This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Prototype-based Causal Intervention for Multi-Label Image Classification

Yanmin Li Zhilong Mao Mao Wang Lihua Liu* Jibing Wu Weidong Bao
Laboratory for Big Data and Decision
National University of Defense Technology
{yanminli, mzl02, wangmao, lihualiu, wujibing, wdbao}@nudt.edu.cn

Abstract

Modern multi-label image classification models suffer from a critical reliance on spurious correlations, failing to learn the underlying causal mechanisms. Many causality-inspired methods are impractical, demanding box-level supervision that is rarely available in real-world datasets. Others rely on static confounder dictionaries, which are inherently inflexible and fail to capture complex biases or adapt to feature space changes during training. To address this, we present prototype-based causal intervention (ProCI), a novel framework that approximates the backdoor adjustment using only image-level supervision. It models confounders as learnable contextual prototypes engineered to represent class-wise co-occurring bias. These prototypes are learned dynamically within a stable memory and leveraged to construct sample-specific bias vectors for an adaptive feature adjustment, effectively counteracting spurious correlations. Experiments on MS-COCO, Pascal VOC, COCO-Stuff, and the challenging Sewer-ML dataset validate our approach. ProCI achieves competitive performance on standard benchmarks while setting a new state-of-the-art on the highly-confounded Sewer-ML. It outperforms the previous best model by a remarkable +5.44 points on the $F2_{CIW}$ metric. These results demonstrate the effectiveness of our approach in mitigating complex real-world biases using only image-level supervision[1].

1. Introduction

While modern multi-label image classification (MLC) models continually achieve state-of-the-art accuracy on standard benchmarks [33, 59], they are notoriously brittle and exhibit poor generalization when deployed in the wild. This critical generalization gap arises because models are optimized to exploit dataset-specific spurious correlations [12, 49], such as transient label co-occurrences, rather than learning the underlying causal mechanisms that govern the visual world. Consequently, model performance often collapses when these learned "shortcuts", which are mere artifacts of dataset biases, do not hold in real-world environments.

Recently, various strategies have been explored, ranging from specialized data augmentation and re-sampling techniques to learning disentangled representations [20, 35]. A principled approach, and one gaining significant traction, leverages causal inference to mitigate the influence of confounders, aiming to learn true causal effects rather than mere statistical associations [7, 24, 29, 40, 44]. However, existing causality-inspired approaches often face two practical limitations. Firstly, their causal interventions highly depend on instance-level bounding boxes [40], which are often unavailable in many real-world datasets, especially in large-scale industrial image classification where data is labeled only at the image level. Secondly, these methods depend on pre-defined, static dictionaries to model potential confounders [37], an approach with limited representative capacity, unable to capture complex and unforeseen biases. By being pre-defined and fixed, this dictionary is inherently "over-fitted" to the specific statistical biases of its construction set, thus lacking the flexibility to adapt to the diverse confounding patterns encountered in real-world scenarios.

To address these gaps, we propose prototype-based causal intervention (ProCI), a novel framework that approximates the backdoor adjustment [28] in feature space using only image-level labels. ProCI uniquely models confounding influences as learnable *contextual prototypes*. Unlike traditional prototypes that capture discriminative class features, our prototypes are designed to represent class-wise co-occurring biases. The framework is composed of two synergistic modules designed to precisely overcome the limitations of prior work: the dynamic confounder memory (DCM) and the adaptive causal intervention (ACI) module.

The DCM module directly confronts the "static dictionary" problem. Instead of relying on pre-defined confounders, it maintains a learnable memory of class-wise contextual prototypes that model latent confounding influences present in multi-label images. To balance adaptabil-

*Corresponding author.
[1]Code is available at https://github.com/JustinLiam/ProCI.

ity and stability, we use a two-fold update strategy. First, it selectively identifies and aggregates the most informative contextual features from the current batch, capturing immediate, salient patterns. Second, this new information is then smoothly integrated into the memory. This allows the DCM to learn and adapt to the complex distributions of confounders, using only image-level supervision provided by class activation map (CAM)-guided feature extraction.

The ACI module then performs the causal intervention, isolating true causal semantics (e.g., the object) from the confounding influences (e.g., its co-occurrence) identified by the DCM. To achieve this, the ACI module first adaptively queries the confounder memory to construct a sample-specific bias vector. Critically, this vector is then used to perform a conceptual feature adjustment, effectively counteracting the confounding component in the original image representation. This de-biased representation then produces the final causal logits, yielding robust predictions that are less reliant on spurious correlations.

We conduct experiments on standard benchmarks and the challenging Sewer-ML dataset [14], a large-scale industrial dataset with severe confounding. Our contributions are threefold: (1) We propose ProCI, a causal intervention framework that operates solely on image-level supervision, eliminating the reliance on instance-level annotations. (2) We introduce the DCM module, which maintains a trainable, continuously updated confounder memory for modeling confounding influences. (3) ProCI achieves competitive results on standard benchmarks and establishes a new state-of-the-art on Sewer-ML, demonstrating its superior ability to mitigate complex, real-world confounding biases.

2. Related work

Spurious correlations in multi-label image classification. The challenge of spurious correlations in MLC has drawn significant attention [7, 26, 33]. Prior to causal approaches, mitigation efforts broadly fall into two categories. The first focuses on data-level manipulation, such as re-sampling or advanced augmentation strategies [35, 38]. The second attempts model-level fixes, such as dedicated representation or explicit regularization to penalize reliance on contextual features [31, 54]. However, these approaches merely "patch" the symptoms of bias rather than modeling the underlying causal mechanism itself. This fundamental limitation motivates the need for a more principled approach, leading to the rise of causal inference for de-confounding.

Causal inference for visual de-biasing. Causal inference, particularly Pearl's backdoor adjustment [28], provides a principled framework for de-biasing. Since the full confounder set is unobservable, the dominant approximation strategy is to model it as a single, static confounder dictionary [7, 23, 37, 40]. This dictionary is computed once by averaging box-level features [37, 40] or clustering global features [7], and then remains fixed during training. Intervention is then performed by marginalizing over this fixed set, often via attention mechanisms [7, 19, 40] or counterfactual subtraction [23]. However, this paradigm is inherently limited because the fixed dictionary cannot adapt to the feature space changes during the training process. While this paradigm also suffers from scalability issues (e.g., reliance on bounding boxes), the core unaddressed challenge is designing a framework that can learn confounder representations *dynamically* from *image-level supervision*.

