Improving Adversarial Robustness for Recommendation Model via Cross-Domain Distributional Adversarial Training

Jingyu Chen
School of Computer Science
Sichuan University
Chengdu, China
chenjingyu@stu.scu.edu.cn

Lilin Zhang
School of Computer Science
Sichuan University
Chengdu, China
zhanglilin@stu.scu.edu.cn

Ning Yang*
School of Computer Science
Sichuan University
Chengdu, China
yangning@scu.edu.cn

ABSTRACT

Recommendation models based on deep learning are fragile when facing adversarial examples (AE). Adversarial training (AT) is the existing mainstream method to promote the adversarial robustness of recommendation models. However, these AT methods often have two drawbacks. First, they may be ineffective due to the ubiquitous sparsity of interaction data. Second, point-wise perturbation used by these AT methods leads to suboptimal adversarial robustness, because not all examples are equally susceptible to such perturbations. To overcome these issues, we propose a novel method called *Cross-domain Distributional Adversarial Training* (CDAT) which utilizes a richer auxiliary domain to improve the adversarial robustness of a sparse target domain. CDAT comprises a *Domain adversarial network* (Dan) and a *Cross-domain adversarial example generative network* (Cdan). Dan learns a domain-invariant preference distribution which is obtained by aligning user embeddings from two domains and paves the way to leverage the knowledge from another domain for the target domain. Then, by adversarially perturbing the domain-invariant preference distribution under the guidance of a discriminator, Cdan captures an aggressive and imperceptible AE distribution. In this way, CDAT can transfer distributional adversarial robustness from the auxiliary domain to the target domain. The extensive experiments conducted on real datasets demonstrate the remarkable superiority of the proposed CDAT in improving the adversarial robustness of the sparse domain. The codes and datasets are available on https://github.com/HymanLoveGIN/CDAT.

CCS CONCEPTS

• **Information systems** → **Recommender systems**;

KEYWORDS

Cross-Domain Recommendation, Adversarial Robustness, Adversarial Training

*Ning Yang is the corresponding author.

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688116

ACM Reference Format:
Jingyu Chen, Lilin Zhang, and Ning Yang. 2024. Improving Adversarial Robustness for Recommendation Model via Cross-Domain Distributional Adversarial Training. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3640457.3688116

1 INTRODUCTION

Deep learning-based recommendation models are widely deployed on online service platforms to deliver tailored recommendations to users. But, these models are vulnerable to adversarial attacks when facing with adversarial examples (AE) [9, 27]. AE is a form of input data that has been meticulously designed to introduce minor yet purposeful perturbations to the original input samples, which can lead to diminished recommendation performance. Therefore, the improvement of adversarial robustness in recommendation models, namely, the ability to defend against AE, has become a focal point of interest for researchers. Adversarial training (AT) [1, 5, 10, 28, 33–35] is the predominant method for enhancing the adversarial robustness of recommendation models, which dynamically generates AEs and integrates them into the training process, equipping recommendation models with adversarial robustness. However, it still faces the following challenges.

Substantial data are needed for AT of a model, otherwise, the model will be suboptimal in adversarial robustness [23]. This is especially true in recommendation models, where the data sparsity issue is widespread. An effective approach to mitigate this issue is cross-domain recommendation (CDR) [3, 11, 14, 38, 40], which enhances the recommendation performance of the sparse target domain by utilizing the knowledge from another auxiliary domain. However, traditional CDRs focus on transferring the knowledge that can improve the recommendation accuracy, without considering the knowledge benefitting adversarial robustness. On the other hand, to bolster the adversarial robustness of recommendation models, the availability of more powerful AEs is crucial, as they present the models with more formidable attack scenarios. However, prevalent AT methods [10, 34, 36] generally adopt the point-wise perturbation strategy to generate AEs, which ignores the varying susceptibility of the samples to such adversarial attacks, thus yielding suboptimal adversarial robustness [2, 25].

To address the above challenges, we propose a novel AT method called ***C**ross-domain **D**istributional **A**dversarial **T**raining* (CDAT). CDAT encompasses a ***D**omain **a**dversarial **n**etwork* (Dan) and a ***C**ross-**d**omain **a**dversarial example generative **n**etwork* (Cdan). The Dan focuses on capturing a domain-invariant user preference distribution. Specifically, the Dan includes a domain-invariant preference

encoder and a domain discriminator. This encoder takes all user data from both domains as input and aligns user preference embeddings from the two different domains within the same embedding space with the assistance of the domain discriminator. This alignment results in a domain-invariant preference distribution, thereby paving the way for the target domain to leverage the knowledge from another domain. Cdan consists of a cross-domain adversarial example generator and a cross-domain adversarial example discriminator. The target of the generator is to learn a distribution of ***C**ross-**D**omain **A**dversarial **E**xample* (CDAE) which is imperceptible and possesses strong attack capability. To be specific, the generator takes the domain-invariant preference distribution obtained by Dan as input and then adversarially perturbs it to get the CDAE distribution. To ensure the imperceptibility of CDAEs, the discriminator enforces a constraint on the distance between the CDAE distribution and the domain-invariant preference distribution and the generator minimizes the distance. Concurrently, to guarantee the aggressiveness of CDAEs, CDAEs generated by the generator are fed into the recommenders of both auxiliary and target domains, and the generator is optimized to maximize the recommendation losses. Finally, we can generate jointly perturbed and imperceptible CDAEs from the CDAE distribution for CDAT. And in this manner, the adversarial robustness from the auxiliary domain can be transferred to the sparse target domain.

The main contributions of this paper can be summarized as follows:

- We propose a Cross-domain Distributional Adversarial Training (CDAT) method, which enhances the adversarial robustness of a recommendation model in the sparse target domain by transferring adversarial robustness from the auxiliary domain.
- We propose a Cross-domain Adversarial Example Generative Network (Cdan). This network takes a domain-invariant preference distribution obtained by Dan as input and captures a high-quality CDAE distribution to support CDAT, under the constraint of the cross-domain adversarial example discriminator and the guidance of maximizing the recommendation losses.
- We conduct extensive experiments in three scenarios constructed by two real datasets to verify the effectiveness of CDAT.

2 PRELIMINARIES

2.1 Base Model

We design a basic recommendation model called base model to combine with CDAT to show its performance. The base model is made up of a preference encoder E and a recommender Rec taking lower user embeddings $\left\{\boldsymbol{e}_u \in \mathbb{R}^d\right\}$ and lower item embeddings $\left\{\boldsymbol{e}_v \in \mathbb{R}^d\right\}$ as input, where d is embedding dimensionality.

