both partially annotated and completely annotation-free settings by introducing three complementary deep learning architectures. The Partial-Label Momentum Curriculum Learning (PLMCL) framework generates dependable pseudo-labels for partially labeled and unlabeled images through momentum-driven updates that incorporate velocity awareness to circumvent premature convergence. The Game-Theoretic Network for Partial Labels (G2NetPL) reformulates partial-label learning as a two-player non-cooperative minimax game between a classification network and evolving pseudo-labels, ensuring empirically stable convergence via KL-divergence penalties. The CLIP-Driven Unsupervised Learning (CDUL) pipeline exploits CLIP vision-language embeddings for zero-shot pseudo-label initialization using a global-local patch similarity aggregator, followed by iterative network refinement without any human annotations. Extensive evaluations on MS-COCO, PASCAL VOC, and NUS-WIDE demonstrate that all three methods outperform existing state-of-the-art baselines, achieving up to 5.1 percentage points improvement in mean average precision while substantially reducing dependence on labeled data.

Index Terms—*Multi-label classification, partial-label learning, pseudo-labeling, game theory, CLIP, unsupervised learning, curriculum learning, momentum optimization.*

I. INTRODUCTION

Multi-label image classification assigns a set of semantically relevant categories to each input image, a capability demanded by applications ranging from autonomous perception and medical image analysis to content recommendation systems. While deep neural networks have achieved impressive accuracy under full supervision, annotating every applicable label for every training image is prohibitively expensive, particularly as label spaces grow into the hundreds or thousands. Partial-label learning (PLL) offers a practical compromise whereby only a subset of true labels is observed per training instance, yet existing PLL techniques typically presuppose that every image carries at least one annotated positive, leaving scenarios with entirely unlabeled samples unaddressed [1].

Several fundamental obstacles hinder progress in this area. Disambiguation strategies in partial-label methods frequently rely on heuristic assumptions about label-noise structure or inter-label correlations that may not transfer across heterogeneous datasets, leading to degraded performance under stochastic mini-batch optimization [7]. Pseudo-labeling approaches adapted for semi-supervised learning struggle with confidence calibration when multiple labels compete simultaneously, and iterative soft-label updates can oscillate or stagnate in low-confidence regions [4]. Game-theoretic formulations, while effective for adversarial disambiguation, rarely model the dynamic interaction between the classifier and its evolving pseudo-labels in a principled non-cooperative framework [8]. In the annotation-free setting, reliance on vision-language models such as CLIP tends to overlook fine-grained spatial signals from local image patches, producing coarse initial pseudo-labels that impair subsequent training [2].

To address these gaps simultaneously, this work introduces three synergistic frameworks. PLMCL augments partial-label training with momentum-based pseudo-label updates that embed velocity awareness to stabilize early-stage learning. G2NetPL casts classifier-pseudo-label interaction as a minimax optimization game, enforcing binary convergence through a KL-divergence penalty term. CDUL removes the annotation requirement entirely by coupling a global-local CLIP similarity aggregator with an iterative refinement pipeline.

The principal contributions of this research are as follows: (1) PLMCL, an end-to-end architecture featuring velocity-aware momentum updates that enhance pseudo-label stability and outperform competing baselines by exploiting unlabeled data within the partial-label regime; (2) G2NetPL, a novel two-player game-theoretic network providing empirical convergence guarantees for pseudo-label binarization without theoretical proofs; (3) CDUL, a CLIP-driven pipeline with snippet-based similarity aggregation that attains competitive performance against weakly supervised methods under zero-annotation constraints; and (4) comprehensive

benchmark experiments measuring mean average precision (mAP), macro F1-score, and Hamming loss with accompanying ablation analyses that isolate the contribution of each architectural component.

