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.

Seeing Through the Shift: Causality-Inspired Robust Generalized Category Discovery

Wei Feng[1,2] Yiwen Jiang[1] Sijin Zhou[1,2] Zhuang Qi[3]
Zhongxing Xu[1] Zhonghua Wang[1] Feilong Tang[1] Zongyuan Ge[1,2]
[1]AIM for Health Lab, Monash University [2]Airdoc–Monash Research, Monash University
[3]School of Software, Shandong University
{wf02429, sjzhou1995}@gmail.com {yiwen.jiang, zongyuan.ge}@monash.edu z_qi@mail.sdu.edu.cn

Abstract

*Generalized Category Discovery (GCD) aims to transfer knowledge from known categories to automatically discover new, unseen ones while preserving recognition of the known classes. Despite recent progress, existing GCD approaches typically assume that all data are drawn from the same distribution, which is rarely valid in real-world scenarios. In practice, data often experience simultaneous domain shifts and novel category emergence, causing severe performance degradation of existing systems. To address this challenge, we propose CausalGCD, a causality-inspired framework designed to mitigate domain-shift bias in category discovery. Specifically, we first analyze the causal graph to uncover the relationships among key variables in cross-domain GCD. We then introduce the concept of causal dependency risk and propose a Causal Dependency Risk Estimator to capture causal semantics, further deriving a theoretically computable upper bound to optimize this risk under cross-domain GCD settings. Furthermore, we propose a Causal Geometric Manifold Constraint that enforces invariant manifold-level associations between known and unknown categories across domains, thereby facilitating robust discovery of novel classes. Extensive experiments on the **SSB-C** and **DomainNet** benchmarks demonstrate the effectiveness of **CausalGCD** and highlight the significance of causal reasoning in open-world category discovery.*

1. Introduction

Deep learning models have achieved remarkable success in closed-world scenarios, where all categories are predefined and the unlabeled data are assumed to originate from known classes [5–10, 16, 18–20]. However, this assumption severely limits their applicability in open-world environments, where unlabeled data may also contain unknown categories [4, 7, 11, 17, 37]. To address this challenge, *Generalized Category Discovery* (GCD) has been proposed to cluster unlabeled samples—potentially containing both known and unknown classes—by transferring knowledge from labeled known categories [37, 41]. From early two-stage optimization pipelines [37] to more recent single-stage strategies [25, 39, 41], numerous GCD methods have achieved encouraging results. Nevertheless, most of them still assume that the labeled and unlabeled datasets share the same underlying distribution, which is often unrealistic in real-world applications. For example, in medical imaging or cross-camera recognition, images captured from different devices or environments exhibit substantial domain shifts, leading to significant performance degradation when models are deployed in unseen domains.

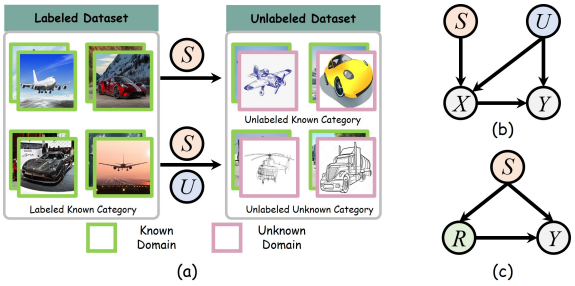


Figure 1. Causal graph of the cross-domain GCD paradigm, where interventions on domain-specific factors (S) lead to domain shift, and interventions on causal semantic factors (U) give rise to novel categories. Changes in S may also confound the relation (R) between known and unknown classes, hindering robust category discovery.

Recent approaches such as HiLo [38] and CDAD-Net [32] extend GCD to cross-domain settings, yet they rely on heuristic feature separation or adversarial alignment, failing to capture the underlying causal mechanisms of domain shifts. Despite the growing interest in causality [28], cross-domain GCD still lacks a unified causal graph and a principled formulation of causal risk. To fill this gap, we propose

CausalGCD—a causality-inspired framework for cross-domain generalized category discovery. From a causal perspective, latent representations can be decomposed into *semantic causal features* (U) and *domain-specific factors* (S), where we assume $S \perp U$ and $S \perp Y|U$. As illustrated in Figure 1, domain distribution shifts arise from interventions on S, which alter the data distribution while keeping the causal mechanism $P(Y|U)$ invariant. When causal features U vary significantly, the label semantics also change, leading to the emergence of novel categories. Building on this insight, we introduce a *Causal Dependency Risk (CDR)* to quantify and optimize the causal dependency between features and predictions, and theoretically derive its upper bound for the unlabeled target domain. Furthermore, by enforcing the causal assumptions of *exogeneity* and *monotonicity* through prototype-level mutual information maximization, we ensures that the extracted features comply with the causal assumptions and enable the identification of counterfactual causal-dependency risk.

Moreover, in GCD tasks, learning robust associations between known and unknown categories is crucial for discovering new ones. However, domain shifts often distort this relational knowledge transfer. As shown in Figure 1(c), the domain-specific variable S acts as a confounder that entangles the relational factor R between known and unknown categories, thereby hindering effective category discovery. To disentangle this confounding effect ($R \nleftarrow S$), we propose a *Causal Geometric Manifold Constraint* that models invariant manifold-level associations across domains, thus enabling robust and domain-invariant category discovery. Our major contributions can be summarized as follows:

- We introduce **CausalGCD**, the first structured causal model for cross-domain GCD. We theoretically derive a computable upper bound for the causal dependency risk, addressing the issues of model **optimizability**, **evaluability**, and **identifiability**, and providing a principled causal framework for category discovery under domain shift.
- We propose a *Causal Geometric Manifold Constraint* that preserves causal-geometric correlations shared across domains, supporting more stable discovery of novel categories.
- Extensive experiments on the **SSB-C** and **DomainNet** benchmarks demonstrate that CausalGCD significantly outperforms state-of-the-art GCD methods. Comprehensive ablation studies further validate the effectiveness and interpretability of each proposed component.

2. Related Works

Category Discovery. Research on category discovery generally spans two paradigms: **Novel Class Discovery (NCD)** and **Generalized Category Discovery (GCD)**. NCD, first proposed by Han et al. [17], focuses on identifying previously unseen categories within unlabeled samples using supervision from known classes only. Early studies employed self-supervised contrastive learning and ranking strategies to extract transferable semantic features [17, 46]. UNO [12] advanced this field by introducing a unified self-labeling framework based on the Sinkhorn–Knopp algorithm with multi-objective optimization, while later works further improved clustering granularity via inter-class semantic modeling [15, 24]. GCD extends NCD by allowing unlabeled data to contain both seen and unseen classes. Vaze et al. [37] jointly optimized supervised and self-supervised contrastive losses followed by semi-supervised clustering. SimGCD [41] further introduced a parametric classifier trained with entropy regularization and self-distillation for end-to-end optimization. SPTNet [39] utilized spatial prompt tuning for cross-category semantic transfer. Recent efforts have generalized GCD to the **cross-domain** scenario. HiLo [38] mitigates domain shift via semantic–domain disentanglement and PatchMix [47], though its independence assumption and mixed patches may introduce noise. CDAD-Net [32] uses entropy-driven adversarial learning but inherits adversarial instability. FREE [4] approaches the problem from the frequency domain using tailored perturbations for more robust category discovery. Different from all the above, **our work is the first to incorporate causal reasoning into cross-domain GCD**, proposing a structured causal model and deriving both theoretical and algorithmic foundations based on counterfactual causal probabilities. This principled framework ensures model *optimability*, *evaluability*, and *identifiability*, providing a causal view of robust discovery across domains.