2.1.1 User and Item Embedding. Let $\mathcal{U}$ and $\mathcal{V}$ (of size $m = |\mathcal{U}|$ and $n = |\mathcal{V}|$) be the sets of users and items, respectively. We use the user one-hot encoding $\boldsymbol{x}_u \in \{0,1\}^m$ and the item one-hot encoding $\boldsymbol{x}_v \in \{0,1\}^n$ to represent a user $u \in \mathcal{U}$ and an item $v \in \mathcal{V}$, respectively. A user u's embedding $\boldsymbol{e}_u$ will be obtained with a lookup over a learnable embedding matrix $\boldsymbol{W}_u \in \mathbb{R}^{d\times m}$, i.e., $\boldsymbol{e}_u = \boldsymbol{W}_u \boldsymbol{x}_u$. Likewise, we can also obtain an item v's embedding $\boldsymbol{e}_v$ by a lookup over a learnable embedding matrix $\boldsymbol{W}_v \in \mathbb{R}^{d\times n}$, i.e., $\boldsymbol{e}_v = \boldsymbol{W}_v \boldsymbol{x}_v$. Furthermore, the preference encoder E parameterized by θ_E will receive a user embedding $\boldsymbol{e}_u$ as input to gain a user's preference embedding $\boldsymbol{z}_u$, which is implemented as a Multi-Layer Perceptron (MLP) with activation function ReLU.

2.1.2 Interaction Prediction Network. To predict the probability that a user u will interact with an item v, we employ a recommender Rec with an MLP and a sigmoid function as output:

$$r_{uv} = Rec\left(\boldsymbol{z}_u \oplus \boldsymbol{e}_v; \theta_{Rec}\right), \tag{1}$$

where θ_{Rec} represents the learnable parameters of Rec and $\oplus$ represents concat operation.

To optimize the model, we construct a training dataset $\mathcal{D} = \{(u, v_+, v_-)\}$, where v_+ and v_- denote a positive sample and a negative sample of a user u, respectively. By applying the popular pair-wise ranking loss of BPR [22, 38], we define the following objective function:

$$\mathcal{L}_{Rec}\left(\theta_E, \theta_{Rec}\right) = -\frac{1}{|\mathcal{D}|} \sum_{(u,v_+,v_-)\in\mathcal{D}} \log\left(r_{uv_+} - r_{uv_-}\right). \tag{2}$$

2.2 Problem Definition

The target problem of this paper is that we want a recommendation model in the sparse target domain t to learn a set of adversarially robust parameters $\theta^{(t)}_{Rec}$ which denotes the parameters of the recommender in the target domain t to defend the adversarial attack with the help of the auxiliary domain s.

2.3 Threat Model

Attack Capability. We consider the threat model as an evasion attack model, which attacks the target recommendation model during inference by crafting AEs. Although in the realm of computer vision, these perturbations are typically added to raw data, in the context of recommendation, they are instead applied to the model parameters that underlie the recommendation strategy. Specifically, we focus on the adversarial perturbations introduced to user embeddings.

Attack Goal. This threat model belongs to an untargeted attack, which seeks to degrade the overall recommendation performance rather than targeting one user or item.

Attack Knowledge. The threat model is regarded as a white-box attack, permitting the method full knowledge of the recommendation model, including its parameters and structures. Although white-box attacks represent an idealized state of attack, they provide the most challenging adversaries for AT, which is beneficial for enhancing the adversarial robustness of the recommendation model.

3 METHODOLOGY

3.1 Overview

Figure 1 shows the architecture of CDAT, where the solid lines represent the paths activated when training, while the gold solid lines represent the paths activated when applying. CDAT is composed of the Domain adversarial network (Dan) and Cross-domain adversarial example generative network (Cdan).

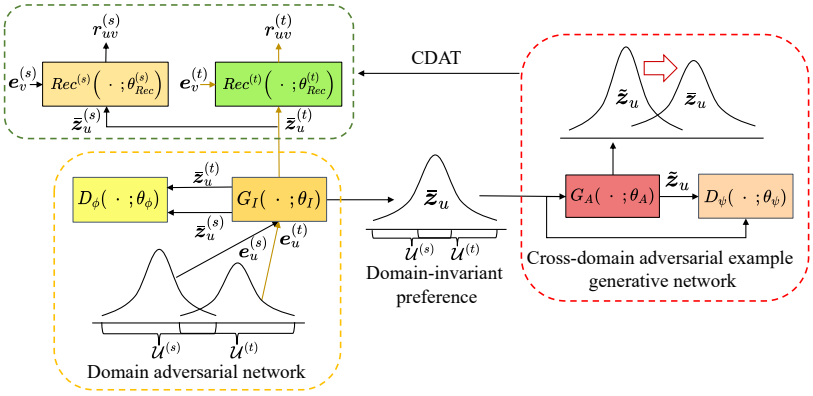


Figure 1: The overview of CDAT.

In Dan, CDAT first transforms the users' one-hot vectors $\left\{\boldsymbol{x}_u^{(\tau)}\right\}$ and the items' one-hot vectors $\left\{\boldsymbol{x}_v^{(\tau)}\right\}$ to their corresponding lower user embeddings $\left\{\boldsymbol{e}_u^{(\tau)}\right\}$ and lower item embeddings $\left\{\boldsymbol{e}_v^{(\tau)}\right\}$ by lookup method. And, $\tau = \{s, t\}$ where s and t denote an auxiliary domain s and a target domain t, respectively. To align the users' preference embeddings from two different domains, CDAT lets domain-invariant preference encoder G_I to capture the distribution $p_{\theta_I}(\bar{\boldsymbol{z}}_u)$ containing users' domain-invariant preferences by means of domain discriminator, where θ_I is the parameters of G_I. In Cdan, with the domain-invariant preference distribution as input, CDAT utilizes cross-domain adversarial example generator G_A parameterized by θ_A to learn a harmful and imperceptible CDAE distribution $p_{\theta_A}(\tilde{\boldsymbol{z}}_u)$.

Lastly, in domain t , the recommender $Rec^{(t)}$ predicts the interaction probability $r_{uv}^{(t)}$ by feeding a user's domain-invariant preference $\bar{\boldsymbol{z}}_u^{(t)}$ and an item embedding $\boldsymbol{e}_v^{(t)}$. At the same time, CDAEs generated by G_A are also fed into the recommender. Then, we adversarially update the parameters of the recommender to obtain a recommendation model with adversarial robustness in the sparse target domain.

3.2 Domain Adversarial Network

Dan is composed of a domain-invariant preference encoder G_I and a domain discriminator D_ϕ. Inspired by DANN [7], we leverage domain discriminator D_ϕ to help encoder G_I to learn a users' domain-invariant preference distribution $p_{\theta_I}(\bar{\boldsymbol{z}}_u)$, where $\bar{\boldsymbol{z}}_u \sim p_{\theta_I}(\bar{\boldsymbol{z}}_u)$, in preparation for further applying the knowledge benefitting AT from the auxiliary domain.

3.2.1 Domain-invariant Preference Embedding. Based on a user embedding $\boldsymbol{e}_u^{(\tau)}$, we generate the domain-invariant preference $\bar{\boldsymbol{z}}_u$ by G_I, as follow:

$$\bar{\boldsymbol{z}}_u = G_I\left(\boldsymbol{e}_u^{(\tau)}; \theta_I\right), \ \tau \in \{s, t\}, \tag{3}$$

where $\bar{\boldsymbol{z}}_u \in \mathbb{R}^d$ and θ_I is the learnable parameters of G_I.