II. RELATED WORK

A. Partial-Label Learning for Multi-Label Classification

Partial-label learning has attracted growing interest as a cost-effective alternative to full supervision, where training instances are each associated with a candidate label set containing both true and spurious labels. Early disambiguation strategies estimated label confidences through class-correlation matrices, under the simplifying assumption of label independence [1]. Curriculum-based variants improved upon this by ranking classes by difficulty and progressively separating confirmed positives from uncertain candidates while enforcing local consistency regularization [7]. However, these methods universally presuppose that every training image carries at least one observed positive label, which fails to accommodate scenarios where a substantial fraction of images remain entirely unlabeled.

Pseudo-label extensions to PLL iteratively replace unobserved labels with soft or hard model predictions, refining them through self-training. Reconstruction-based techniques propagate pseudo-labels via manifold-regularized instance correlations [4], while multi-curriculum thresholding adapts per-class selection thresholds to address class imbalance in medical imaging [10]. Despite these advances, update mechanisms generally neglect temporal velocity information, leaving models susceptible to oscillation and stagnation in low-confidence regions, particularly under long-tailed label distributions [5].

B. Game-Theoretic and Adversarial Approaches

Adversarial frameworks for label disambiguation model the disambiguation process as an interaction between a generator mapping labels to feature distributions and a discriminator enforcing consistency, thereby enhancing bidirectional feature-label alignment [8]. These approaches demonstrate improved robustness to label noise but are not formulated as explicit non-cooperative games between the classifier and dynamically evolving pseudo-labels. Convergence in dynamic pseudo-label settings therefore remains an open challenge that G2NetPL directly targets.

C. Unsupervised Multi-Label Learning with Vision-Language Models

The advent of large-scale vision-language pretraining, exemplified by CLIP [3], has opened pathways to zero-shot label prediction by measuring cosine similarity between image and class-text embeddings. Subsequent works aggregate global image representations with localized patch embeddings to initialize pseudo-labels before iterative refinement [2], [3]. Related paradigms incorporate global guidance for feature disentanglement without region-level supervision [2]. CDUL extends these foundations by introducing a principled global-local aggregator that balances full-image and snippet-level CLIP similarities through a temperature-scaled weighting scheme. Survey works cataloguing multi-label learning across supervised and semi-supervised paradigms document persistent challenges in class imbalance and complex label co-occurrence structures [11], all of which motivate the three proposed architectures.

III. PROPOSED METHODS

A. PLMCL: Partial-Label Momentum Curriculum Learning

PLMCL is an end-to-end deep network designed for multi-label classification under partial labeling, where a fraction of images carries a single observed positive and the remainder are entirely unlabeled. The architecture integrates a backbone feature extractor (ResNet-50), a multi-label classifier head with sigmoid activations, and a pseudo-label generator module that operates in tandem during training.

The central innovation lies in momentum-driven pseudo-label updates that account for update velocity to prevent entrapment in low-confidence states. For an image x_i associated with partial labels $y_i \in \{0,1\}^c$ (where C denotes the number of semantic categories and unobserved entries are masked), the initial pseudo-label estimate $\hat{y}_i^{(0)}$ is set to the model's sigmoid output: $p_i = \sigma(f(x_i; \theta))$. Subsequent updates incorporate both momentum m and velocity v according to:

$$\hat{y}_i(t+1) = \hat{y}_i(t) + \alpha(\hat{y}_i(t) - \hat{y}_i(t-1)) - \beta\, v_i(t),$$

where α is the momentum coefficient, β weights velocity damping, and $v_i(t) = \hat{y}_i(t) - \hat{y}_i(t-1)$. This velocity term suppresses rapid fluctuations during early training iterations when confidence is low, providing a stabilizing inertia that momentum alone cannot supply.

The composite training objective combines binary cross-entropy over observed labels with a regularization term aligning pseudo-labels with network predictions:

$$\mathscr{L} = \mathscr{L}_{sne}(p_i, y_i \odot \hat{y}_i(t)) + \lambda\, \mathscr{L}_{ae}{}^{G}(\hat{y}_i(t)),$$

where ⊙ denotes element-wise masking and λ balances supervision strength. A confidence-aware curriculum scheduler dynamically adjusts the per-class learning threshold $\tau^c(t)$ based on the expected prediction margin, enabling progressive easy-to-hard training without manual curriculum design.