Unsupervised Domain Adaptation. Unsupervised Domain Adaptation (UDA) aims to align labeled source and unlabeled target distributions. Classical approaches rely on *distribution alignment*, either through moment matching [26, 30, 36] or adversarial domain confusion [13, 33, 34], encouraging domain-invariant but discriminative features. More recent methods pursue *architecture-level transferability*. TVT [42] introduces task-aware attention for disentangling generalizable and domain-specific representations, while MTTrans [44] and CDAC [40] ensure layer-wise consistency for stable multi-scale adaptation.

Recently, **causal inference** has emerged as a principled direction for domain adaptation. IRM [1] learns representations where the optimal classifier remains invariant across environments, and iCaRL [27] extends this to nonlinear regimes by identifying latent causal factors under flexible exponential-family priors. EAGLE [45] further models latent spatio-temporal environments to extract invariant patterns on dynamic graphs, while Yang et al. [43] employ the probability of necessary and sufficient causes (PNS) to align representations with causal sufficiency. Lv et al. [28]

formalize domain generalization through structural causal models that disentangle causal and non-causal factors, offering a theoretical lens for invariant mechanism learning. However, explorations of causal graphs and comprehensive risk formulations for **cross-domain GCD** remain scarce — this is precisely the gap our work aims to fill.

3. Method

3.1. Problem Definition

In the cross-domain GCD task, we are given a labeled dataset $\mathcal{D}_l = \{(x_i, y_i)\}_{i=1}^{N_l}$, where $x_i \in \mathcal{X}_l$ are input samples and $y_i \in \mathcal{Y}_l$ are labels of *known categories*. In addition, there exists an unlabeled dataset $\mathcal{D}_u = \{x_j\}_{j=1}^{N_u}$ that contains both known and novel categories. The label spaces of the two datasets can be represented as $\mathcal{Y}_l \subset \mathcal{Y}_u, \mathcal{Y}_u^{\text{known}} = \mathcal{Y}_l, \mathcal{Y}_u^{\text{novel}} = \mathcal{Y}_u \setminus \mathcal{Y}_l$, where $\mathcal{Y}_u^{\text{known}}$ corresponds to the categories shared with the labeled dataset, and $\mathcal{Y}_u^{\text{novel}}$ denotes the *unlabeled novel categories*. It is worth noting that the unlabeled dataset $\mathcal{D}_u$ may come from either the same domain as the labeled dataset or a different domain exhibiting a distribution shift. The key challenge of cross-domain GCD lies in robustly discovering novel categories in the unlabeled data while preserving recognition ability on known categories, especially under potential cross-domain variations.

3.2. Causal Dependency Risk Estimator

Spurious domain-dependent cues often distort the intrinsic relationship between semantic features and labels, causing models to rely on unstable shortcuts. To overcome this limitation, we draw inspiration from causal inference, where counterfactual reasoning evaluates how interventions on underlying variables influence outcomes. Under suitable causal assumptions—such as **exogeneity** and **monotonicity**—these intervention-induced effects become identifiable and can be linked to measurable changes in prediction behavior. Motivated by this perspective, we define a causal dependency measure that quantifies how prediction outcomes respond when the latent representation is subjected to intervention:

$$\mathrm{CDM} = P(Y = y \mid u) - P(Y = y \mid u'), \tag{1}$$

where u and u' denote two specific states of the causal variable U, corresponding to its post-intervention and pre-intervention values, respectively. A larger CDM indicates stronger causal dependency, whereas a smaller value signals unstable or weak influence. Building on this formulation, we introduce the **Causal Dependency Risk** (CDR), which encourages the model to emphasize components with stronger causal relevance and reduces reliance on domain-specific shortcuts:

$$\mathrm{CDR}_{\mathbb{Q}}(u, u') = -\mathrm{CDM} := \mathcal{O}_{\mathbb{Q}}(u) - \mathcal{I}_{\mathbb{Q}}(u'), \tag{2}$$

where $\mathcal{O}_{\mathbb{Q}}(u)$ and $\mathcal{I}_{\mathbb{Q}}(u')$ represent the *observational relevance risk* and the *interventional relevance risk*, respectively:

$$\begin{aligned}
\mathcal{O}_{\mathbb{Q}}(u) &= \mathbb{E}_{(x,y)\sim\mathbb{Q}}\Big[\mathbb{E}_{u\sim P_{\mathbb{Q}}(U|X=x)} P^g(Y \neq y \mid U = u)\Big], \\
\mathcal{I}_{\mathbb{Q}}(u') &= \mathbb{E}_{(x,y)\sim\mathbb{Q}}\Big[\mathbb{E}_{u'\sim P_{\mathbb{Q}}(U'|X=x)} P^g(Y \neq y \mid U' = u')\Big].
\end{aligned} \tag{3}$$

where $g(\cdot)$ is a classifier that models the conditional probability $P^g(Y \mid U)$. $G(\cdot)$ models the posterior $P_{\mathbb{Q}}(U \mid X)$ and yields the invariant causal representation u, while $G'(\cdot)$ specifies $P_{\mathbb{Q}}(U' \mid X)$ and produces the intervened representation u'. $\mathbb{Q}$ denotes the unlabeled target-domain distribution. Intuitively, $\mathcal{O}_{\mathbb{Q}}(u)$ characterizes the predictive risk under the observational distribution, whereas $\mathcal{I}_{\mathbb{Q}}(u')$ captures the risk after a semantic intervention. Their difference, $\mathrm{CDR}_{\mathbb{Q}}(u, u')$, thus measures the model's sensitivity to causal interventions of the representation. Minimizing $\mathrm{CDR}_{\mathbb{Q}}(u, u')$ reduces the discrepancy between observational and interventional predictions. This optimization enforces causal invariance, suppresses domain-specific spurious correlations, and drives the model toward domain-invariant causal representations that remain stable under semantic interventions.

Next, we present several theorems to derive an evaluable upper bound of the causal dependence risk in the unlabeled target domain.

Linking Known-Class and Novel-Class Recognition Risks. In the cross-domain GCD setting, the unlabeled target-domain data contain both known and novel categories. Given the known-class space $\mathcal{Y}_l$ and the novel-class space $\mathcal{Y}_u^{\text{novel}}$, the causal dependency risk associated with unlabeled data can be reformulated as follows.

Theorem 1. *Given the labeled source-domain distribution $\mathbb{P}$ and the unlabeled target-domain distribution $\mathbb{Q}$, the causal dependency risk $\mathrm{CDR}_{\mathbb{Q}}(u, u')$ can be rewritten as*

$$\begin{aligned}
\mathrm{CDR}_{\mathbb{Q}}(u, u') = \mathbb{E}_{(x,y)\sim\mathbb{Q}}\Big[&\mathbb{I}(y \in \mathcal{Y}_l)\big(\mathbb{E}_{u\sim P_{\mathbb{Q}}(U|X=x)} \\
&P(Y \neq y \,|\, U = u) - \mathbb{E}_{u'\sim P_{\mathbb{Q}}(U'|X=x)} P(Y \neq y \,|\, U' = u')\big)\Big] \\
&+ \mathbb{E}_{(x,y)\sim\mathbb{Q}}\Big[\mathbb{I}(y \in \mathcal{Y}_u^{\text{novel}})\big(\mathbb{E}_{u\sim P_{\mathbb{Q}}(U|X=x)} \\
&P(Y \neq y \,|\, U = u) - \mathbb{E}_{u'\sim P_{\mathbb{Q}}(U'|X=x)} P(Y \neq y \,|\, U' = u')\big)\Big] \\
&:= \pi_{\mathbb{Q}\wedge\mathbb{P}}(X, Y) + \pi_{\mathbb{Q}\setminus\mathbb{P}}(X, Y).
\end{aligned} \tag{4}$$

where $\pi_{\mathbb{Q}\wedge\mathbb{P}}(X, Y)$ denotes the causal dependency risk on known categories, and $\pi_{\mathbb{Q}\setminus\mathbb{P}}(X, Y)$ measures the novel-category discovery risk [35]. A complete proof is provided in Appendix 1.1.

