CDAD-NET: Bridging Domain Gaps in Generalized Category Discovery

Sai Bhargav Rongali[1*]   Sarthak Mehrotra[1*]   Ankit Jha [2]   Mohamad Hassan N C[1]

Shirsha Bose[3]   Tanisha Gupta[1]   Mainak Singha[1]   Biplab Banerjee[1]

[1]Indian Institute of Technology Bombay, India

[2]INRIA, Grenoble, France [3]Technical University of Munich, Germany

{rongalisaibhargav002, sarthak2002.mehrotra, ankitjha16, mohdhassannc, shirshabosecs,
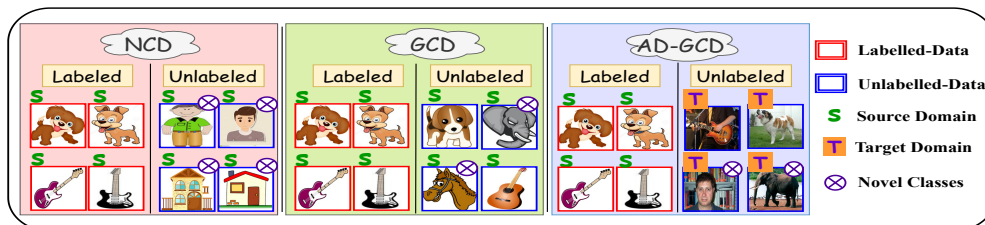tanishagupta072, mainaksingha.iitb, getbiplab}@gmail.com

Figure 1. We introduce the problem setting of **across domain generalized category discovery** (AD-GCD), which is different from the traditional GCD setting [53] in that we consider the labeled and unlabeled data to arise from different data distributions. We call the labeled data the `Source` domain and the unlabeled data the `Target` domain, respectively.

Abstract

In Generalized Category Discovery (GCD), we cluster unlabeled samples of known and novel classes, leveraging a training dataset of known classes. A salient challenge arises due to domain shifts between these datasets. To address this, we present a novel setting: Across Domain Generalized Category Discovery (AD-GCD) and bring forth CDAD-NET *(Class Discoverer Across Domains) as a remedy.* CDAD-NET *is architected to synchronize potential known class samples across both the labeled (source) and unlabeled (target) datasets, while emphasizing the distinct categorization of the target data. To facilitate this, we propose an entropy-driven adversarial learning strategy that accounts for the distance distributions of target samples relative to source-domain class prototypes. Parallelly, the discriminative nature of the shared space is upheld through a fusion of three metric learning objectives. In the source domain, our focus is on refining the proximity between samples and their affiliated class prototypes, while in the target domain, we integrate a neighborhood-centric contrastive learning mechanism, enriched with an adept neighbors-mining approach. To further accentuate the nuanced feature interrelation among semantically aligned images, we champion the concept of conditional image inpainting, underscoring the premise that semantically analogous images prove more efficacious to the task than their disjointed counterparts. Experimentally,* CDAD-NET *eclipses existing lit-erature with a performance increment of 8-15% on three AD-GCD benchmarks we present.*

1. Introduction

Recent times have witnessed the rise of deep learning techniques for visual inference [24, 26, 46]. However, their success predominantly relies on learning from ample supervision, which can be costly at times. One of the solutions is semi-supervised learning [52], which entails learning from both labeled and unlabeled sample sets within a closed-set framework, where both sets encompass identical categories. A more meaningful approach would be not to constrain the classes in the unlabeled set. Accordingly, the concept of novel category discovery (NCD) has recently been introduced [17, 61] that aims to semantically group unlabeled data of unseen categories leveraging supervised knowledge from a distinct set of known categories. However, NCD assumes that all unlabeled instances belong to novel categories with respect to the labeled data, which is often impractical. Vaze *et al*. [53] extended NCD to the GCD setting to address this limitation and allowed unlabeled images to originate from both known and novel classes.

A prevailing assumption in extant GCD methods [53, 58] is that both the labeled and unlabeled datasets uniformly adhere to identical distributions. This assumption facilitates the smooth transfer of discriminative knowledge from labeled to unlabeled sets. Yet, in the real-world context, domain shifts are commonplace. Such shifts can result in the

---
*Equal Contribution

labeled and unlabeled data stemming from markedly different visual domains, impeding the seamless knowledge propagation between the datasets.

Building upon these discussions, this paper introduces a previously unexplored problem setting known as AD-GCD, summarized in Fig. 1. In this context, we designate the domain possessing labeled data as the *source* domain, while its unlabeled counterpart, possibly emanating from disparate data distributions and containing samples from both known and novel classes, is referred to as the *target* domain. Employing a conventional GCD approach may fall short in addressing AD-GCD challenges, as the domain discrepancies could result in inaccurate mappings of novel-class samples from the target to known classes in the source domain, thereby detrimentally impacting performance. Moreover, although augmenting a GCD solution with open-set domain adaptation (OSDA) [44, 48] could enhance the mapping of known-class samples between domains, OSDA tends to hamper the cluster topology of the novel-class samples. The closest existing work to ours is [56], which focuses on clustering target-domain novel-class samples into a predetermined number of known clusters and classifying potential target samples that belong to training classes, using a classifier trained on the source domain. Conversely, our AD-GCD framework endeavors to cluster the target domain accurately without prior knowledge of cluster numbers, positioning our approach as orthogonal to the current relevant body of literature.

AD-GCD has many potential real-world applications. For example, autonomous vehicles navigating diverse environments, often confront unfamiliar objects or road conditions outside their training data. Initially, these models are trained using either synthetic data or a limited set of real-world data, while the actual deployment scenarios may vary significantly. AD-GCD plays a pivotal role in this case by aiding in categorising these unfamiliar elements and recognising known events, enabling more effective handling of these situations.

Consequently, we assert two principal objectives for successfully addressing the challenges of AD-GCD: **i)** Achieve alignment of the known set of classes across domains to felicitate discriminative knowledge transfer from the source to the target domain. **ii)** Ensure that the domains remain discriminative with respect to both known and novel classes.

The proposed solution: We present CDAD-NET, a novel framework to tackle AD-GCD. Towards solving our primary objective, we advocate for a cross-domain alignment, deploying a novel entropy optimization strategy. This is anchored on assessing the similarity distributions of target samples vis-à-vis the class prototypes of the source domain. Underpinning our approach is the theory that target samples aligning with the source domain would show a decline in entropy of the aforesaid distribution, each trending towards a particular prototype. Conversely, samples representing novel classes are projected to display increased entropy of the distribution, highlighting their departure from established class boundaries. Clearly, our approach takes a different path from traditional OSDA practices. Rather than funneling target novel-class samples into one cohesive cluster, our technique deliberately sidesteps this convention.