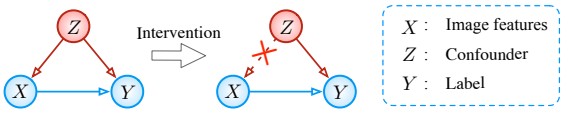


Figure 1. Causal graph in multi-label image classification. The causal intervention blocks the back-door path from Z to X, aiming to learn true causal effects rather than statistical associations.

Prototype-based bias representation. The concept of prototype learning is well-established, notably in few-shot classification, where prototypes are learned to represent the discriminative essence of a target category [10, 36]. Because prototypes inherently capture stable and interpretable representations [53], they provide a natural vehicle for encoding contextual regularities that act as confounding factors. Therefore, we re-purpose this mechanism to model the effect of the unobservable confounder—the contextual bias it induces in the feature space. We argue that in the presence of co-occurring biases, CAMs become "polluted" signals, inevitably mixing target features with contextual cues. Based on this insight, we introduce *class-wise contextual prototypes* as a learnable memory, and propose to dynamically distill and update these prototypes from the CAMs. This reframes prototype learning from representing target semantics to explicitly capturing class-wise contextual biases. Extended comparisons are provided in Appendix A.

3. Our method

3.1. Preliminaries

In the MLC, a standard deep model learns the observational conditional probability $P(Y|X)$, where X represents the image features and Y denotes the label vector. However, this probability merely captures the statistical association, which is often spurious as it may not arise from the intrinsic properties of the objects themselves. Such spurious correlations primarily stem from confounders, where a common "context" causes both the features and the labels to appear (as illustrated in Fig. 1). To learn the true causal effect of X on Y, we need to answer a counterfactual question: "What would the probability of label Y be if we could enforce the features X, imagining it appearing across all possible

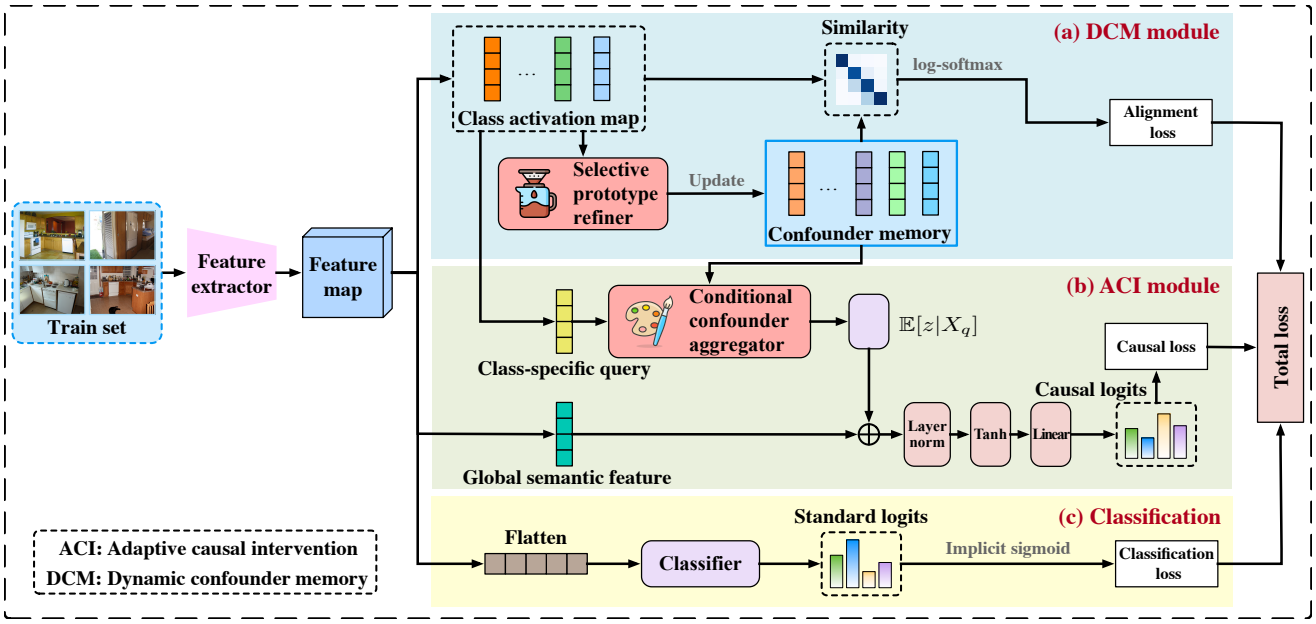


Figure 2. Overview of the ProCI framework. (a) **DCM Module:** Tasked with building the confounder representation, this module first employs CAM to capture per-class contextual features. A *selective prototype refiner* then filters these features, balancing selectivity (to weight high-quality samples) with stability (to ensure long-term consistency), thereby continuously updating the *confounder memory*, which also enforces feature-memory alignment. (b) **ACI Module:** This module approximates the causal intervention. It first disentangles the image into a global semantic feature and a class-specific query. A *conditional confounder aggregator* then uses the query to compute the conditional expected confounder representation. This debiased representation is fused with the global semantic feature to produce the final causal logits. (c) **Standard classification:** This module produces standard observational logits, which act as a complementary signal.

confounders Z?" This concept is formalized in causal inference as an intervention, denoted as $P(Y|\text{do}(X))$, where the $\text{do}(\cdot)$ represents an external manipulation. According to Pearl's backdoor criterion [28], the true causal effect can be identified and computed:

$$P(Y|\text{do}(X)) = \sum_z P(Y|X,z)P(z) \tag{1}$$

However, directly computing the causal effect via Eq. (1) is intractable. The marginalization requires enumerating all possible confounders Z, which are not only unobserved, high-dimensional, and latent, but also represent a complex, unknown distribution of visual contexts. Therefore, the central challenge is to find a practical approximation for this intervention, $P(Y|\text{do}(X))$, using only image-level labels.

3.2. ProCI overview

To address the challenges discussed in Sec. 3.1, we propose the prototype-based causal intervention (ProCI) framework. Our goal is to approximate the intractable causal effect $P(Y|\text{do}(X))$, as defined in Eq. (1), by explicitly modeling and mitigating the influence of latent confounders Z. As illustrated in Fig. 2, ProCI achieves this through two cooperative causal modules:

(a) The DCM module, which dynamically maintains a confounder memory. This memory stores prototypes that represent the expected contextual bias induced by Z.

(b) The ACI module, which performs the backdoor adjustment. It leverages the confounder memory to construct a sample-specific bias representation, then uses it to "debias" the image feature, yielding the final causal logits.