Linking Cross-Domain Dependency Risks. Since only labeled source data and unlabeled target data are available, the causal dependency risk on the target domain cannot be

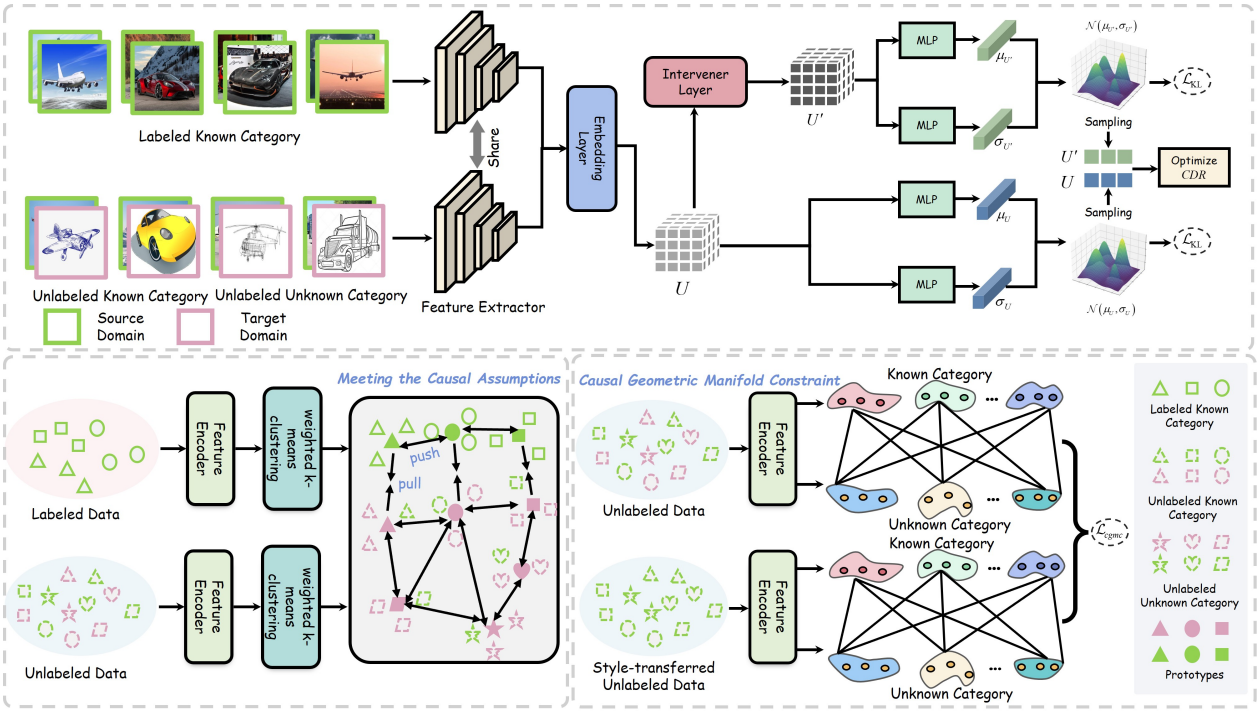


Figure 2. Overall architecture of the proposed **CausalGCD** framework. An intervention module implemented by a multilayer perceptron (MLP) is appended after the feature representation U to perform nonlinear interventions, modeling its distribution as a Gaussian with learnable mean and variance, similar to the VIB formulation [21]. Prototype-level mutual information maximization enforces exogeneity to satisfy causal assumptions, while the **Causal Geometric Manifold Constraint (CGMC)** maintains geometric consistency between known and novel categories, facilitating reliable discovery of unseen classes across domains.

directly computed. To obtain a tractable surrogate, we introduce two quantities that do not require target labels: the **Expected Collaborative Deviation (ECD)** and the **Expected Relational Inconsistency (ERI)** [23]:

$$\varepsilon_{\mathbb{Q}}(u) = \mathbb{E}_{g_1,g_2\sim\varsigma}\mathbb{E}_{(x,y)\sim\mathbb{Q}}\mathbb{I}[g_1(u)\neq y]\,\mathbb{I}[g_2(u)\neq y],$$
$$\delta_{\mathbb{Q}}(u) = \mathbb{E}_{g_1,g_2\sim\varsigma}\mathbb{E}_{x\sim\mathbb{Q}}\mathbb{I}[g_1(u)\neq g_2(u)]. \tag{5}$$

Here, $\varepsilon_{\mathbb{Q}}(u)$ measures correlated prediction errors, while $\delta_{\mathbb{Q}}(u)$ captures hypothesis-space disagreement under any probability distribution ς over g. Together, they characterize the shift of causal dependencies across domains. Using the β-divergence [14] to quantify distributional mismatch and applying Hölder's inequality, we obtain the following upper bound for the causal dependency risk (proof in Appendix 1.2).

Theorem 2. *Let $\mathbb{P}$ and $\mathbb{Q}$ denote the labeled source and unlabeled target distributions. For the inference encoder G, intervention encoder G', and classifier g, the causal dependency risk on $\mathbb{Q}$ is upper bounded by:*

$$\mathrm{CDR}_{\mathbb{Q}}(u,u') < \frac{1}{2}\big(\delta_{\mathbb{Q}}(u)-\delta_{\mathbb{Q}}(u')\big) + \lim_{q\to\infty}\beta_q(\mathbb{Q}\|\mathbb{P})\big(\varepsilon_{\mathbb{P}}(u)-\varepsilon_{\mathbb{P}}(u')\big)+\pi_{\mathbb{Q}\backslash\mathbb{P}}(X,Y), \tag{6}$$

where

$$\mathcal{L}_{\mathbb{Q}} = \delta_{\mathbb{Q}}(u)-\delta_{\mathbb{Q}}(u')+\pi_{\mathbb{Q}\backslash\mathbb{P}}(X,Y), \mathcal{L}_{\mathbb{P}} = \varepsilon_{\mathbb{P}}(u)-\varepsilon_{\mathbb{P}}(u'). \tag{7}$$

This result provides a tractable link between source and target causal risks through relational inconsistency, collaborative deviation, and cross-domain distributional divergence.