To achieve our secondary objective of amplifying the distinctiveness within the shared embedding space, we scrutinize existing techniques. Established GCD methodologies [43, 53] advocate for an augmentation-based contrastive self-supervision method like [9] across both labeled and unlabeled data for representation learning. Nonetheless, as reported in [29], such techniques work well for a single domain, but are unfavorable for multi-domain data.

This drives our pivot towards accentuating domain-specific discriminative features, where we propose to harness neighborhood-centric contrastive learning. Specifically for the unlabeled target domain, we offer to reuse the aforementioned distance dynamics interlinking target samples and class prototypes from the source domain to obtain the positive and negative neighbors per reference image.

Our existing global image embedding-based losses might inadequately capture local image variations per category for different visual domains. To enhance these embeddings with detailed, local attributes for better clustering, we introduce the task of *conditional image inpainting*. In this approach, we base the patch-level reconstruction of an image on another complete image. We hypothesize that a conditioning image with semantically similar traits to the reference will yield superior results compared to one with divergent semantics. This concept is operationalized using a newly developed loss function.

The culmination is a meticulously aligned and discriminative embedding space, paving the way for superior cluster assignment of target domain samples via semi-supervised K-means [37], guided by the aligned source domain. Our significant **contributions** are, therefore,

[-] We present a pragmatic problem framework known as AD-GCD, which involves utilizing labeled and unlabeled samples from distinct data distributions in a GCD setup.

[-] Our solution, CDAD-NET, introduces three key innovations. Our unique domain alignment approach maintains the intrinsic clustering of the target domain. For feature discriminability, we unveil a neighborhood-focused contrastive self-supervision for the target domain. Lastly, we propose a conditional image inpainting task to enhance fine-grained feature association within clusters.

[-] We establish the experimental setup for AD-GCD on three datasets, and thoroughly analyze CDAD-NET both for cross-domain and in-domain GCD tasks.

2. Related Works

Class discovery: Introduced by Han *et al.* [23], NCD focuses on categorizing unlabeled samples from new categories by building upon the knowledge gained from a predefined set of categories. Early attempts to tackle the NCD challenge include [27, 28], both employing two models trained on labeled and unlabeled data separately to facilitate general task transfer learning. There exists a number of follow-up works, including [17, 21, 59, 61], to name a few, which consider mostly parametric classification heads on top of a generic feature extractor for solving NCD.

The GCD task [53] extends beyond NCD by incorporating unlabeled data from both known and novel classes. [5] tackled this by developing an adaptive margin loss to balance intra-class variances, while [53] employed pre-trained Vision Transformers (ViTs) [13] and a modified K-Means algorithm for a semi-supervised approach. Further innovations by [58] and [43] introduced contrastive learning with visual prompts and self-supervised learning strategies to enhance GCD. Meanwhile, [60] utilized Gaussian mixture models for clustering, while [10] introduced an information-theoretic solution to GCD. Compared to these non-parametric clustering-based approaches, [55] analyzed the parametric methods for GCD. However, all these endeavors grapple with covariate shifts between datasets.

Cross-domain learning: Navigating multiple visual domains poses significant challenges, primarily due to appearance drifts across these domains [3, 45]. There are two key strategies for handling multi-domain data effectively. The first, Domain Adaptation (DA) [1, 11, 34, 36], aims to align a source domain with a target domain in a transductive learning manner. The second, Domain Generalization (DG) [20, 33, 47], focuses on building a robust and adaptable model applicable to novel target domains during inference. Meanwhile, multi-domain learning [3, 35, 45] aims to develop a model that leverages labeled data from various domains without letting domain-specific features negatively impact the learning process. Both DA and DG are applicable in closed-set and open-set scenarios [2, 4, 40, 48, 50]. Open-set DA/DG aims to categorize novel target samples into a singular, undefined class. However, in AD-GCD, the emphasis shifts to discerning and organizing the semantic structures within these unlabeled datasets, given partial supervision from a distinct visual domain.

Self-supervised learning: Self-supervision embodies the concept of deriving meaningful representations directly from data, bypassing the reliance on external semantic labels. These techniques can be broadly classified into contrastive learning-based and more conventional approaches. In the realm of traditional self-supervised models, a pretext task is established, such as solving a jigsaw puzzle, predicting image rotations, or performing image inpainting [8, 38, 41, 57], among others. These pretext tasks impart

visual commonsense to the model.

Numerous studies have delved into the efficacy of contrastive loss in the realm of unsupervised representation learning. This exploration includes seminal works such as InfoNCE [39], SimCLR [9], SWaV [6], MoCo [25], SEER [19], to name a few. Here, the idea is to maximize compatibility between different augmentations of the same image while minimizing the same for distinct images. In place of crafting augmented views to create pairs, [14] proposed to consider the neighbors. Finally, the notion of multi-domain self-supervised learning is less studied, and existing few methods like [16, 29] focus on the closed-set scenario.

Existing GCD models generally use a semi-supervised contrastive learning from labeled and unlabeled data for crafting representations. However, this approach falls short for divergent data distributions, leading us to focus on domain-specific metric objectives. Additionally, we propose to bolster the global image representations by ensuring that they accommodate the local image properties well, through the proposed novel conditional inpainting task.

3. CDAD-NET: Proposed Methodology

Consider the dataset $\mathcal{D}$, which encompasses a combination of labeled samples (source domain) denoted as $\mathcal{D}_{\mathcal{L}}$ and unlabeled samples (target domain) referred to as $\mathcal{D}_{\mathcal{U}}$. To elaborate further, $\mathcal{D}_{\mathcal{L}}$ comprises pairs of the form $\{(x_i, y_i)\}_{i=1}^{n_l} \in \mathcal{X} \times \mathcal{Y}_{\mathcal{L}}$, while $\mathcal{D}_{\mathcal{U}}$ encompasses pairs $\{(x_j, y_j)\}_{j=1}^{n_u} \in \mathcal{X} \times \mathcal{Y}_{\mathcal{U}}$, with the constraint that $\mathcal{Y}_{\mathcal{L}} \subset \mathcal{Y}_{\mathcal{U}}$. $\mathcal{Y}_{\mathcal{L}}$ corresponds to the set of labels specifically assigned to the known classes, denoted as $\mathcal{C}_{kwn}$, whereas $\mathcal{Y}_{\mathcal{U}}$ represents the complete label set, encompassing both the known classes and the novel classes $\mathcal{C}_{new}$, denoted collectively as $\mathcal{C}$. As opposed to the original GCD formulation [53], we introduce a new constraint in our proposed AD-GCD setting: $\mathcal{P}(\mathcal{D}_{\mathcal{L}}) \neq \mathcal{P}(\mathcal{D}_{\mathcal{U}})$, which means the domain characteristics of $\mathcal{D}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{U}}$ are different.