In parallel, a standard classification learns the observational probability $P(Y|X)$. We retain this as a complementary signal to enhance the training stability. The specific designs for the DCM and ACI modules are detailed in Sec. 3.3 and Sec. 3.4, followed by the training objective in Sec. 3.5.

3.3. Dynamic confounder memory module

The goal of the DCM module is not to explicitly model the unobservable confounders Z, but to approximate their *observable effect*—the contextual patterns that consistently co-occur with a given class. To this end, we maintain a learnable confounder memory $M \in \mathbb{R}^{K\times D}$, where each entry M_k stores a *class-conditioned contextual prototype*. Rather than assuming that confounders are class-specific, this formulation captures the empirical contextual regularities associated with class k under the training distribution. These prototypes serve as adaptive estimates of class-wise contextual bias and provide the bias representations

required by the ACI module for feature-level intervention.

Contextual feature extraction. To capture class-conditioned contextual signals, we generate per-class features $Z \in \mathbb{R}^{B\times K\times D}$ via CAM-based weighted pooling [5, 56]. Given the feature map F and the CAM $A \in \mathbb{R}^{B\times K\times H\times W}$, we compute:

$$Z_{b,k} = \sum_{h,w} \frac{\text{ReLU}(A[b,k,h,w])}{\sum_{h',w'} \text{ReLU}(A[b,k,h',w'])} \cdot F_b[:,h,w] \quad (2)$$

The resulting $Z_{b,k}$ captures the dominant co-occurring patterns for class k. We treat these CAM-conditioned features as observable contextual evidence, serving as the operational signal for memory construction and intervention.

Selective prototype refiner. Because the CAM-conditioned features may exhibit noise or sample-level variability, M requires a dynamic mechanism that can emphasize reliable contextual evidence while suppressing outliers. Simply averaging the batch features $Z_{:,k}$ is therefore suboptimal. We introduce the selective prototype refiner, which updates prototypes through two complementary processes:

Intra-batch selectivity. Not all features $Z_{b,k}$ (positives $\mathcal{B}_k$) contribute equally to contextual characterization. To prioritize the most representative samples, we compute an attention score using a lightweight, class-specific contextual preference vector $w_k \in \mathbb{R}^D$:

$$s_{b,k} = w_k^T Z_{b,k}. \quad (3)$$

Softmax normalization produces attention weights $a_{b,k}$:

$$\{a_{b,k}\}_{b\in\mathcal{B}_k} = \text{Softmax}(\{s_{b,k}\}_{b\in\mathcal{B}_k}), \quad (4)$$

leading to the refined batch prototype

$$\hat{m}_k = \sum_{b\in\mathcal{B}_k} a_{b,k} Z_{b,k}. \quad (5)$$

Inter-batch stability. To mitigate batch-level fluctuations, the memory prototype M_k is updated using an EMA with momentum $\alpha \in [0,1)$:

$$M_k \leftarrow \alpha M_{k-1} + (1-\alpha)\hat{m}_k. \quad (6)$$

After each update, prototypes are L2-normalized. This EMA-based update maintains a long-term memory of contextual patterns accumulated across training batches, which mitigates the risk of overwriting earlier confounding patterns and stabilizes the training dynamics of the DCM. Through this process, stable class-conditioned contextual prototypes are distilled from noisy CAM features. Importantly, DCM remains robust regardless of CAM localization quality; see Appendix B for a detailed analysis.

Alignment regularization To ensure that the confounder memory M receives reliable contextual evidence, the extracted features Z must avoid excessive dispersion caused by CAM noise. We therefore introduce a lightweight alignment loss [16] to regularize the feature space. This loss encourages each contextual feature $Z_{b,k}$ to remain close to its corresponding prototype M_k while remaining distinguishable from other prototypes M_j. Using L2-normalized features $\tilde{Z}_{b,k}$ and prototypes $\tilde{M}_j$, the loss is:

$$L_{ali} = -\frac{1}{|\mathcal{P}|} \sum_{(b,k)\in\mathcal{P}} \log \frac{\exp(\tilde{Z}_{b,k}\cdot\tilde{M}_k/\tau)}{\sum_{j=1}^{K}\exp(\tilde{Z}_{b,k}\cdot\tilde{M}_j/\tau)} \quad (7)$$

where $\mathcal{P}$ is the set of positive pairs (b,k) in the batch (i.e., $Y_{b,k}=1$), and τ is a temperature hyper-parameter. This regularization helps stabilize memory updates by reducing feature-level noise.

3.4. Adaptive causal intervention module

The ACI module implements a feature-level approximation of the backdoor adjustment in Eq. (1). It first factorizes the image representation into a global semantic feature X_c and a class-specific query X_q. Using X_q, the DCM provides a confounder-aware embedding C that captures the contextual patterns associated with each class. ACI then combines X_c and C through a lightweight modulation function to produce the causal logits. See Appendix C for a formal connection to the backdoor adjustment.

Feature disentanglement. A valid causal adjustment requires separating the contributions of X (content) and Z (confounder). We obtain them by the following:

- *Global semantic feature (X_c):* Represents context-agnostic content. We compute $X_c \in \mathbb{R}^{B\times D}$ via global average pooling on the final backbone feature map F, followed by L2-normalization.
- *Class-specific query (X_q):* Represents localized, class-conditioned evidence used to infer Z. We use the contextualized features $Z \in \mathbb{R}^{B\times K\times D}$ as the query representation, where K is the number of classes.

This separation is crucial: X_c is derived from global features, whereas X_q originates from class-conditioned features that tend to capture co-occurring contexts.

Adaptive confounder estimation. To perform the adjustment, we estimate the confounder z conditioned on the query X_q. We approximate this conditional expectation by treating the normalized memory prototypes $\tilde{M}_j$ as discrete prototype candidates. For each vector $Z_{b,k}$, we compute:

$$(q_{b,k}, k_j) = (\text{Norm}(W_q Z_{b,k}),\ \text{Norm}(W_k \tilde{M}_j)) \quad (8)$$

$$\beta_{b,k,j} = \text{Softmax}_j\left((q_{b,k}\cdot k_j^T)/\tau_{aci}\right) \quad (9)$$

$$C_{b,k} = \sum_{j=1}^{K} \beta_{b,k,j}\tilde{M}_j \approx \mathbb{E}[z \mid Z_{b,k}] \quad (10)$$

where W_q and W_k are linear projection layers and τ_{aci} is a temperature hyperparameter. The resulting tensor $C \in \mathbb{R}^{B\times K\times D}$ provides the estimated bias representation for each sample–class pair.