Relating Expected and Empirical Causal-Dependency Risk. The causal-dependency risk $\mathrm{CDR}_{\mathbb{Q}}(u,u')$ is defined with respect to the underlying target distribution $\mathbb{Q}$ and therefore corresponds to an *expected risk* quantity that cannot be evaluated directly. To construct a computable surrogate, we derive its empirical counterpart using the available unlabeled and labeled sample sets: $\mathcal{D}_u = \{x_i\}_{i=1}^{N_u}, \mathcal{D}_l = \{(x_i,y_i)\}_{i=1}^{N_l}$. Let $\bar{P}(Y\neq y \mid U=u), \bar{P}(Y\neq y \mid U'=u')$ denote disagreement estimates induced by the invariant encoder G and the intervention encoder G'. Using these quantities, the *empirical causal-dependency risk* is defined as:

$$\mathrm{CDR}_{\mathcal{D}_u}(u,u') = \mathbb{E}_{x\sim\mathcal{D}_u}\mathbb{E}_{u\sim\bar{P}(U|X=x)}\bar{P}(Y\neq y\mid U=u) - \mathbb{E}_{x\sim\mathcal{D}_u}\mathbb{E}_{u'\sim\bar{P}(U'|X=x)}\bar{P}(Y\neq y\mid U'=u'). \tag{8}$$

The fundamental objective is to relate this empirical discrepancy to the expected discrepancy $\mathrm{CDR}_{\mathbb{Q}}(u, u')$. To achieve this, we introduce prior distributions over the representation families: $\nu_U = P(U), \nu_{U'} = P(U')$. Following these priors, we derive Corollary 1 using variational inference together with Hoeffding's inequality, Markov's inequality, and Jensen's inequality [43], providing a high-probability control of the gap between the expected and empirical causal-dependency risks (see proof in Appendix 1.3).

Corollary 1. *With probability at least $1-\delta$ over draws $X \sim \mathcal{D}_u/\mathcal{D}_l$, the expected and empirical causal-dependency discrepancies satisfy:*

$$\begin{aligned} &\left|\mathrm{CDR}_{\mathbb{Q}}(u, u') - \mathrm{CDR}_{\mathcal{D}_u}(u, u')\right| \le \mathcal{L}_{\mathrm{KL}}, \\ &\text{where} \quad \mathcal{L}_{\mathrm{KL}} = -\log(\sigma_U) + \log(\sigma_{U'}). \end{aligned} \tag{9}$$

The term L_{KL} acts as a *representation variability index*, quantifying how the stochasticity of U and U' influences the deviation between the expected and empirical causal-dependency risks. Smaller KL divergences tighten this index, ensuring that $\mathrm{CDR}_{\mathcal{D}_u}$ serves as a faithful approximation of $\mathrm{CDR}_{\mathbb{Q}}$.

3.3. Satisfaction of Causal Assumptions

To reliably identify counterfactual causal-dependency risk, the learned representations must satisfy two fundamental causal assumptions: *exogeneity* and *monotonicity*. In this section, we examine how these assumptions are ensured through our representation learning framework.

Monotonicity. The monotonicity assumption requires that intervening on the causal representation induces a consistent directional change in prediction. In particular, replacing the invariant representation u with its intervened counterpart u' should not *increase* the model's confidence in the original label y. This requirement can be expressed as the following optimization objective:

$$\max \; P(Y = y \mid U = u) - P(Y = y \mid U = u'). \tag{10}$$

In fact, monotonicity is naturally satisfied during training, as minimizing the causal-dependency risk directly enforces the inequality in Eq. (10).

Exogeneity. The exogeneity assumption requires that domain factors S do not confound the causal path from U to Y, i.e., $P(Y \mid U, S) = P(Y \mid U)$. A sufficient condition is to maximize the cross-domain semantic consistency of each class by increasing the mutual information between source and target prototypes, i.e., $\max_G I(e^k_{\mathbb{P}}, e^k_{\mathbb{Q}})$ for all $k \in \mathcal{Y}_l$. When $I(e^k_{\mathbb{P}}, e^k_{\mathbb{Q}})$ is maximized, the class-conditional distributions $P(Y{=}k \mid U)$ become aligned across domains, leading to $P(Y \mid U, S) \approx P(Y \mid U)$ (proof in Appendix 1.4).

To enforce exogeneity in practice, we first estimate the unlabelled target-domain prototype $e^k_{\mathbb{Q}}$ using weighted feature averaging, $e^k_{\mathbb{Q}} = \frac{\sum_i p_{i,k} G(x_i)}{\sum_i p_{i,k}}$, where $G(x_i)$ is the feature of sample x_i and $p_{i,k}$ is its soft assignment to class k. The labelled source-domain prototype $e^k_{\mathbb{P}}$ is obtained in a similar manner. We then adopt a prototype-alignment loss that pulls same-class prototypes together while pushing different or novel-class prototypes apart:

$$\mathcal{L}_{\mathrm{e}} = -\sum_{k \in \mathcal{Y}_l} \log \frac{\exp(\mathrm{sim}(e^k_{\mathbb{P}}, e^k_{\mathbb{Q}}))}{\sum_{e \in \mathcal{N}(k)} \exp(\mathrm{sim}(e^k_{\mathbb{P}}, e))}, \tag{11}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity, $\mathcal{N}(k)$ includes $e^k_{\mathbb{Q}}$, all other known-class prototypes $\{e^{k'}_{\mathbb{Q}} : k' \neq k\}$, and all novel-class prototypes $\{e^j_{\mathbb{Q}}\}_{j=1}^{\mathcal{Y}_u^{\text{novel}}}$. This loss increases $I(e^k_{\mathbb{P}}, e^k_{\mathbb{Q}})$ and separates all novel-class prototypes, thereby suppressing domain-specific nuisance factors and ensuring the learned representation satisfies exogeneity. To promote semantic separability under the causal premise [43], where features with different semantics are expected to remain sufficiently apart, we also impose: $\mathcal{L}_{\text{sep}} = \kappa - \|u - u'\|_2^2$, where κ denotes the intervention strength on causal factors. A sufficiently large κ enforces a semantic margin between u and u', mitigating entanglement and unstable learning.

3.4. Causal Geometric Manifold Constraint

Although the above causal-enhanced features capture invariant and semantically meaningful representations, they remain inadequate for reliably discovering novel categories under domain shifts. We argue that the intrinsic relationships between known and unknown categories are essential for novel-class discovery. Inspired by manifold theory [29], we model these relationships through the geometric structures of category distributions, which capture transferable semantic correlations essential for discovering unseen categories. To preserve these structures under domain shifts, we propose the **Causal Geometric Manifold Constraint (CGMC)**, which enforces cross-domain consistency in the geometric relations between every known–unknown category pair. We first generate cross-domain style-transferred unlabeled samples via Fourier-based style transfer: $\tilde{x}^u_i = \mathcal{F}^{-1}(\mathcal{A}^l_i \cdot e^{j\mathcal{P}^u_i})$, where $\mathcal{A}^l_i$ is the amplitude (domain style) of a randomly selected labeled source-domain sample x^l_i, and $\mathcal{P}^u_i$ is the phase (semantics) of an unlabeled target-domain sample x^u_i. $\mathcal{F}^{-1}$ denotes the inverse Fourier transform. This produces a style-shifted version that separates domain factors from geometric semantics. We then feed both the original unlabeled samples and their style-transferred counterparts into the model to obtain their feature embeddings.