During the training phase, our model solely has access to the labeled data contained in $\mathcal{D}_{\mathcal{L}}$, while the labels of $\mathcal{D}_{\mathcal{U}}$ are not accessible. Our main goal is to perform accurate clustering of the data samples in $\mathcal{D}_{\mathcal{U}}$ during inference.

3.1. Model architecture and training details

Learning objectives: As outlined previously, our objective is to address two fundamental challenges of AD-GCD: **i)** How to calculate a shared embedding space using the source domain $\mathcal{D}_{\mathcal{L}}$ and the target domain $\mathcal{D}_{\mathcal{U}}$? In this space, we aim to map the known-class samples within the label space $\mathcal{C}_{kwn}$ from $\mathcal{D}_{\mathcal{U}}$ to coincide with $\mathcal{D}_{\mathcal{L}}$, while simultaneously ensuring that novel-class samples from $\mathcal{D}_{\mathcal{U}}$ are positioned distinctively. **ii)** How do we ensure the class/cluster discriminativeness for both domains in this space? We seek to optimize the within-class compactness while maximizing the between-class separation given $\mathcal{D}_{\mathcal{L}} \cup \mathcal{D}_{\mathcal{U}}$. Following the
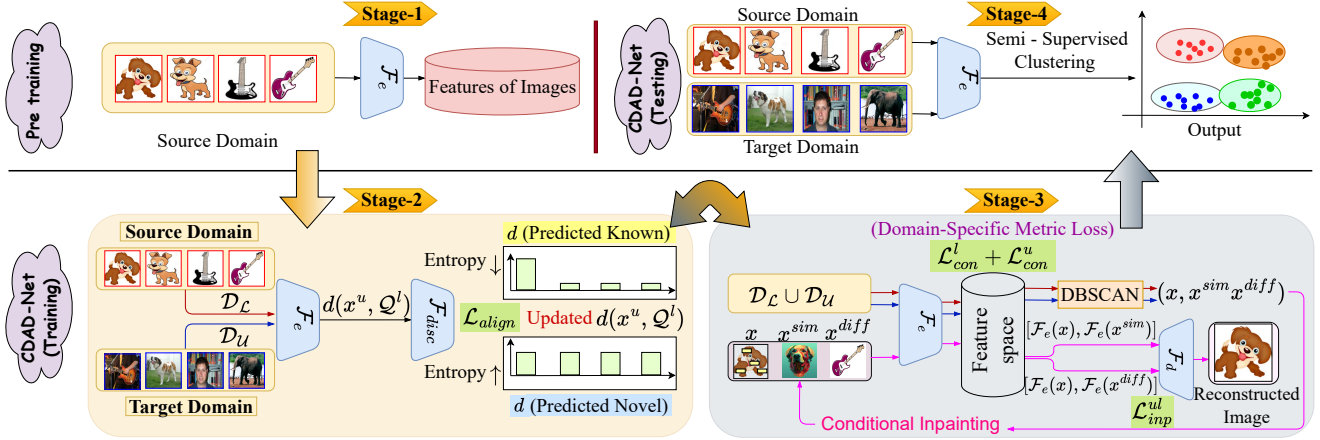
Figure 2. **Architecture overview and training pipeline of CDAD-NET**. The DINO pre-trained ViT encoder $\mathcal{F}_e$ is first fine-tuned on $\mathcal{D}_{\mathcal{L}}$. Subsequently, the model is trained on $\mathcal{D}_{\mathcal{L}} \cup \mathcal{D}_{\mathcal{U}}$ in two stages in each training epoch, using the domain alignment objective $\mathcal{L}_{align}$ given $(\mathcal{F}_e, \mathcal{F}_{disc})$, and the cumulative metric objective $\mathcal{L}_{con}^l + \mathcal{L}_{con}^u + \mathcal{L}_{inp}^{ul}$ given $(\mathcal{F}_e, \mathcal{F}_d)$. Inference is carried our by semi-supervised K-means applied on $\mathcal{F}_e(\mathcal{D}_{\mathcal{L}}) \cup \mathcal{F}_e(\mathcal{D}_{\mathcal{U}})$. The number of target domain clusters is estimated using the `elbow` method.

establishment of this shared representation space, we proceed by employing semi-supervised K-Means to cluster the samples contained in $\mathcal{D}_{\mathcal{L}} \cup \mathcal{D}_{\mathcal{U}}$ while ensuring that the labeled data from $\mathcal{D}_{\mathcal{L}}$ and the confident pseudo-labeled data from $\mathcal{D}_{\mathcal{U}}$ preserve their cluster assignments.

Model architectural overview: Building on the insights from [53], we adopt the ViT with a B-16 backbone that has been pre-trained using the DINO objective [7] on the ImageNet dataset [12], as it provides a robust nearest-neighbor classifier, which is deemed to enhance our clustering performance. We denote this pre-trained model as $\mathcal{F}_e$ and use the CLS token for the image embeddings. As we advance further in our approach, we introduce a discriminator denoted as $\mathcal{F}_{disc}$, which operates on the features extracted by $\mathcal{F}_e$ to oversee the domain alignment objective. In addition to this, we incorporate a decoder network $\mathcal{F}_d$ to aid in the conditional image inpainting task at the patch level.

$\mathcal{F}_d$ denotes the module responsible for inpainting, enacted atop $\mathcal{F}_e$. We infuse the feature encodings from $\mathcal{F}_e$ of the reference image together with that of the companion image without any missing patches to provide the context for reconstruction at different levels of $\mathcal{F}_d$.

Walking through the training of CDAD-NET (Fig. 2):

- **Fine-tuning $\mathcal{F}_e$ on $\mathcal{D}_{\mathcal{L}}$:** At the beginning, we perform a warm-up supervised fine-tuning of $\mathcal{F}_e$ on $\mathcal{D}_{\mathcal{L}}$.