3.2.2 Optimization. To make sure that $\bar{\boldsymbol{z}}_u$ is domain-invariant, we introduce a domain discriminator D_ϕ, which takes $\bar{\boldsymbol{z}}_u$ as input. The loss of the domain discriminator is defined as:

$$\begin{aligned}\mathcal{L}_\phi\left(\theta_I, \theta_\phi\right) = -\Bigg(&\mathbb{E}_{\bar{\boldsymbol{z}}_u^{(s)} \sim p_{\theta_I}\left(\bar{\boldsymbol{z}}_u | \boldsymbol{e}_u^{(s)}\right)} \log\left(D_\phi\left(\bar{\boldsymbol{z}}_u^{(s)}\right)\right) \\ &+\mathbb{E}_{\bar{\boldsymbol{z}}_u^{(t)} \sim p_{\theta_I}\left(\bar{\boldsymbol{z}}_u | \boldsymbol{e}_u^{(t)}\right)} \log\left(1 - D_\phi\left(\bar{\boldsymbol{z}}_u^{(t)}\right)\right)\Bigg),\end{aligned} \tag{4}$$

where θ_ϕ is the learnable parameters of the domain discriminator, $\bar{\boldsymbol{z}}_u^{(s)}$ and $\bar{\boldsymbol{z}}_u^{(t)}$ are the domain-invariant preference based on s and t domains' user embeddings, respectively. And if $\tau = s$,i.e. the domain-invariant preference is from the domain s, we let $D_\phi\left(\bar{\boldsymbol{z}}_u^{(s)}\right) = 1$ as groundtruth, otherwise $D_\phi\left(\bar{\boldsymbol{z}}_u^{(t)}\right) = 0$. The role of the domain discriminator D_ϕ is to help G_I learn an indistinguishable preference distribution. Therefore, the optimization objective of the G_I and D_ϕ is defined as the following minmax game:

$$\min_{\theta_I} \max_{\theta_\phi} -\mathcal{L}_\phi. \tag{5}$$

The optimization objective of the domain discriminator D_ϕ is to strengthen its discriminating ability by maximizing the negative discriminating loss ($-\mathcal{L}_\phi$). The adversarial optimization of Equation (5) results in a good G_I by minimizing the loss to update its parameters. Hence, G_I is able to generate $\{\bar{\boldsymbol{z}}_u\}$ that are indistinguishable enough to fool a powerful domain discriminator.

3.2.3 Constraint. Shared users are both in s and t domains. We can get a user's two domain-invariant preferences from different domains. Then, it is essential to keep the two domain-invariant preferences of a user close in the same space. To achieve this, we introduce an auxiliary loss $\mathcal{L}_{su}$ to constrain them. The $\mathcal{L}_{su}$ is defined as:

$$\mathcal{L}_{su}\left(\theta_I\right) = \left\|\bar{\boldsymbol{z}}_u^{(s)} - \bar{\boldsymbol{z}}_u^{(t)}\right\|, u \in \left\{\mathcal{U}^{(s)} \cap \mathcal{U}^{(t)}\right\}. \tag{6}$$

We will update G_I by minimizing $\mathcal{L}_{su}$ to resist the distance between $\bar{\boldsymbol{z}}_u^{(s)}$ and $\bar{\boldsymbol{z}}_u^{(t)}$.

In the end, we can get the following loss of Dan by combining $\mathcal{L}_\phi$ and $\mathcal{L}_{su}$:

$$\mathcal{L}_{Dan}\left(\theta_I, \theta_\phi\right) = -\mathcal{L}_\phi + \mathcal{L}_{su}. \tag{7}$$

To ensure that the G_I can capture the recommendation information, we combine $\mathcal{L}_{Dan}$ and the loss of recommender with domain-invariant preferences as input to guide the optimization process of Dan jointly. As a result, the optimization objective of Dan can be concluded as follow:

$$\min_{\theta_I} \max_{\theta_\phi} \left(\mathcal{L}_{Dan} + \mathcal{L}_{Rec}^{(s)} + \mathcal{L}_{Rec}^{(t)} \right), \tag{8}$$

where $\mathcal{L}_{Rec}^{(s)}$ is the recommendation loss of the recommender in the auxiliary domain and $\mathcal{L}_{Rec}^{(t)}$ is the recommendation loss of the recommender in the target domain.

3.3 Cross-domain Adversarial Example Generative Network

Inspired by generative adversarial network [8, 38], we propose a novel cross-domain adversarial example generative network (Cdan) containing cross-domain adversarial example generator G_A and cross-domain adversarial example discriminator D_ψ to jointly perturb all samples. G_A will learn a satisfactory distribution of CDAE constrained by D_ψ and maximized recommendation losses.

3.3.1 CDAE. The goal of G_A is to perturb all input samples jointly to generate an imperceptible CDAEs' distribution $p_{\theta_A}(\tilde{\boldsymbol{z}}_u)$ which has enough harmfulness, where $\tilde{\boldsymbol{z}}_u \sim p_{\theta_A}(\tilde{\boldsymbol{z}}_u)$. To realize this, we take domain-invariant preference distribution as input of G_A and generate CDAE $\tilde{\boldsymbol{z}}_u$:

$$\tilde{\boldsymbol{z}}_u = G_A(\bar{\boldsymbol{z}}_u; \theta_A),\ u \in \left\{ \mathcal{U}^{(s)} \cup \mathcal{U}^{(t)} \right\}, \tag{9}$$

where $\bar{\boldsymbol{z}}_u$ implies a user's domain-invariant preference who is from s or t domain and θ_A is the learnable parameters of G_A.

3.3.2 Optimization. In order to guarantee that CDAEs' distribution is imperceptible, we introduce a cross-domain adversarial example discriminator D_ψ parameterized by θ_ψ, with $\bar{\boldsymbol{z}}_u$ and $\tilde{\boldsymbol{z}}_u$ as input. Consequently, the loss of Cdan is defined as:

$$\begin{aligned} \mathcal{L}_{Cdan}\left(\theta_A, \theta_\psi\right) = -\Big(& \mathbb{E}_{\bar{\boldsymbol{z}}_u \sim p_{\theta_I}(\bar{\boldsymbol{z}}_u)} \log\left(D_\psi\left(\bar{\boldsymbol{z}}_u\right)\right) \\ & + \mathbb{E}_{\tilde{\boldsymbol{z}}_u \sim p_{\theta_A}(\tilde{\boldsymbol{z}}_u)} \log\left(1 - D_\psi\left(\tilde{\boldsymbol{z}}_u\right)\right)\Big). \end{aligned} \tag{10}$$

If $\bar{\boldsymbol{z}}_u$ acts as input of D_ψ, we set $D_\psi(\bar{\boldsymbol{z}}_u) = 1$, otherwise $D_\psi(\tilde{\boldsymbol{z}}_u) = 0$.