Causal logit computation. Instead of attempting to "subtract" the bias, we compute causal logits by learning a confounder-aware modulation function g. We posit that a classifier $g(\cdot)$ trained on the disentangled X_c and C is encouraged to rely primarily on the causal path (via X_c) while accounting for the spurious path (via C).

To compute the K logits for each sample b, we first expand the global feature tensor X_c (shape $B \times D$) to match the per-class bias tensor C (shape $B\times K\times D$). We then concatenate these tensors and pass the result through an MLP (LayerNorm, Tanh, Linear) to produce the final logits:

$$h = \text{Tanh}(\text{LayerNorm}([X_{c,\text{expanded}} \oplus C])) \tag{11}$$

$$\hat{Y}^{\text{cau}} = W_{\text{cau}} h + b_{\text{cau}} \tag{12}$$

The final output $\hat{Y}^{\text{cau}} \in \mathbb{R}^{B\times K}$ is the output of this modulation function $g(X_c, C)$, providing a feature-level approximation to the intervened prediction $P(Y|\text{do}(X))$.

3.5. Training objective

We jointly optimize the framework with a total loss L_{total} composed of three terms that reflect our model design:

$$L_{total} = L_{cau} + \lambda_{ali} L_{ali} + \lambda_{cls} L_{cls}. \tag{13}$$

Here, L_{cau} and L_{cls} are the classification losses computed from the causal logits $\hat{Y}^{\text{cau}}$ and the standard logits $\hat{Y}^{cls}$, respectively, both implemented using the asymmetric loss (ASL) [31]. L_{ali} is the representation alignment loss from Eq. (7). The coefficients λ_{ali} and λ_{cls} are weighting factors that balance the alignment and auxiliary classification terms, supporting the end-to-end optimization.

4. Experiments

We evaluate ProCI across standard multi-label benchmarks and a large-scale industrial dataset with severe real-world confounding. We first compare against state-of-the-art methods on the standard benchmarks, then assess performance under real-world confounding, supported by qualitative heatmap visualizations. Finally, we conduct ablations to isolate the contribution of each component.

4.1. Experimental setup

Datasets. We evaluate on four multi-label benchmarks: **MS-COCO** [22], **Pascal VOC** [9], **COCO-Stuff** [1] and **Sewer-ML** [14]. While MS-COCO and Pascal VOC are standard benchmarks for general performance , COCO-Stuff is employed to assess model robustness under explicit co-occurrence shifts. **Sewer-ML** is a large-scale industrial dataset containing over 1.3 million images with severe real-world confounding. This comprehensive setting highlights the necessity of causal de-biasing under weak supervision. Detailed dataset statistics are provided in Appendix D.

Implementation details. We train all models using AdamW with backbone-specific learning rates and weight decay following prior work, under a cosine decay schedule and mixed-precision training. The loss follows Eq. (13) with $\lambda_{cls} = \lambda_{ali} = 1.0$, using ASL for classification. The alignment loss L_{ali} and causal loss L_{cau} are activated after 10 epochs to allow the confounder memory to stabilize. All models are trained for 50 epochs. Baseline results are taken from original publications or reproduced using official implementations when backbone settings match. Detailed hyperparameters are provided in Appendix E.

Evaluation metrics. For **MS-COCO**, we report mean average precision (mAP) together with per-class (CP, CR, CF1) and overall (OP, OR, OF1) precision, recall, and F1 scores. For **Pascal VOC 2007**, we report mAP and per-class average precision (AP). For **COCO-Stuff**, we report mAP across the *exclusive*, *co-occur*, and *all* distributions. For **Sewer-ML**, we report $\text{F1}_{\text{Normal}}$ and the class-importance weighted F2 score (F2_{CIW}). $\text{F1}_{\text{Normal}}$ corresponds to the dataset's "normal" (non-defect) class, while F2_{CIW} is an importance-weighted F2 measure emphasizing defect recall. Detailed formulations are provided in Appendix F.

4.2. Results on standard benchmarks

MS-COCO Tab. 1 reports comparisons with state-of-the-art methods on MS-COCO. ProCI delivers consistent improvements across all backbone families.

CNN backbones. On standard architectures, ProCI reaches 87.0% mAP with ResNet-101 and 90.1% with TResNet-L, exceeding all prior methods using the same backbone settings. Moreover, the gains extend beyond mAP: ProCI improves CF1 and OF1, indicating that our causal intervention enhances both per-class and overall label consistency.

Transformer backbones. The benefits of ProCI become more pronounced with high-capacity models. Using a Swin-Base pretrained only on ImageNet-1k, ProCI achieves 90.7% mAP, surpassing the previously reported Q2L-Swin-L (90.5%) despite using a significantly smaller backbone and a weaker pretraining source. This demonstrates strong parameter efficiency and robustness to pretraining scale. When applied to the same Swin-L (21k) backbone, ProCI further establishes a new state-of-the-art with 91.4% mAP, along with new best CF1 (87.1%) and OF1 (89.0%).

Key observations. These results highlight two important findings: (i) ProCI consistently improves strong baselines without modifying backbones, validating the effectiveness of our confounder modeling strategy; (ii) the gains amplify with transformer models, supporting our design moti-

Table 1. Comparison results on MS-COCO with state-of-the-art methods. Backbones are pretrained on ImageNet-1k unless otherwise specified. "21k", "22k", and "OI" denote ImageNet-21k, ImageNet-22k, and Open Images pretraining, respectively. Input resolution follows the settings used in the corresponding original works. Best results are highlighted in **bold**. All metrics are in %.