Category geometry [29]. Given n D-dimensional embeddings $\mathbf{Z} = [G(x_1), \ldots, G(x_n)] \in \mathbb{R}^{D \times n}$ predicted as category k, their covariance matrix is $\mathbf{\Sigma}_Z = \mathbb{E}\big[\frac{1}{n}\sum_{i=1}^n G(x_i)G(x_i)^\top\big] = \frac{1}{n}\mathbf{Z}\mathbf{Z}^\top \in \mathbb{R}^{D \times D}$. Let $\{\xi_1, \ldots, \xi_m\}$ be the top-m eigenvectors of $\mathbf{\Sigma}_Z$; these form the principal geometric manifold: $\mathrm{CG} = \{\xi_1, \ldots, \xi_m\}$. For a known category k and an unknown category r, let CG^k_d and CG^r_d denote their geometric manifolds in domain

$d \in \{\mathbb{P}, \mathbb{Q}\}$ (corresponding to original and style-transferred unlabeled samples, respectively). Their geometric correlation is defined as $\mathrm{Sim}_d(k,r) = \sum_{i=1}^{m} \left|\langle \xi_{i,d}^k, \xi_{i,d}^r \rangle\right|$. If semantic structure is stable, then $\mathrm{Sim}_{\mathbb{P}}(k,r) \approx \mathrm{Sim}_{\mathbb{Q}}(k,r)$ should hold across domains.

Causal geometric manifold constraint. CGMC enforces cross-domain invariance of the known–unknown geometric correlation by minimizing their discrepancy:

$$\mathcal{L}_{\mathrm{cgmc}} = \sum_{k \in \mathcal{Y}_l} \sum_{r \in \mathcal{Y}_u^{\mathrm{novel}}} \left|\mathrm{Sim}_{\mathbb{P}}(k,r) - \mathrm{Sim}_{\mathbb{Q}}(k,r)\right|. \quad (12)$$

By preserving the geometric relationships between known and unknown categories across domains, CGMC eliminates domain-induced geometric distortions and ensures that category discovery relies on stable causal semantic structure rather than spurious domain correlations.

The overall objective of our framework is formulated as:

$$\mathcal{L} = \mathcal{L}_{\mathbb{P}} + \lambda_1 \mathcal{L}_{\mathbb{Q}} + \lambda_2 \mathcal{L}_{\mathrm{KL}} + \lambda_3 \mathcal{L}_{\mathrm{e}} + \lambda_4 \mathcal{L}_{\mathrm{cgmc}} + \lambda_5 \mathcal{L}_{\mathrm{sep}}, \quad (13)$$

where $\mathcal{L}_{\mathbb{P}}$ measures the empirical causal-dependency risk on the source domain using ground-truth labels, $\mathcal{L}_{\mathbb{Q}}$ estimates the causal-dependency risk on the target domain using predicted pseudo-labels, $\mathcal{L}_{\mathrm{KL}}$ reduces the gap between expected and empirical risks, $\mathcal{L}_{\mathrm{e}}$ enforces the exogeneity assumption, and $\mathcal{L}_{\mathrm{cgmc}}$ imposes the causal geometric manifold constraint for stable novel-category discovery. $\mathcal{L}_{\mathrm{sep}}$ enforces feature separability and contributes to training stability.

4. Experiments

4.1. Experimental Setup

Datasets. To comprehensively evaluate the proposed framework, we conduct experiments on two widely used benchmarks: the Corrupted Semantic Shift Benchmark (SSB-C) [38] and DomainNet [30]. The SSB-C benchmark is derived from SSB by introducing nine types of image corruptions, each applied at five severity levels, to assess model robustness under distributional perturbations. DomainNet consists of approximately 600K images spanning 345 categories across six distinct domains: Real, Clipart, Infograph, Painting, Quickdraw, and Sketch.

For domain partitioning, we follow the protocol in HiLo [38]. In SSB-C, the clean subset of each dataset serves as the source domain, while all corrupted variants are aggregated to form the target domain. For DomainNet, we use the *Real* domain as the source and, in turn, designate each of the remaining domains as the target, forming several domain pairs. We also explore a multi-domain configuration, where all non-Real domains are combined as a single target domain (see Appendix for detailed results).

For category partitioning, we adopt the same strategy as HiLo [38]. Specifically, a subset of source-domain categories is selected as the known classes, from which 50% of the samples constitute the labeled dataset $\mathcal{D}_l$. The remaining samples, including both known and novel classes from either source or target domains, form the unlabeled dataset $\mathcal{D}_u$. A summary of the dataset splits and statistics is provided in Table 1 of the Appendix.

Implementation Details and Evaluation Metrics. We adopt DINO-pretrained ViT-B/16 [3] as the backbone and fine-tune only the last transformer block for 200 epochs using SGD with a weight decay of 5×10^{-5}. The learning rate starts at 0.1 and decays to 0.0001 via cosine annealing, with a batch size of 128. Hyperparameters λ_1, λ_2, λ_3, λ_4, and λ_5 are set to 0.5, 0.5, 0.5, 0.5, and 0.3, respectively. We set the parameters $m = 5$ and $\kappa = 0.5$ in all experiments (**their sensitivity analysis is provided in the Appendix**). All experiments are repeated three times with different random seeds, and mean results are reported. Models are implemented in PyTorch and trained on eight NVIDIA RTX-4090 GPUs.

Following prior works [37, 38], clustering accuracy (ACC) is used as the main evaluation metric. Let $\hat{y}_i$ and y_i denote the predicted and ground-truth labels of sample i from the unlabeled target set $\mathcal{D}_u$. We compute the optimal label assignment $\mathcal{P}^*$ using the Hungarian algorithm [22]: $\mathrm{ACC} = \frac{1}{|\mathcal{D}_u|} \sum_{i=1}^{|\mathcal{D}_u|} \mathbb{I}\{y_i = \mathcal{P}^*(\hat{y}_i)\}$. We report ACC on the source and target domains, including overall, old-class, and new-class accuracies.

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

We compare CausalGCD with leading GCD models—ORCA [2], GCD [37], SimGCD [41], SPTNet [39], and RLCD [25]—as well as adapted NCD methods (RankStats+ [17], UNO+ [12]). We further include domain-oriented approaches such as CDAD-Net [32], HiLo [38], and FREE [4]. Results in Tables 1 and 2 show that most existing methods deteriorate notably once domain shifts are introduced, with unlabeled samples from unseen domains often degrading accuracy even within the known domain. In contrast, CausalGCD provides consistent and substantial improvements. On the corrupted domains of the SSB-C benchmark, it achieves the best results on nearly all domain pairs, outperforming FREE by 2.1% on CUB-C, 2.6% on Scars-C, and 2.2% on FGVC-C. Gains are also observed in the clean domains, indicating stronger cross-domain stability. On DomainNet, our model again improves performance across most domain combinations. For instance, using Real as the known domain and Painting as the unknown domain, CausalGCD exceeds FREE by 2.4% on Painting and 2.2% on Real. Overall, the results confirm that CausalGCD not only withstands domain shifts but also delivers

Table 1. Clustering results of various methods on the DomainNet benchmark. Each experiment uses *Real* as the source domain, with one of *Painting*, *Sketch*, *Quickdraw*, *Clipart*, or *Infograph* serving as the target. Clustering accuracy is reported for both domains.