To make certain that the CDAE distribution possesses adequate aggressivity while avoiding obtaining a trivial Cdan, it is required to maximize the recommendation losses of the auxiliary and target domains' recommender taking CDAEs as input to direct the optimization of Cdan. As a result, we define the following minmax optimization objective for Cdan:

$$\min_{\theta_\psi} \max_{\theta_A} \left(\mathcal{L}_{Cdan} + \lambda \left(\tilde{\mathcal{L}}_{Rec}^{(s)} + \tilde{\mathcal{L}}_{Rec}^{(t)} \right) \right), \tag{11}$$

where λ is a factor to control the contribution of CDAT, $\tilde{\mathcal{L}}_{Rec}^{(s)}$ is the adversarial recommendation loss of the auxiliary domain and $\tilde{\mathcal{L}}_{Rec}^{(t)}$ is the adversarial recommendation loss of the target domain.

In this way, we train the discriminator D_ψ to minimize the cross entropy $\mathcal{L}_{Cdan}$ to yield a strong discriminator to ensure CDAEs' imperceptible. That CDAEs own imperceptibility indicates that domain-invariant preference distribution and CDAE distribution are close. Apart from this, we also adversarially train G_A with the aim of maximizing the loss $\mathcal{L}_{Cdan} + \lambda \left(\tilde{\mathcal{L}}_{Rec}^{(s)} + \tilde{\mathcal{L}}_{Rec}^{(t)} \right)$ to gain aggressive CDAEs for enhancing adversarial robustness of recommendation model in the sparse target domain.

Algorithm 1 Training Algorithm of CDAT.

Input: Datasets $\mathcal{D} = \mathcal{D}^{(s)} \bigcup \mathcal{D}^{(t)}$, λ, η;
Output: Well-trained G_I, $Rec^{(t)}$;

Random initialize G_I, D_ϕ, G_A, D_ψ, $Rec^{(s)}$ and $Rec^{(t)}$;
for T epochs **do**
  Randomly draw samples $\{(u, v_+, v_-)\}$ from $\mathcal{D}$;
  for T_{Dan} epochs **do**
    Compute gradient $\nabla_{\theta_\phi}(\mathcal{L}_{Dan})$ and $\nabla_{\theta_I}\left(\mathcal{L}_{Dan} + \mathcal{L}_{Rec}^{(s)} + \mathcal{L}_{Rec}^{(t)}\right)$;
    $\theta_\phi^{(T+1)} = \theta_\phi^{(T)} + \eta \nabla_{\theta_\phi}(\mathcal{L}_{Dan})$;
    $\theta_I^{(T+1)} = \theta_I^{(T)} - \eta \nabla_{\theta_I}\left(\mathcal{L}_{Dan} + \mathcal{L}_{Rec}^{(s)} + \mathcal{L}_{Rec}^{(t)}\right)$;
  end for
  Get $\tilde{\boldsymbol{z}}_u \sim p_{\theta_I}(\bar{\boldsymbol{z}}_u)$;
  for T_{Cdan} epochs **do**
    Compute gradient $\nabla_{\theta_\psi}(\mathcal{L}_{Cdan})$ and $\nabla_{\theta_A}\left(\mathcal{L}_{Cdan} + \lambda\left(\tilde{\mathcal{L}}_{Rec}^{(s)} + \tilde{\mathcal{L}}_{Rec}^{(t)}\right)\right)$;
    $\theta_\psi^{(T+1)} = \theta_\psi^{(T)} - \eta \nabla_{\theta_\psi}(\mathcal{L}_{Cdan})$;
    $\theta_A^{(T+1)} = \theta_A^{(T)} + \eta \nabla_{\theta_A}\left(\mathcal{L}_{Cdan} + \lambda\left(\tilde{\mathcal{L}}_{Rec}^{(s)} + \tilde{\mathcal{L}}_{Rec}^{(t)}\right)\right)$;
  end for
  Get $\tilde{\boldsymbol{z}}_u \sim p_{\theta_A}(\tilde{\boldsymbol{z}}_u)$;
  Compute gradient $\nabla_{\theta_{Rec}^{(s)}}\left(\mathcal{L}_{Rec}^{(s)} + \lambda \tilde{\mathcal{L}}_{Rec}^{(s)}\right)$ and $\nabla_{\theta_{Rec}^{(t)}}\left(\mathcal{L}_{Rec}^{(t)} + \lambda \tilde{\mathcal{L}}_{Rec}^{(t)}\right)$;
  $\left(\theta_{Rec}^{(s)}\right)^{(T+1)} = \left(\theta_{Rec}^{(s)}\right)^{(T)} - \eta \nabla_{\theta_{Rec}^{(s)}}\left(\mathcal{L}_{Rec}^{(s)} + \lambda \tilde{\mathcal{L}}_{Rec}^{(s)}\right)$;
  $\left(\theta_{Rec}^{(t)}\right)^{(T+1)} = \left(\theta_{Rec}^{(t)}\right)^{(T)} - \eta \nabla_{\theta_{Rec}^{(t)}}\left(\mathcal{L}_{Rec}^{(t)} + \lambda \tilde{\mathcal{L}}_{Rec}^{(t)}\right)$;
end for

3.4 Cross-domain Distributional Adversarial Training

We leverage these well-pleasing CDAEs captured by Cdan to adversarially attack the recommender $Rec^{(t)}$ in the target domain and update the parameters of $Rec^{(t)}$ to obtain a more robust recommendation model. In this process, the adversarial robustness will be transferred from the rich auxiliary domain to the sparse target domain by CDAEs.

To be specific, a CDAE $\tilde{\boldsymbol{z}}_u$ is sampled from $p_{\theta_A}(\tilde{\boldsymbol{z}}_u)$ which fits by G_A. And then, we predict the attacked interaction probability $\tilde{r}_{\tilde{u}v}^{(t)}$ that an attacked user $\tilde{u}$ will interact with an item v:

$$\tilde{r}_{\tilde{u}v}^{(t)} = Rec^{(t)}\left(\tilde{\boldsymbol{z}}_u^{(t)} \oplus \boldsymbol{e}_v^{(t)}; \theta_{Rec}^{(t)}\right), \tag{12}$$

where $\tilde{\boldsymbol{z}}_u^{(t)} \sim p_{\theta_A}\left(\tilde{\boldsymbol{z}}_u \mid \bar{\boldsymbol{z}}_u^{(t)}\right)$ and $\boldsymbol{e}_v^{(t)}$ is an item embedding in domain t.