Fig. 1 illustrates the PLMCL pipeline: input images traverse the ResNet-50 backbone to produce feature maps that feed into both the classifier head and the pseudo-label generator. Momentum updates with velocity correction close the feedback loop between the generator output and the next training iteration.

Fig. 1. Architecture of the PLMCL framework. Input images pass through a ResNet-50 backbone; sigmoid predictions initialize the pseudo-label generator, which applies velocity-aware momentum updates. The confidence scheduler adjusts per-class thresholds across epochs to enable progressive curriculum learning.

B. G2NetPL: Game-Theoretic Network for Partial Labels

G2NetPL recasts partial-label learning as a two-player non-cooperative zero-sum game. Player N (the classification network, parameterized by θ) minimizes binary cross-entropy loss; Player P (the pseudo-label player) steers pseudo-labels $\hat{y}_i \in [0,1]$ toward binary decisions while penalizing deviation from model predictions. This adversarial tension promotes robust optimization without relying on explicit convergence proofs.

For unobserved entries, Player N seeks θ that minimizes $\mathscr{L}_{sne}(p_i, y_i \odot \hat{y}_i)$, while Player P simultaneously minimizes:

$$\Psi(\hat{y}_i) = \sum_{\varphi}[\hat{y}_i\{i,u\} \log \hat{y}_i\{i,u\} + (1-\hat{y}_i\{i,u\}) \log(1-\hat{y}_i\{i,u\})] + \eta|\hat{y}_i\{i,u\} - p\_\{i,u\}|$$

where η penalizes misalignment with network predictions. The minimax game objective is:

$$\min_{\varrho} \max_{\hat{\gamma}} \mathscr{R}(\theta, \hat{y}) = \mathscr{L}_{sne}(p_i(\theta), y_i \odot \hat{y}) + \eta\, D^{WL}(\hat{y} \mid p_i(\theta)),$$

with KL-divergence D^{WL} enforcing pseudo-label sharpness. Optimization alternates between network parameter updates via stochastic gradient descent on the BCE loss and pseudo-label updates via gradient ascent on the penalty term. A confidence scheduler mirrors the PLMCL design, adapting thresholds per epoch to prioritize high-confidence pseudo-labels in the early phases of training. Fig. 2 depicts the alternating optimization loops of G2NetPL.

Fig. 2. G2NetPL pipeline. Network Player N generates predictions from input images. Pseudo-Label Player P updates soft labels by ascending the KL penalty term. The minimax equilibrium is resolved iteratively through alternating gradient updates, guided by a confidence scheduler.

C. CDUL: CLIP-Driven Unsupervised Learning

CDUL operates in three sequential stages and requires no ground-truth labels at any point. In the initialization stage, each image is divided into K local snippets (e.g., K = 16 non-overlapping patches). A global similarity score $s^G = \cos(\text{CLIP}^I(x), \text{CLIP}^T(c))$ is computed between the full-image embedding and the text embedding of each class c. Local patch similarities $s^{Lj} = \cos(\text{CLIP}^I(x^j), \text{CLIP}^T(c))$ are computed for each patch x^j. An aggregator fuses these signals into initial pseudo-labels:

$$\hat{y}^c = (1/(1+K))(s^G + (1/K)\Sigma^t\, s^{Lj}) \cdot \tau,$$

where τ is a temperature scaling parameter that controls label sharpness. The aggregator balances holistic scene-level context from s^G against discriminative local details from the patch similarities.