Method	Backbone	Resolution	mAP	CP	CR	CF1	OP	OR	OF1
ML-GCN [6]	ResNet-101	448 × 448	83.0	85.1	72.0	78.0	85.8	75.4	80.3
CSRA [50]	ResNet-101	448 × 448	83.5	84.1	72.5	77.9	85.6	75.7	80.3
MS-CMA [51]	ResNet-101	448 × 448	83.8	82.9	74.4	78.4	84.4	77.9	81.0
MCAR [11]	ResNet-101	448 × 448	83.8	85.0	72.1	78.0	88.0	73.9	80.3
KSSNet [41]	ResNet-101	448 × 448	83.7	84.6	73.2	77.2	87.8	76.2	81.5
TDRG [55]	ResNet-101	448 × 448	84.6	86.0	73.1	79.0	86.6	76.4	81.2
IDA [24]	ResNet-101	448 × 448	84.8	-	-	78.7	-	-	80.9
SpliceMix-CL [38]	ResNet-101	448 × 448	84.9	87.4	73.2	79.7	88.2	76.3	81.7
PAT-T [44]	ResNet-101	448 × 448	85.0	87.6	74.4	80.5	88.4	77.9	82.8
C-Trans [21]	ResNet-101	576 × 576	85.1	86.3	74.3	79.9	87.7	76.5	81.7
C-TMS [43]	ResNet-101	448 × 448	85.1	87.2	74.2	79.1	88.7	76.5	81.4
MambaML [59]	ResNet-101	448 × 448	85.7	86.1	75.1	80.3	87.7	77.6	82.3
OPRL [54]	ResNet-101	448 × 448	85.7	85.9	75.3	79.8	87.1	77.9	82.3
CMLL [37]	ResNet-101	576 × 576	86.5	**88.8**	76.7	81.2	88.1	78.9	82.8
ProCI (Ours)	ResNet-101	448 × 448	86.5	88.2	76.5	81.9	88.4	78.5	83.2
ProCI (Ours)	ResNet-101	576 × 576	**87.0**	88.4	**77.0**	**82.3**	**88.9**	**79.1**	**83.7**
SPA [8]	ViT-B/32	224 × 224	76.8	-	-	-	-	-	-
ASL [31]	TResNet-L	448 × 448	86.6	87.2	76.4	81.4	88.2	79.2	81.8
CPSD [47]	TResNet-L	448 × 448	87.3	85.5	78.9	82.1	85.7	81.5	83.7
Q2L [25]	TResNet-L	448 × 448	87.3	87.6	76.5	81.6	88.4	78.5	83.1
OPRL [54]	ViT-L	448 × 448	88.1	89.6	76.6	81.7	89.3	80.0	84.4
PAT-T [44]	TResNet-L	448 × 448	88.7	89.0	78.8	83.6	89.3	81.3	85.1
ADDS [48]	ViT-L	224 × 224	89.8	-	-	-	-	-	-
CMLL [37]	TResNet-L	576 × 576	89.2	86.3	**81.4**	83.8	88.5	**83.3**	85.6
ProCI (Ours)	TResNet-L	448 × 448	**90.1**	**89.8**	79.2	**84.2**	**91.2**	81.5	**86.1**
TResNet-L [30]	TResNet-L (21k)	448 × 448	88.4	-	-	-	-	-	-
ML-Decoder [33]	TResNet-L (OI)	448 × 448	90.0	-	-	-	-	-	-
Q2L-TResL [25]	TResNet-L (OI)	448 × 448	90.1	83.6	**84.2**	83.9	85.0	**85.9**	85.4
OPRL [54]	ViT-L (22k)	448 × 448	90.3	92.5	78.5	84.3	92.7	80.2	86.0
Q2L-SwinL [25]	Swin-L (21k)	384 × 384	90.5	89.4	81.7	85.4	89.8	83.2	86.4
ProCI (Ours)	**Swin-Base (1k)**	384 × 384	90.7	92.2	81.5	86.5	92.7	84.6	88.5
ProCI (Ours)	Swin-L (21k)	384 × 384	**91.4**	**92.7**	82.2	**87.1**	**93.1**	85.2	**89.0**

Table 2. Comparison of class-wise average precision (AP) and mean average precision (mAP) on the Pascal VOC 2007 dataset. All metrics are in %. The symbol ⋆ denotes using a large input resolution (576×576). The best performance is highlighted in **bold**.

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP
RMIC [17]	97.1	91.3	94.2	57.1	86.7	90.7	93.1	63.3	83.3	76.4	92.8	94.4	91.6	95.1	92.3	59.7	86.0	69.5	96.4	79.0	84.5
RLSD [52]	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	90.0	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
RDAR [42]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	95.8	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
RARL [3]	98.6	97.1	97.1	95.5	75.6	92.8	97.3	97.3	78.3	92.2	87.6	97.6	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
RCP [39]	99.3	97.6	98.0	96.4	79.3	93.8	96.6	97.1	78.0	88.7	87.1	97.1	96.3	95.4	99.1	82.1	93.6	82.2	98.4	92.8	92.5
SSGRL⋆ [4]	99.5	97.1	97.6	97.8	82.4	94.8	96.7	98.1	78.0	97.0	85.6	97.8	97.4	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
ML-GCN [6]	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	97.5	93.7	98.4	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
TSGCN [46]	98.9	98.5	96.8	97.3	**87.5**	94.2	97.4	97.7	84.1	92.6	89.3	98.4	98.0	96.1	98.7	84.9	96.6	87.2	98.4	93.7	94.3
ASL [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	94.4
CSRA [58]	**99.9**	98.4	98.1	**98.9**	82.2	95.3	97.8	97.9	84.6	94.8	**90.8**	98.1	97.6	96.2	99.1	86.4	95.9	88.3	98.9	94.4	94.7
DATran [57]	**99.9**	98.7	98.6	98.4	82.6	96.0	97.7	98.6	85.0	96.2	84.7	97.9	96.7	96.9	98.8	85.0	97.9	86.7	99.2	**95.2**	94.6
C-TMS [43]	99.5	96.2	97.6	96.5	86.4	95.8	95.2	**98.9**	**88.7**	97.5	87.6	99.2	**99.2**	97.3	99.0	85.0	98.1	86.1	99.1	94.2	94.8
MambaML [59]	99.8	98.6	97.8	98.0	82.8	96.3	98.1	98.3	84.0	96.7	88.3	98.2	98.6	96.8	99.0	87.5	96.8	**89.8**	99.2	95.1	95.0
OPRL [54]	**99.9**	98.5	98.4	98.4	83.9	97.3	97.7	98.6	80.7	98.1	86.9	99.1	99.1	**98.2**	**99.3**	87.0	**99.6**	86.7	**99.7**	93.8	95.1
ProCI (Ours)	**99.9**	**99.0**	**98.8**	98.2	**87.5**	**98.4**	**98.5**	98.5	84.0	**98.2**	89.7	**99.3**	98.8	**98.2**	**99.3**	**88.2**	99.1	**89.8**	**99.7**	**95.2**	**96.0**

vation that reducing context-induced spurious correlations particularly benefits high-capacity feature extractors.