To accomplish CDAT, we build an adversarial training set $\tilde{\mathcal{D}}^{(t)} = \{(\tilde{u}, v_+, v_-)\}$, where v_+ and v_- represent a positive sample and a negative sample of an attacked user $\tilde{u}$, respectively. We define the following adversarial recommendation loss of the recommender

for attacked domain t like Equation (2):

$$\tilde{\mathcal{L}}_{Rec}^{(t)}\left(\theta_A, \theta_{Rec}^{(t)}\right) = -\frac{1}{\left|\tilde{\mathcal{D}}^{(t)}\right|}\sum_{(\tilde{u},v_+,v_-)\in\tilde{\mathcal{D}}^{(t)}} \log\left(\tilde{r}_{\tilde{u}v_+}^{(t)} - \tilde{r}_{\tilde{u}v_-}^{(t)}\right). \quad (13)$$

In order to strengthen the adversarial robustness of the recommendation model in the sparse target domain t by CDAT, the optimization objective is defined as the following minmax game:

$$\min_{\theta_{Rec}^{(t)}} \max_{\theta_A} \tilde{\mathcal{L}}_{Rec}^{(t)}. \quad (14)$$

θ_A is updated by maximizing the loss $\tilde{\mathcal{L}}_{Rec}^{(t)}$ to destroy the performance of $Rec^{(t)}$. Meanwhile, $\theta_{Rec}^{(t)}$ is also updated by minimizing the loss $\tilde{\mathcal{L}}_{Rec}^{(t)}$ to make $Rec^{(t)}$ be able to defend the adversarial attack.

3.5 Overall Optimization Objective

Finally, by combining $\mathcal{L}_{Dan}$, $\mathcal{L}_{Cdan}$, $\mathcal{L}_{Rec}^{(s)}$, $\mathcal{L}_{Rec}^{(t)}$, $\tilde{\mathcal{L}}_{Rec}^{(s)}$ and $\tilde{\mathcal{L}}_{Rec}^{(t)}$, we can get the following overall optimization objective:

$$\min_{\theta_I,\theta_\psi,\theta_{Rec}^{(s)},\theta_{Rec}^{(t)}} \max_{\theta_\phi,\theta_A} \Big(\mathcal{L}_{Dan} + \mathcal{L}_{Cdan} + \mathcal{L}_{Rec}^{(s)} + \mathcal{L}_{Rec}^{(t)} + \lambda\left(\tilde{\mathcal{L}}_{Rec}^{(s)} + \tilde{\mathcal{L}}_{Rec}^{(t)}\right)\Big). \quad (15)$$

We summarize the training procedure of CDAT in Algorithm 1.

Table 1: Statistics of datasets.

Datasets	Douban				Amazon	
Scenarios	Movie-Book		Movie-Music		TV-CD	
Domains	Movie	Book	Movie	Music	TV	CD
#Users	1,873	2,063	2,049	1,624	17,894	15,733
#Shared users	1,231		967		1,859	
#Items	9,468	6,703	9,496	5,508	9,881	12,482
#Interactions	460,028	69,211	487,122	51,985	520,166	262,709
Density	2.59%	0.50%	2.51%	0.58%	0.29%	0.13%

4 EXPERIMENTS

The experiments aim to answer the following research questions:

- **RQ1** How does CDAT perform compared to the state-of-the-art baselines in terms of recommendation accuracy and adversarial robustness?
- **RQ2** How do the different parts of CDAT affect the performance?
- **RQ3** How to illustrate the superiority of CDAT with visualizable case studies?

4.1 Experimental Settings

4.1.1 Datasets. We conduct the experiments on two public datasets: Douban [39] and Amazon [21]. Specifically, we build three cross-domain scenarios: Movie-Book and Movie-Music in Douban, and TV-CD in Amazon, where Book, Music and CD are the target domains. Table 1 shows the statistics of the three cross-domain scenarios (after all pre-processing steps), where the density is defined as the ratio of the observed interactions over all possible interactions. For the interaction data of the target domain in different scenarios, the data is randomly divided into training set, validation set, and testing set. For each user, we randomly select an interacted history to the validation set, another one to the testing set and the remaining ones to the training set. We repeat such procedure three times and report the average results with standard deviation.

4.1.2 Baseline Methods. We compare our CDAT with the following baseline methods, comprising one basic recommendation model, three AT methods in recommendation (APR, DAT, ACDN), and two CDR methods (MLP++, ACDR) which are briefly described as follows:

- **Base Model** is a basic recommendation model including an encoder and a recommender, which are both implemented as an MLP with activation function ReLU.
- **APR** [10] is an AT method, which generates AEs based on the user embeddings to maximize the BPR loss and then forces the recommendation model to minimize the BPR loss on AEs.
- **DAT** [33] further considers to add perturbations to both user and item embeddings with proper restriction.
- **ACDN** [34] leverages the users shared by the target and auxiliary domains to construct AEs for AT.
- **MLP++** [11] combines two MLPs by sharing user embedding matrix to realize cross-domain recommendation.
- **ACDR** [14] incorporates global user preferences and domain-specific user preferences via adversarial learning.

For fairness, APR, DAT, ACDN and proposed CDAT use the same Base Model to compare their performances.

4.1.3 Evaluation Protocols. We adopt the widely used metrics Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) to evaluate the recommendation accuracy of CDAT and the baseline methods. HR measures whether the testing item is in the top-K list. While NDCG is a position-sensitive metric, which assigns higher weights to hits at higher positions. In a word, for both metrics, the higher values, the better performance can indicate. We also employ the popular *leave-one-out* evaluation protocol [4, 10]. Specifically, we rank a positive item of a testing user among her 99 negative items which are randomly sampled.

In order to fully evaluate the defense capability of these methods against different attacks, we employ three attacking methods FGSM [9], PGD [20] and advGAN [31] to attack the recommendation model trained by CDAT and baseline methods. Then, we evaluate the adversarial robustness also by HR and NDCG. For FGSM and PGD, we set up two attack intensities, e.g. $\epsilon \in \{0.1, 1\}$, to test the robustness against different attack strengths, which are denoted as FGSM-0.1, FGSM-1, PGD-0.1 and PGD-1. The detailed settings of all attacks are presented in Table 2.

4.1.4 Hyper-Parameter Setting. All the hyper-parameters of baselines and CDAT are tuned on validation sets. The specific settings of CDAT are presented in Table 3. All modules in CDAT are implemented as an MLP with activation function ReLU. We choose Adam as the optimizer. To ensure the stability of the adversarial process, we employ the hinge gan loss [17, 41].

Table 2: Hyper-parameters setting of different attacking methods in different scenarios

Attacking Method	Scenarios		
	Movie-Book	Movie-Music	TV-CD
advGAN	α=0.3,β=0.1,c=10	α=0.3,β=0.1,c=10	α=0.5,β=0.1,c=10
FGSM-0.1	ϵ=0.1, step size=0.1		
FGSM-1	ϵ=1, step size=1		
PGD-0.1	ϵ=0.1, step size=0.01, iterations=20		
PGD-1	ϵ=1, step size=0.1, iterations=20		

Table 3: Hyper-parameters setting of CDAT in different scenarios

Hyper-parameter	Scenarios		
	Movie-Book	Movie-Music	TV-CD
learning rate of D_ϕ and G_I	0.0003	0.0003	0.0001
learning rate of D_ψ and G_A	0.0001	0.0001	0.0001
learning rate of $Rec^{(\tau)}$	0.0001	0.0001	0.0001
λ	0.1	0.3	0.1
size of embedding	128		
size of domain-invariant preference	32		
size of CDAE	32		

4.2 Performance Comparison (RQ1)

Tables 4 and 5 present the target domain's performance in different scenarios from which we can make the following observations and analyses.