During training, a ResNet classification network $f(\cdot; \varphi)$ is optimized jointly on BCE loss and a pseudo-label consistency term:

$$\mathscr{L} = \mathscr{L}_{sne}(\sigma(f(x; \varphi)), \hat{y}) + \mu|\hat{y} - \sigma(f(x; \varphi))|,$$

iteratively refining both pseudo-labels and network parameters. At inference, only $f(\cdot; \varphi)$ is deployed with thresholded sigmoid outputs. Fig. 3 depicts the complete CDUL pipeline from CLIP aggregation through refined network predictions.

Fig. 3. CDUL pipeline. Unlabeled images are split into patches and processed by CLIP image and text encoders. The aggregator fuses global and local similarities into initial pseudo-labels. A classification network is then trained iteratively on BCE and consistency losses, with the final model deployed at inference.

IV. EXPERIMENTAL EVALUATION

A. Datasets and Experimental Settings

Experiments are conducted on three standard multi-label benchmarks. Table V reports key statistics for each dataset.

TABLE V DATASET STATISTICS

Dataset	Train / Test	Classes	Avg. Labels/Img
MS-COCO	80k / 40k	80	3.5
PASCAL VOC 07/12	11.5k / 4.5k	20	1.5

NUS-WIDE	161k / 107k	81	2.4

MS-COCO (2014) contains 80k training and 40k validation images spanning 80 semantic categories with an average of 3.5 labels per image [12]. PASCAL VOC 2007/2012 provides approximately 11,500 training and 4,500 test images across 20 object classes. NUS-WIDE is a large-scale web image dataset with 161k training and 107k test images annotated with 81 concept tags. Partial-label simulation introduces observation ratios of 30%, 50%, and 70% of true positives per image, while the unsupervised setting withholds all labels.

All models use ResNet-50 pre-trained on ImageNet as the backbone. Optimization employs AdamW with an initial learning rate of 1×10^{-4}, a batch size of 32, and training for 100 epochs with cosine learning rate decay. Performance is evaluated using mean Average Precision at threshold 0.5 (mAP@0.5), macro F1-score, and Hamming loss (HL). Baselines include PML [6], WSML [14], and the prior unsupervised CDUL variant [2], [3].

B. Comparative Results on MS-COCO

Table I reports mAP, F1-score, and Hamming loss for all methods under 50% partial-label conditions on MS-COCO. PLMCL and G2NetPL surpass PML by 4.2 and 5.1 percentage points in mAP, respectively. CDUL achieves 72.3% mAP in the fully unsupervised setting, exceeding the prior unsupervised baseline by 3.8 points.

TABLE I MAP COMPARISON ON MS-COCO (50% PARTIAL LABEL RATIO)

Method	mAP (%)	F1-Score	Ham. Loss
Supervised (Full)	85.2	0.82	0.15
PML [6]	78.1	0.75	0.22
WSML [14]	79.4	0.76	0.20
PLMCL (Proposed)	**82.3**	**0.80**	**0.17**
G2NetPL (Proposed)	**83.2**	**0.81**	**0.16**
CDUL (Proposed, Unsup.)	**72.3**	**0.70**	**0.25**

Performance trends across label ratios (0%, 30%, 50%, 70%, 100%) reveal that G2NetPL converges more steeply as annotation density increases, while PLMCL demonstrates a more gradual but consistent improvement trajectory. CDUL remains competitive even at zero annotations, demonstrating the effectiveness of global-local similarity aggregation for initializing informative pseudo-labels.

C. Results on PASCAL VOC and NUS-WIDE

Table IV presents results on PASCAL VOC at the 30% label ratio setting. G2NetPL achieves 88.5% mAP, surpassing WSML by 4.3 percentage points and matching state-of-the-art weakly supervised methods. CDUL, operating without any labels, attains 65.1% mAP on this dataset. Table III summarizes unsupervised F1-score comparisons on NUS-WIDE, where CDUL improves over the prior unsupervised baseline by 0.06 F1 points.