- **Proposed alignment of $\mathcal{D}_{\mathcal{L}}$ and $\mathcal{D}_{\mathcal{U}}$:** Given the feature extractor $\mathcal{F}_e$ and the domain discriminator $\mathcal{F}_{disc}$, we propose an entropy-based adversarial learning technique to align the domains, solving our first objective. To elucidate, let's denote $\mathcal{Q}^l = \{\mathcal{Q}_k^l\}_{k=1}^{|\mathcal{C}_{kwn}|}$ as the class prototypes for the source domain classes within $\mathcal{D}_{\mathcal{L}}$. Each $\mathcal{Q}_k^l$ is obtained by averaging the image embeddings given the labeled samples from the kth category. Now, given a sample x^u from $\mathcal{D}_{\mathcal{U}}$, we compute the normalized distance distribution of x^u with

respect to $\mathcal{Q}^l$, $d(x^u, \mathcal{Q}^l)$, using the cosine similarity (δ).

$$d(x^u, \mathcal{Q}^l) = \text{Normalize}[\delta(\mathcal{F}_e(x^u), \mathcal{Q}_1^l), \cdots ,$$
$$\delta(\mathcal{F}_e(x^u), \mathcal{Q}_{|\mathcal{C}_{kwn}|}^l)] \quad (1)$$

Under this setting, our proposed domain alignment objective focuses on the entropy of $d(x^u, \mathcal{Q}^l)$. In essence, when x^u belongs to one of the known classes in $\mathcal{C}_{kwn}$, we anticipate $d(x^u, \mathcal{Q}^l)$ to exhibit a one-hot type distribution, leading to low entropy. Conversely, suppose x^u originates from one of the novel classes in $\mathcal{C}_{new}$. In that case, we aim for the entropy of $d(x^u, \mathcal{Q}^l)$ to be high, indicating that the sample is distant from the prototypes within $\mathcal{Q}^l$. We represent the corresponding loss objective as $\mathcal{L}_{align}$.

Existing adversarial OSDA techniques, such as [48], utilize a threshold probability of 0.5 to distinguish unknown samples, typically herding them into a single class label —a practice our proposed method circumvents. Additionally, our feature discrimination tactic, outlined subsequently, enhances cluster density by leveraging $d(x^u, \mathcal{Q}^l)$.

- **Enforcing discriminativeness objectives:** We introduce two distinct contrastive objectives for both domains alongside a novel image inpainting formulation to improve the cluster density of the learned embeddings. In the source domain, our primary aim is to bring the embeddings of each x^l closer to their respective class prototypes from $\mathcal{Q}^l$ while simultaneously pushing them away from prototypes associated with negative classes. This specific loss is $\mathcal{L}_{con}^l$. However, achieving this in the target domain presents a significant challenge due to the absence of direct supervision. As a remedy, our method focuses on enhancing contrastive learning by considering near and far neighbors relative to a reference image. To address this, we introduce a technique that leverages the relationships between distances $d(x_1^u, \mathcal{Q}^l)$ and $d(x_2^u, \mathcal{Q}^l)$. Specifically, if x_1^u and x_2^u are similar, they

are deemed to show identical relation with the source domain prototypes. Accordingly, within a batch, we define a pair (x_1^u, x_2^u) as positive if the Manhattan distance between $d(x_1^u, \mathcal{Q}^l)$ and $d(x_2^u, \mathcal{Q}^l)$ is the smallest among all pairs in the batch. Similarly, we pick $\mathcal{M}$ negative samples based on high Manhattan distance. Subsequently, we apply the contrastive loss on this set, referred to as $\mathcal{L}_{con}^u$. In order to avoid any erroneous positive pairs, we include some image-specific augmentations into our data-pool of the batch.

Moreover, we aim to refine the global CLS embeddings from $\mathcal{F}_e$ to be more responsive to nuanced local image features, a task not directly achieved by $\mathcal{L}_{con}^l + \mathcal{L}_{con}^u$, as they primarily focus to optimize the global image embeddings. To enhance the compactness of our clusters, our strategy is to amplify feature correlations among similar images and reduce them for dissimilar ones, thereby strengthening the clarity and distinction of the clusters. We approach this through the proposed conditional image inpainting task, using an unmasked full image to assist in the patch-level inpainting of a reference image. Precisely, we begin by over-clustering the embeddings obtained from $\mathcal{F}_e(\mathcal{D}_\mathcal{L}) \cup \mathcal{F}_e(\mathcal{D}_\mathcal{U})$ using DBSCAN [15], a density-based non-parametric clustering technique. Within these clusters, we select a quadruplet of samples denoted as $(x, x', x^{sim}, x^{diff})$ where samples in pairs (x, x^{sim}) and (x, x^{diff}) share similar and different cluster labels, respectively. x' is obtained by removing a random patch from x, and we are interested in the inpainting task for x' using $\mathcal{F}_d$. To do this, we enforce $\mathcal{F}_e(x^{sim})$ to provide better assistance than $\mathcal{F}_e(x^{diff})$ since x shares superior semantic similarity to x^{sim} than x^{diff}, realised through a novel loss $\mathcal{L}_{inp}^{ul}$.

3.2. Loss function, training, and inference

$\mathcal{L}_{align}$: **Proposed adversarial domain alignment loss:** To implement the proposed adversarial entropy optimization-based domain alignment method, we introduce the concept of a pseudo-ground-truth for each target domain sample. This pseudo-ground-truth, denoted by y^g, is defined as a uniform distribution across $|\mathcal{C}_{kwn}|$ classes. For potential samples from known classes in $\mathcal{D}_\mathcal{U}$, our objective is to learn features $\mathcal{F}_e(x^u)$ in such a way that we maximize the cross-entropy loss $\mathcal{L}_{ce}$ between y^g and the distribution $d(x^u, \mathcal{Q}^l)$. Conversely, our goal for potential samples from novel classes is to minimize $\mathcal{L}_{ce}$. In this way, we separate the known and novel class samples of $\mathcal{D}_\mathcal{U}$ and align the known class samples of both domains. To achieve our goal, we propose the following loss function:

$$\mathcal{L}_{align} = \min_{\mathcal{F}_e} \max_{\mathcal{F}_{disc}} \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_\mathcal{L} \cup \mathcal{D}_\mathcal{U})} \mathcal{L}_{ce}(y^g, d(x^u, \mathcal{Q}^l)) \quad (2)$$

$\mathcal{L}_{con}^l$: **Source domain supervised contrastive loss:** We employ the traditional distance optimization technique when provided with a source domain sample and all the class prototypes in $\mathcal{Q}^l$. Our objective is to maximize the

similarity between the sample and prototypes of the same class while minimizing it simultaneously for all negative-class prototypes. This results in the following loss objective, given a kth-class sample x^k and class prototype $\mathcal{Q}_k^l$:

$$\mathcal{L}_{con}^l = \min_{\mathcal{F}_e} \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_\mathcal{L})} -\log \frac{\exp(\delta(\mathcal{F}_e(x^k), \mathcal{Q}_k^l))}{\sum_{n=1}^{|\mathcal{C}_{kwn}|} \exp(\delta(\mathcal{F}_e(x^k), \mathcal{Q}_n^l))} \quad (3)$$

$\mathcal{L}_{con}^u$: **Target domain contrastive loss with the proposed neighbors selection strategy:** Aligned with our approach in the source domain, we adopt a domain-specific contrastive loss for the target domain, using neighboring sample relationships. The absence of labels for $\mathcal{D}_\mathcal{U}$ poses a challenge. To overcome this, we utilize $d(x^u, \mathcal{Q}^l)$ to discern positive from negative target sample pairings for contrastive loss. We strive to enhance the similarity for positive pairs and reduce the same for negative pairs. With an anchor, a positive, and a collective of negative samples, denoted as $(x^u, x^{+u}, \{x_m^{-u}\}_{m=1}^{\mathcal{M}})$, we define $\mathcal{L}_{con}^u$ using the conventional InfoNCE [39] based contrastive objective.