Methods	Real + Painting						Real + Sketch						Real + Quickdraw						Real + Clipart						Real + Infograph					
	Real			Painting			Real			Sketch			Real			Quickdraw			Real			Clipart			Real			Infograph		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
RankStats+	34.1	62.0	19.7	29.7	**49.7**	9.6	34.2	62.0	19.8	17.1	**31.1**	6.8	34.1	62.5	19.5	4.1	4.4	3.9	34.0	62.4	19.4	24.1	**45.1**	6.2	34.2	62.4	19.6	12.5	**21.9**	6.3
UNO+	44.2	72.2	29.7	30.1	45.1	17.2	43.7	72.5	28.9	12.5	17.0	9.2	31.1	60.0	16.1	6.3	5.8	6.8	44.5	66.1	33.3	21.9	35.6	10.1	42.8	69.4	29.0	10.9	15.2	8.0
ORCA	31.9	49.8	23.5	28.7	38.5	7.1	32.5	50.0	23.9	11.4	14.5	7.2	19.2	39.1	15.3	3.4	3.5	3.2	32.0	49.7	23.9	19.1	31.8	4.3	29.1	47.7	20.1	8.6	13.7	7.1
GCD	47.3	53.6	44.1	32.9	41.8	23.0	48.0	53.8	45.3	16.6	22.4	11.1	37.6	41.0	35.2	5.7	4.2	6.9	47.7	53.8	44.3	22.4	34.4	16.0	41.9	46.1	39.0	10.9	17.1	8.8
SimGCD	61.3	77.8	52.9	34.5	35.6	33.5	62.4	77.6	54.6	16.4	20.2	13.6	47.4	64.5	37.4	6.6	5.8	7.5	61.6	77.2	53.6	23.9	31.5	17.3	52.7	67.0	44.8	11.6	15.4	9.1
SPTNet	61.6	76.9	54.7	35.2	35.9	35.1	63.3	77.8	55.3	16.7	26.0	11.3	47.1	65.6	35.4	6.9	5.7	7.7	62.5	76.5	55.4	24.7	30.9	18.8	54.5	67.9	46.2	11.9	19.4	7.9
RLCD	62.1	**78.3**	53.8	36.9	35.7	36.2	62.8	77.4	55.7	17.0	20.4	15.2	49.1	67.8	38.0	7.0	5.8	7.8	62.3	77.1	54.7	24.5	38.0	13.9	57.2	68.3	48.1	12.0	15.9	9.8
CDAD-Net	63.6	77.8	56.3	38.4	38.4	37.5	61.9	76.3	52.1	17.3	20.9	15.9	48.5	66.5	36.7	6.4	5.6	7.3	61.3	77.0	53.1	25.2	31.9	19.0	56.5	68.0	47.1	11.8	15.6	9.4
HiLo	64.4	77.6	57.5	42.1	42.9	41.3	63.3	77.9	55.9	19.4	22.4	17.1	58.6	76.4	52.5	7.4	6.9	8.0	63.8	77.6	56.6	27.7	34.6	21.7	64.2	78.1	57.0	13.7	16.4	11.9
FREE	67.7	78.1	61.2	45.6	46.1	44.8	67.8	78.2	61.6	22.5	25.8	20.9	61.4	**78.1**	55.1	8.9	**7.8**	9.0	66.4	**78.1**	60.1	29.3	37.2	26.3	68.1	**78.9**	60.2	16.1	18.6	13.4
CausalGCD	**69.9**	77.9	**64.1**	**48.0**	47.3	**46.4**	**69.3**	**78.7**	**63.5**	**24.3**	25.1	**23.2**	**62.9**	77.7	**57.6**	**9.3**	7.6	**9.9**	**68.0**	77.9	**62.6**	**31.2**	37.0	**28.4**	**70.0**	78.2	**62.1**	**17.5**	19.9	**15.0**

Table 2. Clustering results of various methods on the SSB-C benchmark. For each dataset (CUB, Scars, and FGVC), the clean set serves as the source domain, while its corrupted counterpart is treated as the target domain. Clustering accuracy is reported for both domains.

Methods	CUB-C						Scars-C						FGVC-C					
	Original			Corrupted			Original			Corrupted			Original			Corrupted		
	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New	All	Old	New
RankStats+	19.3	22.0	15.4	13.6	23.9	4.5	14.8	20.8	7.8	11.5	22.6	1.0	14.4	16.4	14.5	8.3	15.6	5.0
UNO+	25.9	40.1	21.3	21.5	33.4	8.6	22.0	41.8	7.0	16.9	29.8	4.5	22.0	33.4	15.8	16.5	25.2	8.8
ORCA	18.2	22.8	14.5	21.5	23.1	18.9	19.1	28.7	11.2	15.0	22.4	8.3	17.6	19.3	16.1	13.9	17.3	10.1
GCD	26.6	27.5	25.7	25.1	28.7	22.0	22.1	35.2	20.5	21.6	29.2	10.5	25.2	28.7	23.0	21.0	23.1	17.3
SimGCD	31.9	33.9	29.0	28.8	31.6	25.0	26.7	39.6	25.6	22.1	30.5	14.1	26.1	28.9	25.1	22.3	23.2	21.4
SPTNet	33.0	34.5	31.2	30.1	33.1	26.1	28.0	40.2	27.9	24.2	32.1	16.3	28.7	30.2	27.9	24.8	25.7	23.9
RLCD	35.9	35.1	33.2	32.3	34.8	28.5	29.8	41.2	30.4	25.3	33.4	18.1	27.9	30.1	26.8	24.4	26.8	22.7
CDAD-Net	40.4	38.9	39.3	37.7	39.1	34.2	32.1	42.9	32.2	28.8	35.6	21.4	33.8	35.5	31.2	27.8	29.6	25.6
HiLo	56.8	54.0	60.3	52.0	53.6	50.5	39.5	44.8	37.0	35.6	42.9	28.4	44.2	50.6	47.4	31.2	29.0	33.4
FREE	60.4	58.5	63.2	55.7	**57.1**	53.7	43.6	48.1	40.8	38.9	**46.1**	32.6	48.5	54.9	51.2	35.0	32.4	38.9
CausalGCD	**62.2**	**60.1**	**64.9**	**57.8**	56.6	**56.9**	**45.8**	**49.1**	**42.2**	**41.5**	45.6	**36.0**	**49.8**	**56.1**	**53.1**	**37.2**	**33.4**	**40.2**

Table 3. Ablation study on different loss components. We use *Real* from DomainNet as the source domain and *Painting* as the target domain, and report clustering performance on both domains.

Components	Real			Painting		
	All	Old	New	All	Old	New
w/o $\mathcal{L}_{\mathbb{P}}$	63.4	67.7	60.1	42.8	39.0	43.4
w/o $\mathcal{L}_{\mathbb{Q}}$	64.5	72.8	58.3	42.4	41.8	38.1
w/o $\mathcal{L}_{\mathrm{KL}}$	66.7	75.4	61.5	45.8	44.6	43.8
w/o $\mathcal{L}_{\mathrm{e}}$	65.5	74.1	59.9	44.6	42.5	43.4
w/o $\mathcal{L}_{\mathrm{cgmc}}$	67.6	77.2	61.4	45.0	45.9	42.8
w/o $\mathcal{L}_{\mathrm{sep}}$	68.2	77.0	62.9	46.3	45.7	44.5
CausalGCD	**69.9**	**77.9**	**64.1**	**48.0**	**47.3**	**46.4**

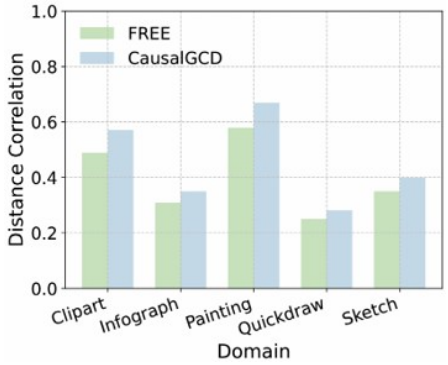
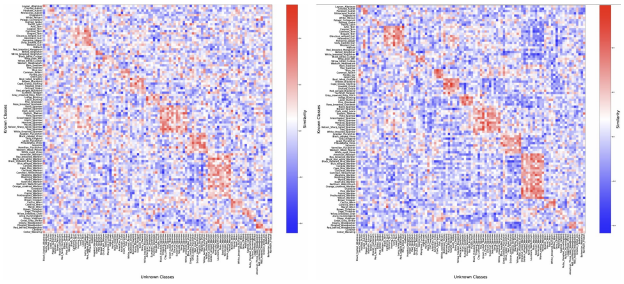


Figure 3. **Distance Correlation** results across five non-real DomainNet domains.

superior clustering accuracy in both known and unseen domains.