First, in all scenarios, CDAT almost consistently outperforms the baselines in terms of all the metrics under adversarial robustness evaluation. These results verify the superiority of CDAT which can enhance the adversarial robustness of the sparse target domain by transferring distributional robustness from a denser auxiliary domain.

And, these results also imply that our CDAT using joint perturbation is superior to the baselines using point-wise perturbation.

Then, notably, in certain cases, the performance of APR and DAT, designed for single-domain recommendations, slightly outperforms our CDAT. We argue that it may be attributed to negative transfer of recommendation information in CDAT in certain sparse cases.

Finally, in Figure 2, we can observe that the performance of general CDR in the target domain declines significantly with the PGD's perturbation constrain ϵ increasing. Consequently, we confirm that the general CDRs just transfer recommendation information to have a reliable recommendation performance and cannot transfer the knowledge benefiting adversarial robustness, which leads to unpromising adversarial robustness of the general CDR model.

4.3 Ablation Study (RQ2)

4.3.1 Performance of Different Parts. Now we investigate the effectiveness of the Cdan, unshared users, and D_ψ. For this purpose, we compare CDAT with its three variants as follows:

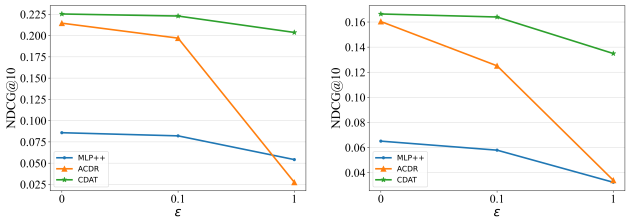
- **CDAT-w/o-Cdan** is the variant without Cdan, where $\mathcal{L}_{Cdan}$, $\tilde{\mathcal{L}}^{(s)}_{Rec}$, and $\tilde{\mathcal{L}}^{(t)}_{Rec}$ are removed from the Equation (15).
- **CDAT-Shared** is the variant that only utilizes the shared users' data as input.
- **CDAT-w/o-D_ψ** is the variant without the discriminator D_ψ, where $\mathcal{L}_{Cdan}$ is removed from the Equation (15).

Table 4: Performance comparison in different scenarios. Clean represents the performance of the recommendation model trained by the proposed CDAT and baselines without adversarial attacks on the test set. The metric is HR@10 in this table.

Scenarios	Attack	Base Model	APR	DAT	ACDN	CDAT
Movie-Book	Clean	0.3652±0.0087	0.3822±0.0127	0.3844±0.0120	0.3404±0.0256	**0.3940±0.0066**
	FGSM-0.1	0.3081±0.0153	0.3714±0.0125	0.3774±0.0108	0.3390±0.0244	**0.3920±0.0063**
	FGSM-1	0.1640±0.0141	0.3313±0.0187	0.3510±0.0086	0.3364±0.0247	**0.3850±0.0042**
	PGD-0.1	0.3013±0.0156	0.3614±0.0117	0.3708±0.0102	0.3394±0.0233	**0.3904±0.0061**
	PGD-1	0.1067±0.0143	0.2529±0.0313	0.2739±0.0346	0.3263±0.0259	**0.3698±0.0039**
	advGAN	0.1403±0.0244	0.1015±0.0110	0.1472±0.0174	0.3360±0.0237	**0.3886±0.0079**
Movie-Music	Clean	0.2789±0.0167	0.3018±0.0027	0.3135±0.0184	0.2802±0.0117	**0.3180±0.0146**
	FGSM-0.1	0.2252±0.0120	0.2956±0.0013	0.3126±0.0180	0.2743±0.0042	**0.3169±0.0159**
	FGSM-1	0.0877±0.0056	0.2805±0.0025	0.2928±0.0259	0.2208±0.0612	**0.3004±0.0156**
	PGD-0.1	0.2184±0.0092	0.2928±0.0026	0.3120±0.0185	0.2710±0.0073	**0.3155±0.0141**
	PGD-1	0.0488±0.0018	0.2244±0.0102	0.2502±0.0494	0.1906±0.0884	**0.2724±0.0170**
	advGAN	0.0570±0.0133	0.1115±0.0221	0.1563±0.0055	0.2271±0.0741	**0.2976±0.0148**
TV-CD	Clean	0.3079±0.0052	0.3025±0.0222	0.2959±0.0098	0.2326±0.0153	**0.3176±0.0047**
	FGSM-0.1	0.2826±0.0084	0.2867±0.0496	0.2814±0.0095	0.2067±0.0310	**0.3162±0.0047**
	FGSM-1	0.1946±0.0112	**0.3265±0.0246**	0.2538±0.0088	0.1409±0.0420	0.3100±0.0074
	PGD-0.1	0.2819±0.0081	0.2860±0.0504	0.2803±0.0094	0.2032±0.0350	**0.3159±0.0049**
	PGD-1	0.1765±0.0094	0.2072±0.1381	0.2382±0.0089	0.1017±0.0396	**0.3027±0.0073**
	advGAN	0.2100±0.0047	0.2515±0.1077	0.2827±0.0076	0.1883±0.0676	**0.3167±0.0049**

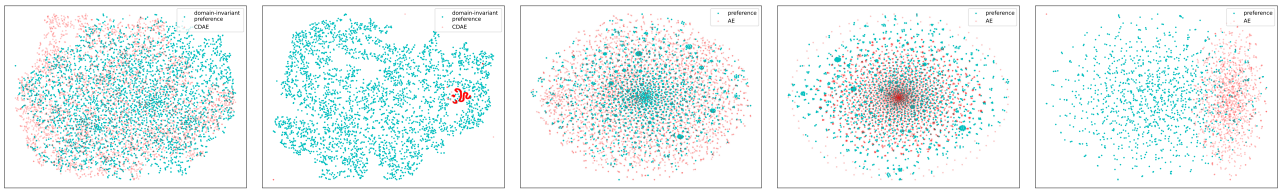
Table 5: Performance comparison in different scenarios. Clean represents the performance of the recommendation model trained by the proposed CDAT and baselines without adversarial attacks on the test set. The metric is NDCG@10 in this table.