Pascal VOC We further evaluate ProCI on the Pascal VOC 2007 dataset. Following common practice [54, 57,

59], we employ a TResNet-L backbone pretrained on ImageNet-1k with a 448×448 input resolution to ensure a fair comparison with recent strong baselines.

ProCI sets a new state-of-the-art, achieving **96.0% mAP**, which is a clear improvement of **+0.9 points** over the previous best method, OPRL (95.1%). This gain is supported by strong class-wise performance: ProCI attains the best or tied-best AP on 14 out of 20 categories, including challenging classes such as *bus*, *car*, *plant*, *sofa*, and *tv*.

Notably, VOC 2007 is a highly saturated benchmark with sparse label co-occurrence (approximately 1.5 labels per image), implying limited contextual bias. The significant gain in this low-correlation regime suggests that our ProCI does more than suppress dataset-specific co-occurrences. It instead strengthens object-centric representations and prevents the model from overfitting to incidental context.

4.3. Performance under real-world confounding

To assess robustness under severe real-world confounding, we evaluate ProCI on the Sewer-ML dataset, which exhibits strong background–defect correlations induced by pipe materials, lighting conditions, and imaging artifacts [14]. Following prior work [18], we adopt a TResNet-L backbone with 224×224 input resolution to ensure a fair comparison. Most COCO/VOC-oriented MLC methods (e.g., ML-Decoder, OPRL and MambaML) are tailored to natural-image statistics, balanced label distributions, and large-scale pretraining, which do not hold in Sewer-ML. Consistent with established practice in this domain [18], we therefore compare against the strongest task-relevant baselines.

As shown in Tab. 3, ProCI establishes a new state-of-the-art by a clear margin, achieving **68.82%** $F2_{CIW}$ and **92.65%** $F1_{Normal}$. Compared to the strongest end-to-end non-causal baseline (TResNet-L), ProCI delivers a substantial gain of **+14.19 points** in $F2_{CIW}$. It also surpasses the previous leading model, SPM [18], by **+5.44 points** on this primary metric, highlighting the benefit of our dynamic confounder memory and adaptive intervention mechanism.

ProCI consistently outperforms all two-stage approaches, including the best-performing KSSNet [41]. These results underscore the effectiveness of ProCI in mitigating inherent dataset biases and enhancing model robustness in confounded real-world scenarios.

4.4. Deconfounding under co-occurrence shifts

Following [7] and [34], we evaluate ProCI on the COCO-Stuff dataset under the exclusive and co-occur settings, where label pairs that co-occur during training are decoupled at test time to simulate distribution shifts in co-occurrence patterns. Since the official implementations of Chen et al. [7] and feature-split [34] are not publicly available, we use reported results and run a publicly available reimplementation[2] using the same ResNet-101 backbone and protocol. As shown in Tab. 4, ProCI achieves substantial improvements over the prior state-of-the-art, with **+8.3 points** mAP in the exclusive setting and **+5.6 points** in the co-occur setting. These results indicate that ProCI effectively mitigates spurious correlations and generalizes under explicit confounder shifts.

Table 3. Comparison on Sewer-ML. The best performance is highlighted in **bold**.

	Method	Validation	
		$F2_{CIW}$ [%]	$F1_{Normal}$ [%]
Two-stage	Xie [45]	48.57	91.08
	Chen [2]	42.03	3.96
	Hassan* [13]	18.08	91.08
	Myrans* [27]	27.87	91.08
	ResNet-101* [15]	54.45	91.28
	KSSNet* [41]	55.37	91.30
	TResNet-L* [32]	55.36	91.32
End-to-end	Hassan [13]	13.14	0.00
	Myrans [27]	4.01	26.03
	ResNet-101 [15]	53.26	79.55
	KSSNet [41]	54.42	80.60
	TResNet-L [32]	54.63	81.22
	SPM [18]	63.38	91.57
	ProCI (Ours)	**68.82**	**92.65**

* denotes two-stage training protocol applied to an end-to-end architecture.

Table 4. Deconfounding performance (mAP) under co-occurrence shifts on COCO-Stuff. The best performance is highlighted in **bold**.

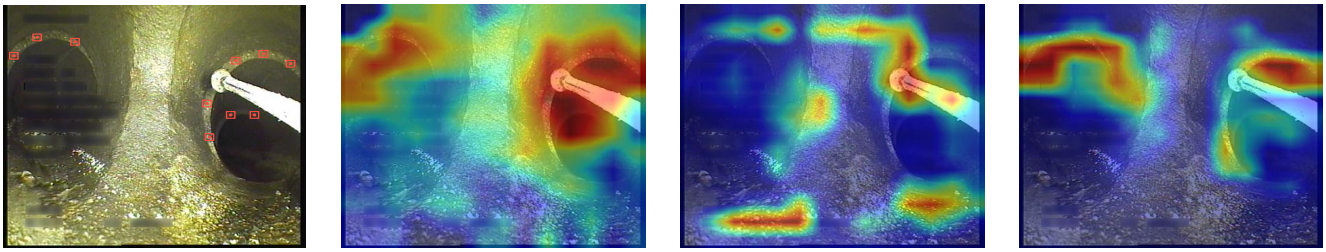
Method	Exclusive	Co-occur	All
ResNet-101 [15]	24.4	66.5	55.8
CAM-Based [34]†	25.6	63.5	55.9
feature-split [34]†	27.9	64.9	56.1
Chen et al. [7]‡	31.0	70.3	61.4
ProCI (Ours)	**39.3**	**75.9**	**64.5**

† Results obtained using a publicly available reimplementation[2].
‡ Results reported in the original paper.

4.5. Qualitative analysis

To better understand why ProCI yields large gains on Sewer-ML, we visualize model attention in Fig. 3. We compare the CAM of the *TResNet-L baseline* with two complementary heatmaps from our method: *Confounder-CAM*, obtained by projecting the learned DCM prototypes

[2]https://github.com/princetonvisualai/ContextualBias/tree/main



(a) Original image (b) Baseline CAM (c) Confounder-CAM (d) ProCI-CAM

Figure 3. Qualitative analysis on Sewer-ML. One of the true labels is "displaced joint". (a) Original image containing the defect (marked in red for visualization) and salient contextual confounders (e.g., the inspection probe). (b) Baseline CAM: the TResNet-L model incorrectly focuses its attention on the spurious probe rather than the defect region. (c) Confounder-CAM: visualizes contextual patterns captured by the DCM prototypes, which correspond to confounding signals rather than object evidence. (d) ProCI-CAM: suppresses the confounder response and concentrates on the joint region, suggesting that our model relies more on the defect region for prediction.

back to the image space to visualize modeled confounders, and *ProCI-CAM*, computed from the standard classification branch after applying our causal intervention.