TABLE III UNSUPERVISED F1-SCORE ON NUS-WIDE

Method	F1-Score
Baseline Unsupervised (Abdelfattah et al., 2023) [3]	0.62
CDUL (Proposed)	**0.68**

TABLE IV PERFORMANCE COMPARISON ON PASCAL VOC (30% LABEL RATIO)

Method	mAP (%)	Label Ratio	F1-Score
WSML [14]	84.2	30%	0.81
G2NetPL (Proposed)	**88.5**	30%	**0.85**
CDUL (Proposed, Unsup.)	**65.1**	0%	**0.63**

D. Ablation Study

Table II presents component-level ablation results on MS-COCO at the 50% partial-label ratio. Removing the momentum term from PLMCL degrades mAP by 2.1 points, confirming that momentum-driven pseudo-label updates provide essential stability. Disabling the velocity damping term causes a further 1.2-point drop, isolating its independent contribution. For G2NetPL, setting $\eta = 0$ (eliminating the KL game penalty) reduces mAP by 1.8 points, validating the importance of adversarial pseudo-label sharpening. Removing the confidence scheduler incurs a 1.2-point loss. Within CDUL, discarding local patch similarities reduces mAP by 3.2 points, and removing the aggregator causes a 1.8-point drop, confirming that global-local fusion is the most critical design element.

TABLE II ABLATION STUDY ON MS-COCO (50% PARTIAL LABEL RATIO)

Model Variant	mAP (%)	F1-Score
PLMCL (Full Model)	82.3	0.80
PLMCL w/o Momentum	80.2	0.78
PLMCL w/o Velocity Term	81.1	0.79
G2NetPL (Full Model)	83.2	0.81
G2NetPL w/o Game Penalty (η=0)	81.4	0.79
G2NetPL w/o Conf. Scheduler	82.0	0.80
CDUL (Full Model)	72.3	0.70
CDUL w/o Local Similarity	69.1	0.68
CDUL w/o Aggregator	70.5	0.69

V. DISCUSSION

While all three frameworks demonstrate robust performance under annotation scarcity, limitations remain. PLMCL's empirically stable momentum updates lack formal convergence guarantees and may oscillate on heavily imbalanced datasets. G2NetPL assumes a well-defined non-cooperative equilibrium, whereas real-world label noise can disturb minimax dynamics, potentially necessitating adaptive penalty schedules. CDUL's dependence on CLIP's pretraining distribution may introduce biases toward frequently occurring Internet concepts, causing underperformance on domain-specific categories without fine-tuning.

Comparison with the state of the art confirms the gains reported in Tables I–IV: PLMCL surpasses PML [6] by integrating unlabeled samples through momentum scheduling; G2NetPL improves over adversarial PLL [8] through explicit pseudo-label game objectives; and CDUL matches WSML [14] despite requiring no human annotations, highlighting the practical power of large-scale vision-language pretraining. Promising future directions include deriving theoretical guarantees for pseudo-label optimality, scaling to million-image datasets via distributed training, and hybridizing game-theoretic constraints with CLIP-driven initialization for enhanced zero-shot adaptation.

VI. CONCLUSION

This paper introduced three complementary deep learning frameworks for multi-label image classification under annotation-constrained conditions. PLMCL advances partial-label learning by embedding velocity-aware momentum scheduling into pseudo-label updates, effectively utilizing unlabeled training samples. G2NetPL establishes a game-theoretic paradigm for stable pseudo-label convergence, addressing optimization pitfalls inherent in standard PLL methods. CDUL pioneers annotation-free classification by coupling CLIP-based global-local similarity aggregation with iterative self-refinement, rivaling weakly supervised approaches without any labeled data. Experimental validation across MS-COCO, PASCAL VOC, and NUS-WIDE consistently demonstrates superior mAP, F1-score, and Hamming loss, underscoring the capacity of these frameworks to democratize multi-label recognition systems by minimizing annotation burdens. These contributions lay a foundation for scalable, efficient solutions to real-world vision tasks with broad implications for weakly supervised learning research.