4.3. Ablation Studies and Analysis

Effectiveness of Causal Dependency Risk. We first evaluate the effectiveness of the proposed Causal Dependency Risk by verifying its ability to capture intrinsic causal relationships. To approximate ground-truth causal components, we employ the state-of-the-art zero-shot saliency segmentation model IS-Net [31] to extract object regions from each domain in DomainNet, which correspond to label-relevant causal factors. **The extracted object regions are provided in the Appendix**. Both the original and object-masked images are fed into our model and the strongest baseline, FREE, and we compute the distance correlation between the learned representations and the extracted causal features.



(a) FREE (b) CausalGCD

Figure 4. Visualization of geometric associations between the embedding distributions of known and unknown categories under different algorithms.

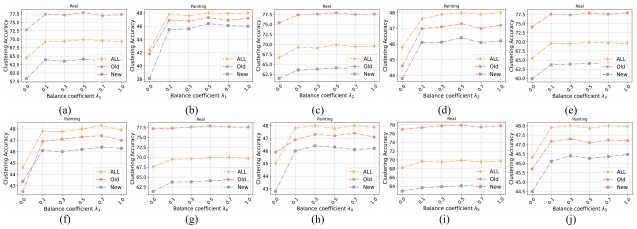


Figure 5. Sensitivity analysis of key hyperparameters.

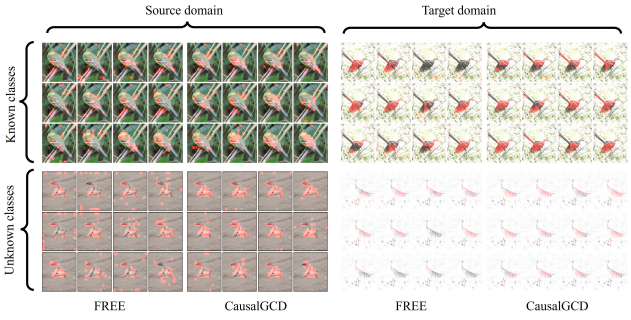


Figure 6. Visualization of attention maps on the *CUB-C* dataset.

As shown in Figure 3, CausalGCD consistently achieves higher correlation with key causal factors across domains, demonstrating that it focuses on semantically causal regions while suppressing spurious correlations. These results confirm that incorporating causal dependency risk effectively drives the model toward invariant and semantically aligned representations.

Effectiveness of Individual Components. We further conduct detailed ablation studies on DomainNet to assess the contribution of each component in our framework. As shown in Table 3, removing either of the domain-specific causal dependency losses ($\mathcal{L}_{\mathbb{P}}$ or $\mathcal{L}_{\mathbb{Q}}$) leads to a marked drop in clustering accuracy, highlighting the necessity of optimizing causal risk across domains. Eliminating the exogeneity loss ($\mathcal{L}_{\mathrm{e}}$) results in reduced recognition of known categories, indicating that this constraint improves model stability and discriminability for seen classes. Removing $\mathcal{L}_{\mathrm{sep}}$ causes unstable training, while excluding $\mathcal{L}_{\mathrm{KL}}$ also degrades performance, highlighting the importance of bridging empirical and expected risks. Similarly, removing the causal geometric manifold constraint ($\mathcal{L}_{\mathrm{cgmc}}$) lowers the discovery accuracy of novel categories, verifying its importance in maintaining structural consistency for unseen-class separation. To provide further insight, we visualize the learned geometric manifolds between known and unknown categories in Figure 4. Interestingly, categories under the same superclass exhibit tighter geometric manifold relationships, suggesting that the learned geometric consistency facilitates effective knowledge transfer and robust novel-class discovery under domain shifts.

Hyperparameter Sensitivity Analysis. We evaluate the sensitivity of our framework to the trade-off coefficients λ_1, λ_2, λ_3, λ_4, and λ_5, which balance the different loss components in the model. Experiments are conducted on the *DomainNet* dataset, using *Real* as the known domain and *Painting* as the unknown domain. As shown in Figure 5, performance remains stable when these coefficients vary within a reasonable range, indicating that our method is robust and not sensitive to hyperparameter selection.

Attention Map Visualization. To further examine whether our model truly focuses on causal features, we visualize the attention maps from the last layer of the ViT backbone. For each input, we aggregate attention across all heads and highlight the top 10% regions with the highest attention values in red to indicate the most informative areas. As shown in Figure 6, our **CausalGCD** primarily attends to the foreground objects that are semantically relevant to category identity, while ignoring background textures and domain-specific artifacts. This demonstrates that our causal learning framework effectively suppresses spurious domain cues and emphasizes truly causal, category-relevant features, leading to more robust and interpretable category discovery.

5. Conclusion

This paper presents **CausalGCD**, a causal-theoretic framework for robust category discovery under domain shifts. CausalGCD explicitly models causal dependencies among key variables through a counterfactual-based *Causal Dependency Risk*, for which we derive a theoretically computable upper bound to guide optimization. In addition, a *Causal Geometric Manifold Constraint* is introduced to maintain geometric consistency between known and novel categories, enabling more reliable discovery of unseen classes. Experiments on DomainNet and SSB-C show that CausalGCD surpasses state-of-the-art methods, achieving higher clustering accuracy and cross-domain robustness.

References

[1] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019. 2

[2] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *arXiv preprint arXiv:2102.03526*, 2021. 6

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[4] Wei Feng and Zongyuan Ge. Generalized category discovery under domain shift: A frequency domain perspective. *arXiv preprint arXiv:2511.00573*, 2025. 1, 2, 6

[5] Wei Feng, Lin Wang, Lie Ju, Xin Zhao, Xin Wang, Xiaoyu Shi, and Zongyuan Ge. Unsupervised domain adaptive fundus image segmentation with category-level regularization. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, pages 497–506. Springer, 2022. 1

[6] Wei Feng, Meihan Duan, Bingjie Wang, Yu Du, Yiran Zhao, Bin Wang, Lin Zhao, Zongyuan Ge, and Yuntao Hu. Automated segmentation of choroidal neovascularization on optical coherence tomography angiography images of neovascular age-related macular degeneration patients based on deep learning. *Journal of Big Data*, 10(1):111, 2023.

[7] Wei Feng, Lie Ju, Lin Wang, Kaimin Song, and Zongyuan Ge. Towards novel class discovery: A study in novel skin lesions clustering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–33. Springer, 2023. 1

[8] Wei Feng, Lie Ju, Lin Wang, Kaimin Song, Xin Zhao, and Zongyuan Ge. Unsupervised domain adaptation for medical image segmentation by selective entropy constraints and adaptive semantic alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 623–631, 2023.