As shown in Fig. 3, although one of the true labels is "displaced joint", the baseline CAM behaves like a shortcut learner: its strongest response is on the bright inspection probe, a salient but non-defect object. Our Confounder-CAM reveals that the DCM has successfully internalized this biased pattern, as its prototypes also activate on the probe region despite the probe never being labeled as a separate defect category. In contrast, the ProCI-CAM illustrates the effect of our causal intervention: the response on the probe is strongly suppressed, and the attention shifts towards the joint region where the actual defect is located.

This example highlights how ProCI disentangles confounders from true causal evidence: the confounder memory explicitly models spurious context, and the ACI module downweights its influence when forming the final prediction. More qualitative results are provided in Appendix G.

4.6. Ablation study

We evaluate the components of ProCI on Sewer-ML in Table 5 where our full model achieves **68.82%** $F2_{CIW}$. Notably, training without the classification loss L_{cls} results in a performance of 53.04% which is inferior to the baseline. This validates that L_{cls} provides the foundational semantic anchors required for stable confounder discovery under image-level supervision. The causal intervention module provides the most significant individual contribution because its removal leads to a substantial performance drop of **9.45 points** in $F2_{CIW}$.

We further analyze the internal mechanisms of our ProCI. Replacing the attention-based selectivity with mean pooling yields a clear degradation in accuracy. Similarly, replacing the prototype stabilization with hard updates reduces the performance. These results demonstrate that each component contributes to the final representation quality. Specifically, the dynamic confounder memory provides reliable representations while the intervention mechanism performs effective feature-level deconfounding.

Table 5. Ablation of components and mechanisms on Sewer-ML. (I) Loss term combinations; (II) Structural design variants.

Variant	L_{cls}	L_{ali}	L_{cau}	$F2_{CIW}$	$F1_{Normal}$
Baseline	✓			54.63	81.22
w/o L_{cls}		✓	✓	53.04	77.43
w/o ACI	✓	✓		59.37	85.96
w/o L_{ali}	✓		✓	65.92	91.63
ProCI (Full)	✓	✓	✓	**68.82**	**92.65**
– Mean pooling	✓	✓	✓	64.38	89.70
– Hard update	✓	✓	✓	67.49	90.03

Due to space constraints, we provide extended scalability analyses regarding label density and category abstractions in Appendix H.

5. Conclusion and discussion

We introduced ProCI, a causal intervention framework for multi-label image classification that mitigates spurious correlations using only image-level labels. By modeling confounders as dynamic, learnable prototypes through the DCM module and performing adaptive feature-level adjustment via the ACI module, ProCI mitigates confounding influences in an end-to-end manner.

Although experiments validate the scalability of ProCI, scaling to large-vocabulary datasets with many co-occurring labels requires further investigation. Representing abstract categories such as travel, street, or sunset without salient anchors also remains a challenge when activation maps become diffuse. Developing more expressive confounder representations to address these bottlenecks remains a vital direction for future research.

References

[1] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018. 5

[2] Kefan Chen, Hong Hu, Chaozhan Chen, Long Chen, and Caiying He. An intelligent sewer defect detection method based on convolutional neural network. In *IEEE ICIA*, pages 1301–1306, 2018. 7

[3] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. Recurrent attentional reinforcement learning for multi-label image recognition. In *AAAI*, 2018. 6

[4] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *ICCV*, pages 522–531, 2019. 6

[5] Zhaozheng Chen and Qianru Sun. Extracting class activation maps from non-discriminative features as well. In *CVPR*, pages 3135–3144, 2023. 4

[6] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *CVPR*, pages 5177–5186, 2019. 6

[7] Zhao-Min Chen, Xin Jin, Yisu Ge, and Sixian Chan. In pursuit of causal label correlations for multi-label image recognition. In *NeurIPS*, pages 51634–51654, 2024. 1, 2, 7

[8] Jiacheng Cheng and Nuno Vasconcelos. Towards calibrated multi-label deep neural networks. In *CVPR*, pages 27589–27599, 2024. 6

[9] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 111(1):98–136, 2015. 5

[10] Meijun Fu, Xiaomin Wang, Jun Wang, and Zhang Yi. Prototype bayesian meta-learning for few-shot image classification. *IEEE TNNLS*, 36(4):7010–7024, 2025. 2

[11] Bin-Bin Gao and Hong-Yu Zhou. Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE TIP*, 30:5920–5932, 2021. 6

[12] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020. 1

[13] Syed Ibrahim Hassan, L Minh Dang, Irfan Mehmood, Suhyeon Im, Changho Choi, Jaemo Kang, Young-Soo Park, and Hyeonjoon Moon. Underground sewer pipe condition assessment based on convolutional neural networks. *Autom. Construct.*, 106:102849, 2019. 7

[14] Joakim Bruslund Haurum and Thomas B Moeslund. Sewer-ml: A multi-label sewer defect classification dataset and benchmark. In *CVPR*, pages 13456–13467, 2021. 2, 5, 7

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[16] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, pages 9729–9738, 2020. 4

[17] Shiyi He, Chang Xu, Tianyu Guo, Chao Xu, and Dacheng Tao. Reinforced multi-label image classification by exploring curriculum. In *AAAI*, 2018. 6

[18] Chuanfei Hu, Bo Dong, Hang Shao, Jiapeng Zhang, and Yongxiong Wang. Toward purifying defect feature for multi-label sewer defect classification. *IEEE TIM*, 72:1–11, 2023. 7

[19] Changqin Huang, Jili Chen, Qionghao Huang, Shijin Wang, Yaxin Tu, and Xiaodi Huang. Atcaf: Attention-based causality-aware fusion network for multimodal sentiment analysis. *Information Fusion*, 114:102725, 2025. 2

[20] Jie Huang, Zhao-Min Chen, Zhao-Min Chen, Xiaoqin Zhang, Xiaoqin Zhang, Yisu Ge, Yisu Ge, Lusi Ye, Lusi Ye, Guodao Zhang, Guodao Zhang, Huiling Chen, and Huiling Chen. Label decoupling and reconstruction: A two-stage training framework for long-tailed multi-label medical image recognition. In *ACM MM*, page 2861–2869, 2024. 1

[21] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *CVPR*, pages 16478–16488, 2021. 6

[22] 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 *ECCV*, pages 740–755, 2014. 5