REFERENCES

[1]. R. Abdelfattah, X. Zhang, Z. Wu, X. Wu, X. Wang, and S. Wang, "PLMCL: Partial-Label Momentum Curriculum Learning for Multi-Label Image Classification," arXiv (Cornell University), 2022. https://doi.org/10.48550/arxiv.2208.09999

[2]. R. Abdelfattah, Q. Guo, X. Li, X. Wang, and S. Wang, "CDUL: CLIP-Driven Unsupervised Learning for Multi-Label Image Classification," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023, pp. 1348–1357. https://doi.org/10.1109/iccv51070.2023.00130

[3]. R. Abdelfattah, Q. Guo, X. Li, X. Wang, and S. Wang, "CDUL: CLIP-Driven Unsupervised Learning for Multi-Label Image Classification," arXiv (Cornell University), 2023. https://doi.org/10.48550/arxiv.2307.16634

[4]. Y. Chen, F. Li, N. Han, G. Li, H. Gao, S. Chan, and X. Fang, "Pseudo-Label Reconstruction for Partial Multi-Label Learning," in Proc. 33rd Int. Joint Conf. Artificial Intelligence (IJCAI), 2024, pp. 4896–4904. https://doi.org/10.24963/ijcai.2024/545

[5]. K. Duarte, Y. Rawat, and M. Shah, "PLM: Partial Label Masking for Imbalanced Multi-label Classification," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition Workshops (CVPRW), 2021, pp. 2733–2742.

[6]. T. Durand, N. Mehrasa, and G. Mori, "Learning a Deep ConvNet for Multi-Label Classification with Partial Labels," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2019, pp. 647–657.

[7]. F. Sun, M.-K. Xie, and S.-J. Huang, "A Deep Model for Partial Multi-label Image Classification with Curriculum-based Disambiguation," Machine Intelligence Research, vol. 21, no. 4, pp. 801–814, 2024.

[8]. Y. Yan and Y. Guo, "Adversarial Partial Multi-Label Learning with Label Disambiguation," in Proc. AAAI Conf. Artificial Intelligence, vol. 35, no. 12, pp. 10568–10576, 2021.

[9]. H. Çevıkalp, B. Benligiıray, and Ö. N. Gerek, "Semi-Supervised Robust Deep Neural Networks for Multi-Label Image Classification," Pattern Recognition, vol. 100, p. 107164, 2019.

[10]. Z. Peng, D. Zhang, S. Tian, W. Wu, L. Yu, S. Zhou, and S. Huang, "FaxMatch: Multi-Curriculum Pseudo-Labeling for Semi-Supervised Medical Image Classification," Medical Physics, vol. 50, no. 5, pp. 3210–3222, 2023.

[11]. M. Han, H. Wu, Z. Chen, M. Li, and X. Zhang, "A Survey of Multi-Label Classification Based on Supervised and Semi-Supervised Learning," Int. J. Machine Learning and Cybernetics, vol. 14, no. 3, pp. 697–724, 2022.

[12]. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in Proc. European Conf. Computer Vision (ECCV), 2014, pp. 740–755.

[13]. N. Xu, Y.-P. Liu, and X. Geng, "Partial Multi-Label Learning with Label Distribution," in Proc. AAAI Conf. Artificial Intelligence, vol. 34, no. 04, pp. 6510–6517, 2020.

[14]. Y. Kim, J. M. Kim, Z. Akata, and J. Lee, "Large Loss Matters in Weakly Supervised Multi-Label Classification," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2022, pp. 14136–14145.

[15]. W. Zhang, C. Liu, L. Zeng, B. Ooi, S. Tang, and Y. Zhuang, "Learning in Imperfect Environment: Multi-Label Classification with Long-Tailed Distribution and Partial Labels," in Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV), 2023, pp. 1423–1432.