Scenarios	Attack	Base Model	APR	DAT	ACDN	CDAT
Movie-Book	Clean	0.2086±0.0076	0.2225±0.0086	0.2251±0.0066	0.2094±0.0126	**0.2254±0.0061**
	FGSM-0.1	0.1767±0.0112	0.2164±0.0091	0.2208±0.0058	0.2084±0.0116	**0.2237±0.0060**
	FGSM-1	0.0885±0.0059	0.1901±0.0102	0.2068±0.0057	0.2068±0.0091	**0.2153±0.0056**
	PGD-0.1	0.1735±0.0118	0.2122±0.0084	0.2186±0.0056	0.2085±0.0113	**0.2230±0.0053**
	PGD-1	0.0623±0.0050	0.1511±0.0187	0.1664±0.0187	0.2021±0.0089	**0.2037±0.0090**
	advGAN	0.0973±0.0155	0.0686±0.0052	0.1034±0.0127	0.2073±0.0111	**0.2223±0.0058**
Movie-Music	Clean	0.1478±0.0103	0.1576±0.0041	0.1643±0.0067	0.1491±0.0059	**0.1664±0.0038**
	FGSM-0.1	0.1191±0.0085	0.1547±0.0032	0.1644±0.0059	0.1442±0.0117	**0.1646±0.0042**
	FGSM-1	0.0480±0.0058	0.1453±0.0033	**0.1545±0.0099**	0.1142±0.0401	0.1508±0.0023
	PGD-0.1	0.1165±0.0073	0.1534±0.0039	**0.1645±0.0058**	0.1422±0.0142	0.1640±0.0036
	PGD-1	0.0289±0.0021	0.1193±0.0029	0.1343±0.0214	0.0995±0.0534	**0.1349±0.0018**
	advGAN	0.0369±0.0098	0.0719±0.0111	0.0956±0.0062	0.1172±0.0495	**0.1559±0.0046**
TV-CD	Clean	0.1547±0.0040	0.1589±0.0102	**0.1618±0.0050**	0.1259±0.0118	0.1609±0.0018
	FGSM-0.1	0.1396±0.0053	0.1512±0.0233	0.1518±0.0047	0.1112±0.0188	**0.1602±0.0017**
	FGSM-1	0.0955±0.0060	**0.1703±0.0131**	0.1335±0.0050	0.0731±0.0241	0.1561±0.0030
	PGD-0.1	0.1405±0.0050	0.1509±0.0236	0.1516±0.0049	0.1097±0.0202	**0.1599±0.0018**
	PGD-1	0.0886±0.0046	0.1068±0.0671	0.1231±0.0058	0.0527±0.0220	**0.1514±0.0031**
	advGAN	0.1138±0.0024	0.1316±0.0565	0.1544±0.0043	0.1043±0.0327	**0.1607±0.0019**



(a) Movie-Book (b) Movie-Music

Figure 2: Performance comparison with general cross-domain recommendation methods.



(a) constraint with D_ψ (b) constraint without D_ψ (c) APR (d) DAT (e) ACDN

Figure 3: Visualization comparison between AE distribution and preference distribution in Movie-Book.

Table 6: Ablation Studies in Movie-Book. Clean represents the performance of the recommendation model trained by the variants without adversarial attacks on the test set.

Attack	Metric	CDAT-w/o-Cdan	CDAT-w/o-D_ψ	CDAT-Shared	CDAT
Clean	HR@10	0.3906±0.0087	0.3918±0.0044	0.3816±0.0086	**0.3940±0.0066**
	NDCG@10	0.2219±0.0061	0.2216±0.0061	0.2057±0.0082	**0.2254±0.0061**
FGSM-0.1	HR@10	0.3848±0.0119	0.3880±0.0054	0.3799±0.0076	**0.3920±0.0063**
	NDCG@10	0.2137±0.0070	0.2191±0.0056	0.2021±0.0053	**0.2237±0.0060**
FGSM-1	HR@10	0.3426±0.0530	0.3762±0.0103	0.3538±0.0171	**0.3850±0.0042**
	NDCG@10	0.1805±0.0397	0.2074±0.0096	0.1843±0.0093	**0.2153±0.0056**
PGD-0.1	HR@10	0.3828±0.0121	0.3866±0.0062	0.3786±0.0071	**0.3904±0.0061**
	NDCG@10	0.2112±0.0077	0.2182±0.0047	0.2011±0.0049	**0.2230±0.0053**
PGD-1	HR@10	0.2971±0.0623	0.3525±0.0100	0.3152±0.0107	**0.3698±0.0039**
	NDCG@10	0.1534±0.0411	0.1909±0.0064	0.1638±0.0030	**0.2037±0.0090**
advGAN	HR@10	0.3524±0.0141	0.3868±0.0060	0.3484±0.0081	**0.3886±0.0079**
	NDCG@10	0.1911±0.0193	0.2188±0.0065	0.1910±0.0120	**0.2223±0.0058**

The results of the Movie-Book scenario are shown in Table 6. First, it can be observed that the performance of CDAT is remarkably better than CDAT-w/o-Cdan, which shows that Cdan generating high-quality CDAEs plays an important role in enhancing the sparse domain's adversarial robustness. Second, we can see that CDAT also outperforms CDAT-Shared. This verifies that leveraging sufficient unshared users is better than only employing shared users' data. In the end, it is observable that the performance of CDAT-w/o-D_ψ is inferior to CDAT, indicating that CDAEs possessing imperceptibility are good for CDAT.

Figure 3 employs the t-SNE algorithm [29] to visualize the user preference distribution and AE distribution under CDAT, CDAT-w/o-D_ψ and three baselines. By comparing subfigures (a) and (b) in Figure 3, we observe that with the constraint of the discriminator D_ψ, the CDAEs' distribution and domain-invariant preference distribution are closer, indicating the imperceptibility of our CDAEs. Furthermore, comparing subfigures (a) with (c), (d), and (e), it is evident that the AEs produced by CDAT are more balanced in terms of quality compared to other point-wise perturbation-based methods. This result validates the effectiveness of our CDAT which jointly perturbs all samples for generating AEs.

4.3.2 Hyper-parameter and Model Convergence Studies. In this section, we study the influence of hyper-parameter, the balance factor λ between standard generalization and adversarial robustness generalization, of which the results are shown in Figure 4 in Movie-Book. We can see that the performance of CDAT under adversarial robustness evaluation significantly increases with the increasing of λ from 0 to 0.1, which demonstrates that the model's adversarial robustness will get promoted with the increasing contribution of CDAT.

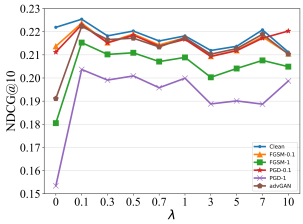


Figure 4: Tuning of hyper-parameter λ in Movie-Book.

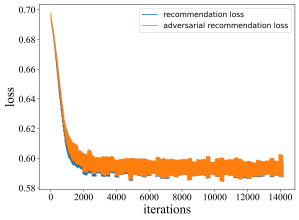
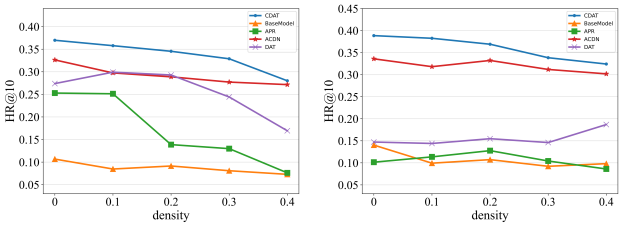


Figure 5: Convergence analysis



(a) PGD-1 (b) advGAN

Figure 6: Density Analysis in Movie-Book.