[9] Wei Feng, Qiujing Huang, Tong Ma, Lie Ju, Zongyuan Ge, Yuzhong Chen, and Peiquan Zhao. Development and validation of a semi-supervised deep learning model for automatic retinopathy of prematurity staging. *Iscience*, 27(1), 2024.

[10] Wei Feng, Duong Nhu, Alison Anderson, Daniel Thom, Sarah N Barnard, Robert Zeibich, Emma Foster, Mark Howard, Susannah T Bellows, Rosemary Burgess, et al. Integrative deep learning of genomic and clinical data for predicting treatment response in newly diagnosed epilepsy. *Neurology*, 105(10):e214315, 2025. 1

[11] Wei Feng, Sijin Zhou, Yiwen Jiang, and Zongyuan Ge. Prism: Progressive robust learning for open-world continual category discovery. In *The Fourteenth International Conference on Learning Representations*, 2026. 1

[12] Enrico Fini, Enver Sangineto, Stéphane Lathuilière, Zhun Zhong, Moin Nabi, and Elisa Ricci. A unified objective for novel class discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9284–9292, 2021. 2, 6

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2

[14] Pascal Germain, Amaury Habrard, François Laviolette, and Emilie Morvant. A new pac-bayesian perspective on domain adaptation. In *International conference on machine learning*, pages 859–868. PMLR, 2016. 4

[15] Peiyan Gu, Chuyu Zhang, Ruijie Xu, and Xuming He. Class-relation knowledge distillation for novel class discovery. *lamp*, 12(15.0):17–5, 2023. 2

[16] Haris Hakeem, Wei Feng, Zhibin Chen, Jiun Choong, Martin J Brodie, Si-Lei Fong, Kheng-Seang Lim, Junhong Wu, Xuefeng Wang, Nicholas Lawn, et al. Development and validation of a deep learning model for predicting treatment response in patients with newly diagnosed epilepsy. *JAMA neurology*, 79(10):986–996, 2022. 1

[17] Kai Han, Sylvestre-Alvise Rebuffi, Sebastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Autonovel: Automatically discovering and learning novel visual categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6767–6781, 2021. 1, 2, 6

[18] Wenlong Hang, Wei Feng, Shuang Liang, Lequan Yu, Qiong Wang, Kup-Sze Choi, and Jing Qin. Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23*, pages 562–571. Springer, 2020. 1

[19] Yiwen Jiang, Deval Mehta, Wei Feng, and Zongyuan Ge. Enhancing interpretable image classification through llm agents and conditional concept bottleneck models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12285–12297, 2025.

[20] Lie Ju, Yicheng Wu, Wei Feng, Zhen Yu, Lin Wang, Zhuoting Zhu, and Zongyuan Ge. Universal semi-supervised learning for medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 355–365. Springer, 2024. 1

[21] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015. 4

[22] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 6

[23] Alexandre Lacasse, François Laviolette, Mario Marchand, Pascal Germain, and Nicolas Usunier. Pac-bayes bounds for the risk of the majority vote and the variance of the gibbs classifier. *Advances in Neural information processing systems*, 19, 2006. 4

[24] Wenbin Li, Zhichen Fan, Jing Huo, and Yang Gao. Modeling inter-class and intra-class constraints in novel class discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3449–3458, 2023. 2

[25] Duo Liu, Zhiquan Tan, Linglan Zhao, Zhongqiang Zhang, Xiangzhong Fang, and Weiran Huang. Generalized category discovery via reciprocal learning and class-wise distribution regularization. *arXiv preprint arXiv:2506.02334*, 2025. 1, 6

[26] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017. 2

[27] Chaochao Lu, Yuhuai Wu, José Miguel Hernández-Lobato, and Bernhard Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021. 2

[28] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8046–8056, 2022. 1, 2

[29] Yanbiao Ma, Wei Dai, Bowei Liu, Jiayi Chen, Wenke Huang, Guancheng Wan, Zhiwu Lu, and Junchi Yan. Calibrating biased distribution in vfm-derived latent space via cross-domain geometric consistency. *arXiv preprint arXiv:2508.13518*, 2025. 5

[30] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 2, 6

[31] Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In *European Conference on Computer Vision*, pages 38–56. Springer, 2022. 7

[32] Sai Bhargav Rongali, Sarthak Mehrotra, Ankit Jha, Shirsha Bose, Tanisha Gupta, Mainak Singha, Biplab Banerjee, et al. Cdad-net: Bridging domain gaps in generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2616–2626, 2024. 1, 2, 6

[33] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732, 2018. 2

[34] Swami Sankaranarayanan, Yogesh Balaji, Carlos D Castillo, and Rama Chellappa. Generate to adapt: Aligning domains using generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8503–8512, 2018. 2

[35] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012. 3

[36] Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *Computer vision–ECCV 2016 workshops: Amsterdam, the Netherlands, October 8-10 and 15-16, 2016, proceedings, part III 14*, pages 443–450. Springer, 2016. 2

[37] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7492–7501, 2022. 1, 2, 6

[38] Hongjun Wang, Sagar Vaze, and Kai Han. Hilo: A learning framework for generalized category discovery robust to domain shifts. *arXiv preprint arXiv:2408.04591*, 2024. 1, 2, 6

[39] Hongjun Wang, Sagar Vaze, and Kai Han. Sptnet: An efficient alternative framework for generalized category discovery with spatial prompt tuning. *arXiv preprint arXiv:2403.13684*, 2024. 1, 2, 6

[40] Kaihong Wang, Donghyun Kim, Rogerio Feris, and Margrit Betke. Cdac: Cross-domain attention consistency in transformer for domain adaptive semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11519–11529, 2023. 2

[41] Xin Wen, Bingchen Zhao, and Xiaojuan Qi. Parametric classification for generalized category discovery: A baseline study. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16590–16600, 2023. 1, 2, 6

[42] Jinyu Yang, Jingjing Liu, Ning Xu, and Junzhou Huang. Tvt: Transferable vision transformer for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 520–530, 2023. 2

[43] Mengyue Yang, Zhen Fang, Yonggang Zhang, Yali Du, Furui Liu, Jean-Francois Ton, Jianhong Wang, and Jun Wang. Invariant learning via probability of sufficient and necessary causes. *Advances in Neural Information Processing Systems*, 36:79832–79857, 2023. 2, 5

[44] Jinze Yu, Jiaming Liu, Xiaobao Wei, Haoyi Zhou, Yohei Nakata, Denis Gudovskiy, Tomoyuki Okuno, Jianxin Li, Kurt Keutzer, and Shanghang Zhang. Mttrans: Cross-domain object detection with mean teacher transformer. In *European Conference on Computer Vision*, pages 629–645. Springer, 2022. 2

[45] Haonan Yuan, Qingyun Sun, Xingcheng Fu, Ziwei Zhang, Cheng Ji, Hao Peng, and Jianxin Li. Environment-aware dynamic graph learning for out-of-distribution generalization. *Advances in Neural Information Processing Systems*, 36:49715–49747, 2023. 2

[46] Bingchen Zhao and Kai Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. *Advances in Neural Information Processing Systems*, 34:22982–22994, 2021. 2

[47] Jinjing Zhu, Haotian Bai, and Lin Wang. Patch-mix transformer for unsupervised domain adaptation: A game perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3561–3571, 2023. 2