$\mathcal{L}_{inp}^{ul}$: **Proposed conditional image inpainting loss:** To consolidate our hypothesis that $\mathcal{F}_e(x^{sim})$ should aid in better patch reconstruction of x' than $\mathcal{F}_e(x^{diff})$, we introduce the following novel loss objective, where $\mathcal{L}_{recon}$ defines the inpainting loss. The loss has two components. In order to train $\mathcal{F}_d$ to learn inpainting, we aim to minimize $\mathcal{L}_{recon}(x'|x)$, *i.e.* we condition $\mathcal{F}_d$ on the original reference image itself. Besides, we introduce a hinge objective that states if $\mathcal{L}_{recon}(x'|x^{sim}) - \mathcal{L}_{recon}(x'|x^{diff}) \le 0$, the loss component should return zero, else we seek to minimize the obtained error. As a whole, the combined loss is expected to attain the value zero at optimality, strengthening the fine-grained correlation between x and x^{sim}.

$$\mathcal{L}_{inp}^{ul} = \min_{\mathcal{F}_e, \mathcal{F}_d} \mathop{\mathbb{E}}_{\mathcal{P}(\mathcal{D}_\mathcal{L} \cup \mathcal{D}_\mathcal{U})} \Big[\max(0, \mathcal{L}_{recon}(x'|x^{sim}) - \\ \mathcal{L}_{recon}(x'|x^{diff})) \Big] + \mathcal{L}_{recon}(x'|x) \quad (4)$$

Training: In CDAD-NET, we adopt an iterative optimization technique. First, the warm-up supervised fine-tuning is performed on $\mathcal{F}_e$. Subsequently, each training epoch alternates between $\mathcal{L}_{align}$ for domain alignment and a composite loss $\mathcal{L}_{con}^l + \mathcal{L}_{con}^u + \mathcal{L}_{inp}^{ul}$ for contrastive learning and inpainting. This balanced optimization ensures both domain independence and discriminability in the CLS tokens of $\mathcal{F}_e$.

Inference: Consistent with [53], we employ $\mathcal{F}_e$'s feature embeddings for applying semi-supervised K-means clustering on $\mathcal{D}_\mathcal{L} \cup \mathcal{D}_\mathcal{U}$. We ensure accurate categorization for $\mathcal{D}_\mathcal{L}$ samples and apply this rigor to confidently pseudo-labeled $\mathcal{D}_\mathcal{U}$ samples, preserving cluster consistency and merging supervised insights with clustering. The confidence is calculated based on a sample's cosine similarity to the cluster centre for the known classes. We set the threshold to be

0.9 to avoid any possible misclassification. For estimating optimal cluster count, we utilize Brent's algorithm [53].

4. Experimental Evaluations

Datasets: We run our experiments using six prominent benchmark datasets from the domain adaptation and GCD literature to analyze the cross-domain and within-domain performance of CDAD-NET. Specifically, for AD-GCD, they include: Office-Home [54], PACS [33], and Domain-Net [42], respectively. We consider all the pairs of domains to define the source and the target and report the average scores for Office-Home and PACS. For DomainNet, we consider the source-target pairs as listed below:

1. Real world → Sketch
2. Painting → Real World
3. Sketch → Clip Art
4. Sketch → Painting
5. Quickdraw → Real World
6. Sketch → Quickdraw
7. Painting → Quickdraw
8. Painting → Infograph
9. Real World → Clip Art

Also, we consider $|\mathcal{C}_{kwn}| : |\mathcal{C}_{new}|$ to be $40 : 25$, $250 : 95$, and $4 : 3$ for Office-Home, DomainNet, and PACS, respectively. Besides, we show the within-domain class-discovery performance of CDAD-NET on the following datasets: CIFAR10-100 [31] and ImageNet100 [12], using the established protocols [53].

Training and implementation details: In our experimental setup, we initialize $\mathcal{F}_e$ with the pretrained weights of the ViT-B-16 backbone from DINO [7]. Subsequently, we fine-tune the last block of $\mathcal{F}_e$ with an initial learning rate of 0.01. This learning rate follows a cosine annealed schedule, with $\mathcal{D}_{\mathcal{L}}$ serving as the reference dataset, as outlined in [53]. To implement the contrastive paradigms, we adhere to the standard practice of using a non-linear projector on top of $\mathcal{F}_e$. On the other hand, $\mathcal{F}_{disc}$ employs an MLP based architecture. We incorporate a random patch masking technique to mask one of the 16 patches in the original image for the conditional inpainting task. $\mathcal{F}_d$ consists of deconvolution blocks to increase the resolution of the feature maps. When calculating $\mathcal{L}_{con}^u$, we consider the number of negative samples $\mathcal{M}$ to be 20. Regarding DBSCAN clustering, we set the radius value to 1. This choice intentionally encourages over-clustering, aligning with our goal of creating semantically well-tied local clusters for selecting the triplets. When estimating the value of K for semi-supervised K-means, we adopt an iterative approach in which we consider K within the range $[|\mathcal{C}_{kwn}|, 1000]$ and employ Brent's algorithm. Finally, we perform warm-up training for 30 iterations, followed by main training for 500 iterations, with a learning rate of 0.01 using the ADAM [30] optimizer.

Evaluation protocols: Our primary evaluation metric, denoted as All, assesses the clustering performance across

the entire dataset $\mathcal{D}_{\mathcal{U}}$. In addition to this global evaluation, we also present results for two specific subsets: Old classes, which consist of instances in $\mathcal{D}_{\mathcal{U}}$ belonging to the $\mathcal{C}_{kwn}$ (known) classes, and New classes, comprising instances in $\mathcal{D}_{\mathcal{U}}$ that belong to the $\mathcal{C}_{new}$ (new) classes. We utilize the Hungarian optimal assignment algorithm [32] to compute these subset-specific accuracies to determine the optimal permutation of labels that maximizes accuracy.