[23] Ruyang Liu, Hao Liu, Ge Li, Haodi Hou, TingHao Yu, and Tao Yang. Contextual debiasing for visual recognition with causal mechanisms. In *CVPR*, pages 12755–12765, 2022. 2

[24] Ruyang Liu, Jingjia Huang, Thomas H. Li, and Ge Li. Causality compensated attention for contextual biased visual recognition. In *ICLR*, 2023. 1, 6

[25] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021. 6

[26] Lei Meng, Xiangxian Li, Xiaoshuo Yan, Haokai Ma, Zhuang Qi, Wei Wu, and Xiangxu Meng. Causal inference over visual-semantic-aligned graph for image classification. In *AAAI*, pages 19449–19457, 2025. 2

[27] Joshua Myrans, Richard Everson, and Zoran Kapelan. Automated detection of fault types in cctv sewer surveys. *J. Hydroinformatics*, 21(1):153–163, 2019. 7

[28] Judea Pearl. *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009. 1, 2, 3

[29] Zhuang Qi, Sijin Zhou, Lei Meng, Han Hu, Han Yu, and Xiangxu Meng. Federated deconfounding and debiasing learning for out-of-distribution generalization. In *IJCAI*, pages 6084–6092, 2025. 1

[30] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. *arXiv preprint arXiv:2104.10972*, 2021. 6

[31] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *ICCV*, pages 82–91, 2021. 2, 5, 6

[32] Tal Ridnik, Hussam Lawen, Asaf Noy, Emanuel Ben Baruch, Gilad Sharir, and Itamar Friedman. Tresnet: High performance gpu-dedicated architecture. In *WACV*, pages 1400–1409, 2021. 7

[33] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. Ml-decoder: Scalable and versatile classification head. In *WACV*, pages 32–41, 2023. 1, 2, 6

[34] Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *CVPR*, pages 11070–11078, 2020. 7

[35] Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Toward robustness in multi-label classification: A data augmentation strategy against imbalance and noise. In *AAAI*, pages 21592–21601, 2024. 1, 2

[36] Jing Tian, Runze Wan, Zhaokui Li, Yan Wang, Zhuoqun Fang, Jiaxu Guo, and Mingtai Qi. A multimodal prototype correction and multidimensional knowledge distillation framework for cross-domain few-shot hyperspectral image classification. *IEEE TGRS*, 63:1–18, 2025. 2

[37] Yingjie Tian, Kunlong Bai, Xiaotong Yu, and Siyu Zhu. Causal multi-label learning for image classification. *Neural Networks*, 167:626–637, 2023. 1, 2, 6

[38] Lei Wang, Yibing Zhan, Leilei Ma, Dapeng Tao, Liang Ding, and Chen Gong. Splicemix: A cross-scale and semantic blending augmentation strategy for multi-label image classification. *IEEE TMM*, 2025. 2, 6

[39] Meng Wang, Changzhi Luo, Richang Hong, Jinhui Tang, and Jiashi Feng. Beyond object proposals: Random crop pooling for multi-label image recognition. *IEEE TIP*, 25(12):5678–5688, 2016. 6

[40] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, pages 10760–10770, 2020. 1, 2

[41] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *AAAI*, pages 12265–12272, 2020. 6, 7

[42] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *ICCV*, pages 464–472, 2017. 6

[43] Yanan Wu, Songhe Feng, Gongpei Zhao, and Yi Jin. Transformer driven matching selection mechanism for multi-label image classification. *IEEE TCSVT*, 34(2):924–937, 2023. 6

[44] Ming-Kun Xie, Jia-Hao Xiao, Pei Peng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. Counterfactual reasoning for multi-label image classification via patching-based training. In *ICML*, pages 54576–54589, 2024. 1, 6

[45] Qian Xie, Dawei Li, Jinxuan Xu, Zhenghao Yu, and Jun Wang. Automatic detection and classification of sewer defects via hierarchical deep learning. *IEEE TASE*, 16(4): 1836–1847, 2019. 7

[46] Jiahao Xu, Hongda Tian, Zhiyong Wang, Yang Wang, Wenxiong Kang, and Fang Chen. Joint input and output space learning for multi-label image classification. *IEEE TMM*, 23:1696–1707, 2020. 6

[47] Jiazhi Xu, Sheng Huang, Fengtao Zhou, Luwen Huangfu, Daniel Zeng, and Bo Liu. Boosting multi-label image classification with complementary parallel self-distillation. *arXiv preprint arXiv:2205.10986*, 2022. 6

[48] Shichao Xu, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Zhu Qi. Open vocabulary multi-label classification with dual-modal decoder on aligned visual-textual features. *arXiv preprint arXiv:2208.09562*, 2022. 6

[49] Zhen Yang, Jun Yue, Pedram Ghamisi, Shiliang Zhang, Jiayi Ma, and Leyuan Fang. Open set recognition in real world. *IJCV*, 132(8):3208–3231, 2024. 1

[50] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *CVPR*, pages 13440–13449, 2020. 6

[51] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *AAAI*, pages 12709–12716, 2020. 6

[52] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu. Multilabel image classification with regional latent semantic dependencies. *IEEE TMM*, 20(10):2801–2813, 2018. 6

[53] Xiaokang Zhang, Wanjing Wu, Mi Zhang, Weikang Yu, and Pedram Ghamisi. Prototypical unknown-aware multiview consistency learning for open-set cross-domain remote sensing image classification. *IEEE TGRS*, 62:1–16, 2024. 2

[54] Haifeng Zhao, Shuo Xu, Leilei Ma, Yufei Zhang, Lei Wang, and Dengdi Sun. Towards space and semantics: Object-purified representation learning for multi-label image classification. In *ACM MM*, page 3270–3279. Association for Computing Machinery, 2025. 2, 6

[55] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li. Transformer-based dual relation graph for multi-label image recognition. In *ICCV*, pages 163–172, 2021. 6

[56] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016. 4

[57] Wei Zhou, Zhijie Zheng, Tao Su, and Haifeng Hu. Datran: Dual attention transformer for multi-label image classification. *IEEE TCSVT*, 34(1):342–356, 2023. 6

[58] Ke Zhu and Jianxin Wu. Residual attention: A simple but effective method for multi-label recognition. In *ICCV*, pages 184–193, 2021. 6

[59] Xuelin Zhu, Jian Liu, Jiuxin Cao, and Bing Wang. Mambaml: Exploring state space models for multi-label image classification. In *ICCV*, pages 4743–4753, 2025. 1, 6, 7