4.1. Comparison to the state-of-the-art

Competitors: We conduct a comparative analysis of CDAD-NET against several existing GCD methods found in the literature. These methods include the following: Baseline K-means [37], ORCA [5], GCD [53], DCCL [43], SimGCD [55], and GPC [60], respectively. Furthermore, to address domain shifts effectively, we explore extensions of the GCD, GPC, SimGCD, and DCCL methods where we integrate these techniques with the off-the-shelf OSDA method, OSDA-BP, which is based on adversarial learning [48]. We adapt the available codebase, and implement DCCL on our own. While integrating OSDA, we follow an alternate optimization strategy between the OSDA and the main objectives. We consider the default parameterization for all these methods.

Discussions on cross-domain GCD performance: Table 1 presents a comprehensive comparison of AD-GCD performance across the Office-Home, DomainNet, and PACS datasets when dealing with domains stemming from different data distributions. CDAD-NET consistently outperforms other methods for all domain pairs within these datasets. We performed the experiments on all the domain pairs in OfficeHome and PACS datasets and took the average across all the pairs.

In the case of the Office-Home, CDAD-NET achieves an impressive All accuracy of 67.55%, surpassing DCCL by 25.83%, SimGCD by 11.88%, GPC by 25.40%, and GCD by 17.32%. For the DomainNet , CDAD-NET attains an All accuracy of 70.28%, outperforming GCD by 17.7%, SimGCD by 16.01%, and GPC by 31%. Finally, in the PACS, CDAD-NET showcases its prowess with an All accuracy of 83.25%, surpassing GCD by 26.54%, SimGCD by 25.47%, and GPC by 46.42%, respectively. Moreover, the inclusion of OSDA-BP [48] did not prove beneficial when integrated with the existing models, given the inherent bias of the OSDA model to cluster the open samples. In this regard, SimGCD+OSDA-BP shows the best performance among the competitors but lags CDAD-NET by $3-18\%$ in All values. In contrast, CDAD-NET effectively combines domain alignment with cluster topology preservation for novel classes, leading to improved clustering performance. Qualitatively, we showcase the t-SNE [51] visualizations in Fig. 3, confirming that CDAD-NET can produce the most discriminative target domain clustering.

Table 1. Comparison of proposed CDAD-NET on AD-GCD with the referred literature (highest in green, second highest in red and **bold** is the proposed method)

| Methods	Office-Home			DomainNet			PACS		
	All	Old	New	All	Old	New	All	Old	New
K-Means [37]	29.48	32.58	24.82	24.73	30.67	22.52	25.46	28.12	19.03
ORCA [5] (ICLR'22)	38.23	40.59	29.07	32.58	38.53	30.77	40.62	43.82	35.58
GCD [53] (CVPR'22)	50.32	54.49	45.93	52.58	55.43	46.12	56.71	61.66	49.32
DCCL [43] (CVPR'23)	41.72	44.43	33.37	40.87	43.64	31.39	38.17	42.52	35.45
GPC [60] (ICCV'23)	42.15	43.33	30.67	39.28	44.45	29.72	36.83	44.74	31.63
SimGCD [55] (ICCV'23)	55.67	60.36	49.76	54.27	65.83	41.67	57.78	66.82	48.76
ORCA + OSDA [48]	45.25	47.61	40.35	42.38	45.31	39.36	45.67	47.21	44.37
GPC + OSDA [48]	47.84	50.62	42.16	46.52	49.46	41.84	47.53	49.14	44.84
GCD + OSDA [48]	61.49	65.38	53.84	58.79	66.41	51.28	60.52	67.69	54.50
DCCL + OSDA [48]	52.24	52.34	45.60	52.12	54.52	43.76	58.32	60.25	55.67
SimGCD + OSDA [48]	64.04	68.82	58.25	51.57	52.38	43.69	64.83	69.78	60.74
CDAD-NET	**67.55**	**72.42**	**63.44**	**70.28**	**76.46**	**65.19**	**83.25**	**87.58**	**77.35**
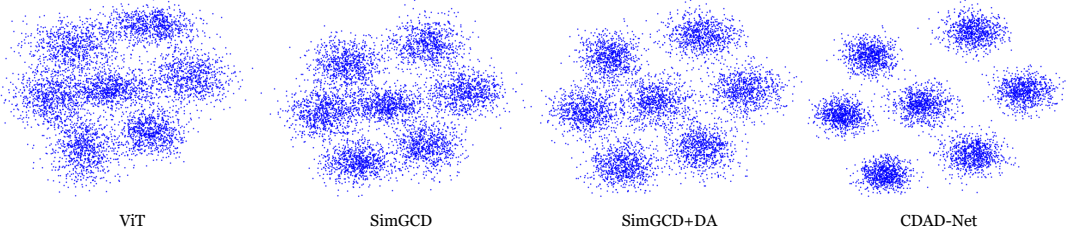
Figure 3. **t-SNE visualizations** of the target domain clusters, as produced by pre-trained ViT, SimGCD [55], SimGCD with OSDA [48], and CDAD-NET for the PACS dataset, clearly demonstrate that CDAD-NET excels in generating the most distinctive embedding space.

Table 2. Comparison of the proposed CDAD-NET to the literature for within-domain GCD task (highest in green, second highest in red and **bold** is the proposed method)

Methods	CIFAR10			CIFAR100			ImageNet-100		
	All	Old	New	All	Old	New	All	Old	New
k-means [37]	83.60	85.70	82.50	52.00	52.20	50.80	72.70	75.50	71.30
RankStats+ [22]	46.80	19.20	60.50	58.20	77.60	19.30	37.10	61.60	24.80
UNO+ [17]	68.60	98.30	53.80	69.50	80.60	47.20	70.30	95.00	57.90
ORCA [5] (ICLR'22)	88.90	88.20	89.20	55.10	65.50	34.40	67.60	90.90	56.00
GCD [53] (CVPR'22)	91.50	97.90	88.20	73.20	76.20	66.53	74.10	89.80	66.30
DCCL [43] (CVPR'23)	96.30	96.50	96.90	75.30	76.80	70.20	80.50	90.50	76.20
CDAD-NET	**96.50**	**97.20**	**96.00**	**75.6**	**76.29**	**70.95**	**80.52**	**80.76**	**80.28**

Table 3. Ablation analysis of the proposed model components of CDAD-NET on the OfficeHome dataset.

Conf.	Fine-tuned ViT	$\mathcal{L}_{con}^{l}$	$\mathcal{L}_{con}^{u}$	$\mathcal{L}_{inp}^{ul}$	$\mathcal{L}_{align}$	Office-Home		
						All	Old	New
C-1	×	×	×	×	×	66.87	69.94	60.48
C-2	✓	×	×	×	×	67.60	68.90	65.35
C-3	✓	✓	×	×	×	72.12	76.64	64.87
C-4	✓	✓	✓	×	×	72.15	76.05	65.29
C-5	✓	✓	✓	✓	×	72.93	76.46	66.61
C-6	✓	✓	✓	✓	✓	**73.10**	**75.83**	**68.40**

Discussions on within-domain GCD performances: We present a performance comparison of CDAD-NET with several existing literature on within-domain GCD datasets, as shown in Table 2. Notably, our method demonstrates significant superiority over competitors, consistently outperforming them in all the metrics, although the competitors are explicitly designed to handle within-domain GCD.

4.2. Ablation analysis

(i) Dissecting the proposed model components: In Table 3, we ablate the performance of various components within CDAD-NET using two domain pairs from the Office-Home dataset, i.e., Art → Real World and Clip Art → Product. It is evident that relying solely on source fine-tuned ViT features

proves to be suboptimal for achieving effective clustering of $\mathcal{D}_{\mathcal{U}}$. However, the inclusion of the contrastive domain-specific objectives shows further improvements by $\approx 5\%$ individually in All, while combining them shows some improvements in the New metric. The use of the proposed inpainting objective further improves the performance by $\approx 1.5\%$ in New metric, signifying its capability to learn more fine-level features into the CLS tokens. The incorporation of $\mathcal{L}_{align}^{ul}$ leads to a notable enhancement, as it assists in domain alignment while not distorting the cluster structure of the classes.. When all the losses are integrated, the model outputs an All value of 73.10%, and New performance of 68.40%.

(ii) Our proposed domain alignment against the existing closed-set and open-set DA strategies: In Table 4, we assess the effectiveness of our entropy optimization-based DA

strategy in comparison to existing closed-set and open-set DA techniques, specifically DANN [18] and OSDA-BP [48] on AD-GCD by integrating these models to CDAD-NET. While closed-set DA models aim to achieve a global alignment between the two domain distributions, open-set DA techniques focus on isolating outliers to prevent them from affecting the alignment of overlapping classes. However, our observations indicate that both approaches fall short in preserving the cluster structure, particularly on novel classes, thereby diminishing the overall clustering performance. In contrast, our method does not constrain the variability of the novel class samples, thus less affecting the clustering performance.

Table 4. Comparative analysis of the proposed domain alignment in the CDAD-NET and the state-of-the-art DA approaches.

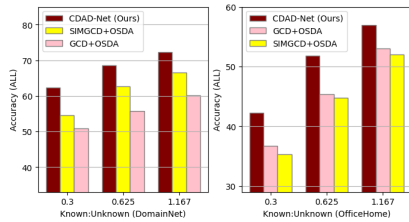| Method	DomainNet			OfficeHome		
	Sketch→Painting			Product→Real World		
	All	Old	New	All	Old	New
DANN [18]	57.36	64.31	53.44	36.25	40.38	25.67
OSDA [48]	56.65	60.57	53.44	28.85	31.22	19.51
CDAD-NET	73.42	76.49	67.70	73.35	79.37	63.08

Figure 4. Openness analysis of CDAD-NET and competitors on Office-Home and Domain-Net datasets.

(iii) Sensitivity of CDAD-NET to the number of novel categories in $\mathcal{D}_{\mathcal{U}}$: We perform a sensitivity analysis of CDAD-NET regarding the number of novel categories in $\mathcal{D}_{\mathcal{U}}$, compared to the known classes overlapped with $\mathcal{D}_{\mathcal{L}}$ (referred to as the openness factor), as depicted in Fig. 4. We evaluate the openness of CDAD-NET using domain pairs from Sketch → Painting and from Product → Real World for the DomainNet and OfficeHome datasets, respectively. For all openness values, CDAD-NET consistently surpasses its competitors. As the number of known classes decreases in $\mathcal{D}_{\mathcal{U}}$, performance tends to decrease as well. However, the decline in accuracy is less pronounced in CDAD-NET, as evident from the plots. In contrast, comparative techniques exhibit a significant drop in performance as the number of novel classes increases in $\mathcal{D}_{\mathcal{U}}$.

(iv) Qualitative visualizations: As outlined in [49], we showcase the GRADCAM outputs to visualize attention maps for diverse images across domains in Fig. 5. The findings indicate that the model adeptly focuses on the dis-
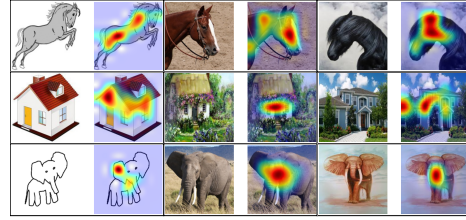

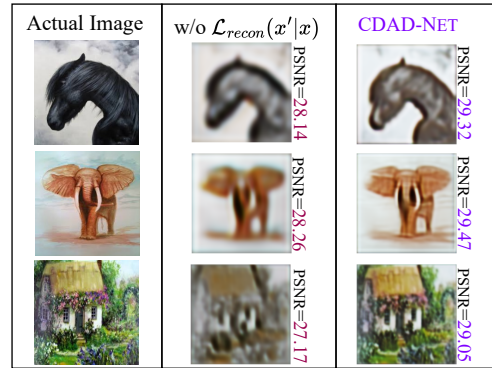Figure 5. The attention maps produced using GRADCAM.

Figure 6. Image inpainting outcomes in the absence of $\mathcal{L}_{recon}(x'|x)$ and the full CDAD-NET.

criminative features crucial for classification. Similarly, we showcase the results of the inpainting with respect to our proposed loss measures in Fig. 6. It is evident from the achieved PSNR scores that the use of the hinge objective aids in better patch reconstruction.

5. Takeaways and Future Scope

In this paper, we propose the challenging new task of AD-GCD, wherein disparate distributions between labeled and unlabeled samples in the GCD setting can thwart the transfer of insights from labeled to unlabeled sets for successful clustering. We tackle this with our CDAD-NET framework, introducing a targeted domain alignment goal that aligns domains while preserving the integrity of cluster structures across all known and emerging categories. Concurrently, we devise a unique domain-centric metric learning approach, employing contrastive methods to bolster discriminability across domains. We further incorporate a conditional image inpainting mechanism to enrich fine-grained feature associations within image embeddings of similar locales. Our empirical evaluations verify CDAD-NET's superior performance in cross-domain and within-domain GCD tasks, laying the groundwork for future applications like multi-camera person re-identification and other cross-domain learning-driven applications.

References

[1] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014. 3

[2] Prathmesh Bele, Valay Bundele, Avigyan Bhattacharya, Ankit Jha, Gemma Roig, and Biplab Banerjee. Learning class and domain augmentations for single-source open-domain generalization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1816–1826, January 2024. 3

[3] Hakan Bilen and Andrea Vedaldi. Universal representations: The missing link between faces, text, planktons, and cat breeds. *arXiv preprint arXiv:1701.07275*, 2017. 3

[4] Shirsha Bose, Ankit Jha, Hitesh Kandala, and Biplab Banerjee. Beyond boundaries: A novel data-augmentation discourse for open domain generalization. *Transactions on Machine Learning Research*, 2023. 3

[5] Kaidi Cao, Maria Brbic, and Jure Leskovec. Open-world semi-supervised learning. *ArXiv*, abs/2102.03526, 2021. 3, 6, 7

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 4, 6

[8] Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International conference on machine learning*, pages 1691–1703. PMLR, 2020. 3

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. 2, 3

[10] Florent Chiaroni, Jose Dolz, Ziko Imtiaz Masud, Amar Mitiche, and Ismail Ben Ayed. Parametric information maximization for generalized category discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1729–1739, 2023. 3

[11] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009. 3

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4, 6

[13] 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. 3

[14] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. With a little help from my friends: Nearest-neighbor contrastive learning of visual representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9588–9597, 2021. 3

[15] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996. 5

[16] Zeyu Feng, Chang Xu, and Dacheng Tao. Self-supervised representation learning from multi-domain data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3245–3255, 2019. 3

[17] 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 (ICCV)*, pages 9284–9292, October 2021. 1, 3, 7

[18] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 8

[19] Priya Goyal, Mathilde Caron, Benjamin Lefaudeux, Min Xu, Pengchao Wang, Vivek Pai, Mannat Singh, Vitaliy Liptchinsky, Ishan Misra, Armand Joulin, et al. Self-supervised pretraining of visual features in the wild. *arXiv preprint arXiv:2103.01988*, 2021. 3

[20] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020. 3

[21] K. Han, Sylvestre-Alvise Rebuffi, Sébastien Ehrhardt, Andrea Vedaldi, and Andrew Zisserman. Automatically discovering and learning new visual categories with ranking statistics. *ArXiv*, abs/2002.05714, 2020. 3

[22] 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, Oct. 2022. 7

[23] K. Han, Andrea Vedaldi, and Andrew Zisserman. Learning to discover novel visual categories via deep transfer clustering. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8400–8408, 2019. 3

[24] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 2022. 1

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2020. 3

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[27] Yen-Chang Hsu, Zhaoyang Lv, and Zsolt Kira. Learning to cluster in order to transfer across domains and tasks. *ArXiv*, abs/1711.10125, 2017. 3

[28] Yen-Chang Hsu, Zhaoyang Lv, Joel Schlosser, Phillip Odom, and Zsolt Kira. Multi-class classification without multi-class labels. *ArXiv*, abs/1901.00544, 2019. 3

[29] Neha Mukund Kalibhat, Yogesh Balaji, C. Bayan Bruss, and Soheil Feizi. Multi-domain self-supervised learning, 2022. 2, 3

[30] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[31] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

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

[33] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Deeper, broader and artier domain generalization. In *Proceedings of the IEEE international conference on computer vision*, pages 5542–5550, 2017. 3, 6

[34] Jingjing Li, Erpeng Chen, Zhengming Ding, Lei Zhu, Ke Lu, and Heng Tao Shen. Maximum density divergence for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):3918–3930, 2020. 3

[35] Yunsheng Li and Nuno Vasconcelos. Efficient multi-domain learning by covariance normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5424–5433, 2019. 3

[36] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31, 2018. 3

[37] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Math., Stat., and Prob*, page 281, 1965. 2, 6, 7

[38] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 3

[39] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 5

[40] Pau Panareda Busto and Juergen Gall. Open set domain adaptation. In *Proceedings of the IEEE international conference on computer vision*, pages 754–763, 2017. 3

[41] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 3

[42] 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. 6

[43] Nan Pu, Zhun Zhong, and Nicu Sebe. Dynamic conceptional contrastive learning for generalized category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7579–7588, 2023. 2, 3, 6, 7

[44] Sayan Rakshit, Dipesh Tamboli, Pragati Shuddhodhan Meshram, Biplab Banerjee, Gemma Roig, and Subhasis Chaudhuri. Multi-source open-set deep adversarial domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16*, pages 735–750. Springer, 2020. 2

[45] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017. 3

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[47] Alexander Robey, George J Pappas, and Hamed Hassani. Model-based domain generalization. *Advances in Neural Information Processing Systems*, 34:20210–20229, 2021. 3

[48] Kuniaki Saito, Shohei Yamamoto, Yoshitaka Ushiku, and Tatsuya Harada. Open set domain adaptation by backpropagation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 153–168, 2018. 2, 3, 4, 6, 7, 8

[49] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2), 2019. 8

[50] Yang Shu, Zhangjie Cao, Chenyu Wang, Jianmin Wang, and Mingsheng Long. Open domain generalization with domain-augmented meta-learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9624–9633, 2021. 3

[51] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[52] Jesper E Van Engelen and Holger H Hoos. A survey on semi-supervised learning. *Machine learning*, 109(2):373–440, 2020. 1

[53] 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, 3, 4, 5, 6, 7

[54] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5018–5027, 2017. 6

[55] 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*, pages 16590–16600, 2023. 3, 6, 7

[56] Qing Yu, Daiki Ikami, Go Irie, and Kiyoharu Aizawa. Self-labeling framework for novel category discovery over domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3161–3169, 2022. 2

[57] Richard Zhang, Phillip Isola, and Alexei A Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1058–1067, 2017. 3

[58] Shengxiang Zhang, Salman A. Khan, Zhiqiang Shen, Muzammal Naseer, Guangyi Chen, and Fahad Shahbaz Khan. Promptcal: Contrastive affinity learning via auxiliary prompts for generalized novel category discovery. *ArXiv*, abs/2212.05590, 2022. 1, 3

[59] Bingchen Zhao and K. Han. Novel visual category discovery with dual ranking statistics and mutual knowledge distillation. In *Neural Information Processing Systems*, 2021. 3

[60] Bingchen Zhao, Xin Wen, and Kai Han. Learning semi-supervised gaussian mixture models for generalized category discovery, 2023. 3, 6, 7

[61] Zhun Zhong, Enrico Fini, Subhankar Roy, Zhiming Luo, Elisa Ricci, and Nicu Sebe. Neighborhood contrastive learning for novel class discovery. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10867–10875, 2021. 1, 3