However, when λ is bigger than 0.1, the performance gradually declines. This is attributed to the excessive contribution of CDAT, which hinders the model's acquisition of adversarial robustness.

Additionally, Figure 5 presents the variation of recommendation loss $\mathcal{L}_{Rec}^{(t)}$ and adversarial recommendation loss $\tilde{\mathcal{L}}_{Rec}^{(t)}$ during the training phase in Movie-Book with λ set to 0.1. It can be observed that both $\mathcal{L}_{Rec}^{(t)}$ and $\tilde{\mathcal{L}}_{Rec}^{(t)}$ decrease gradually with the increasing of iterations, eventually plateauing within a certain range with minor fluctuations. This trend reflects that our CDAT enables the model to achieve a balance between recommendation performance and adversarial robustness as the iterations increases.

4.3.3 Density Analysis. Figure 6 illustrates the variation in performance of our CDAT and baseline methods as the density decreases under different attacks, where the horizontal axes represent the proportion of the data that are randomly removed from the target domain's training set. It can be observed that the performance of almost all methods declines as the density decreases. However, the performance of our CDAT remains superior to all baselines. The

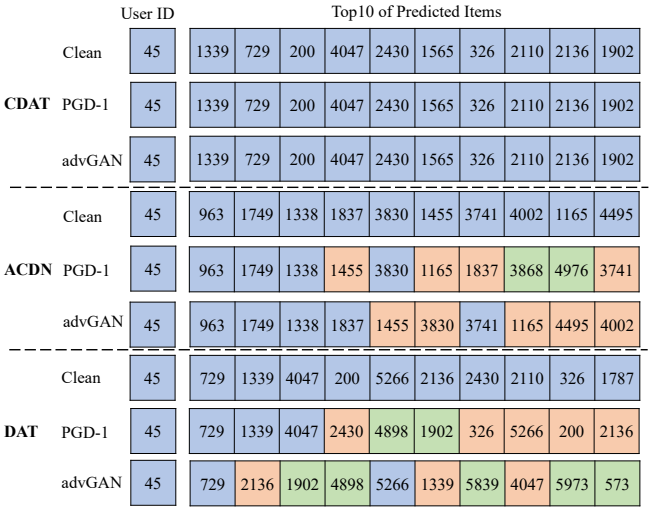


Figure 7: Visualization of adversarial robustness of the recommendation model comparison between our CDAT and two baselines ACDN and DAT in Movie-Book.

result indicates that our CDAT which leverages the dense auxiliary domain to enhance the adversarial robustness of the sparse target domain is effective.

4.4 Case Study (RQ3)

In Figure 7, we visually present the adversarial robustness of recommendation models trained by CDAT, ACDN and DAT in Movie-Book. We select a user whose ID is 45 at random. For this user, we obtain her top-10 predicted item lists from the recommendation models trained by CDAT, ACDN and DAT on the test set, respectively. We evaluate these predicted item lists through three evaluation methods which are clean, PGD-1 and advGAN.

For each AT method, in the PGD-1 and advGAN evaluation, if the cell is filled with blue, it signifies that the item at that position is consistent with the item in the corresponding position of the clean evaluation. If the cell is filled with orange, it indicates that the item at that position in the PGD-1 or advGAN evaluation appears in the clean evaluation but not in the corresponding position. A green fill means that the item at that position is present in the PGD-1 or advGAN evaluation but absent in the clean evaluation.

As we can see, the three evaluations in CDAT are identical, demonstrating that our CDAT method is able to enhance the adversarial robustness of the sparse recommendation model. By contrast, under the ACDN and DAT, the PGD-1 and advGAN evaluations exhibit significant changes compared to the clean evaluation, indicating that the two methods are less effective at improving the model's adversarial robustness.

5 RELATED WORKS

5.1 Adversarial Training Method in Recommendation

Adversarial training [28, 35, 36] is an effective method to promote the adversarial robustness of recommendation models. It typically involves two steps: generating AEs and optimizing model's parameters. APR [10] is the first AT method proposed specifically for recommendation. It aims to strengthen the robustness and generalization capabilities of recommendations with BPR. Building upon APR, the DAT [33] approach further refines the perturbation by restricting its direction. Additionally, ACDN [34] dynamically generates adversarial examples based on shared user embeddings.

In summary, due to the widespread sparsity of data in recommendation, existing AT methods may not effectively improve the adversarial robustness of recommendation models in sparse settings. Moreover, existing AT methods generate AEs by point-wise perturbation, which can lead to uneven AEs' quality due to varying sensitivities to such perturbations, thus hindering the efficacy of AT. In this paper, we address these issues by utilizing the knowledge benefitting AT from the auxiliary domain and joint perturbation to generate AEs.

5.2 Cross-domain Recommendation

In the pursuit of mitigating data sparsity in recommendation models, researchers introduce CDR methods [11, 13, 18, 32, 37]. According to different CDR tasks, these methods can be categorized into single-target domain and dual-target domain recommendations.

The single-target domain recommendations are designed to enhance the recommendation performance in the sparse target domain by leveraging the richer data from the auxiliary domain. Early techniques employ joint factorization of the user-item interaction matrices from both the auxiliary and target domains to capture the common preferences among shared users [12, 16, 24]. In recent years, many methods based on deep neural networks have been proposed for better improving performance in the sparse target domain by transfer learning [6, 15]. Dual-target domain recommendations represent a burgeoning trend in recent years, aiming to simultaneously promote the recommendation performance in two relevant domains [11, 14, 38]. Early methods transfer the knowledge between two domains by jointly modeling a user's preference in two relevant domains like CoNet. However, the users' representations from such methods may be harmful to dual-CDR's performance due to mixed representations. Recently, researchers have also proposed some methods [3, 19, 26, 30] to disentangle preferences.

Although the above CDR methods demonstrate promising performance in generalization ability, they almost ignore transferring adversarial robustness from the auxiliary domain to the target domain, which leads to degraded performance of CDR when facing adversarial attacks. In this paper, we propose the CDAT to tackle this issue.

6 CONCLUSION

In this paper, we propose a novel AT method called Cross-domain Distributional Adversarial Training (CDAT) for enhancing the adversarial robustness of the sparse recommendation domain, which transfers distributional adversarial robustness from the dense auxiliary domain to the sparse target domain. In particular, we propose a *Domain adversarial network* (Dan) that captures a distribution containing domain-invariant preferences, aligning the user embeddings of two domains and paving the way for the use of the knowledge from another domain. We also propose a *Cross-domain adversarial*

example generative network (Cdan) that learns a strongly attacking CDAEs' distribution of good imperceptibility by joint perturbation with the guidance of maximizing recommendation losses and the constraint of the cross-domain adversarial example discriminator to support CDAT. At last, remarkable improvements in adversarial robustness on the sparse target domain demonstrate the superiority of our CDAT.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of Sichuan Province under grant 2024NSFSC0516 and National Natural Science Foundation of China under grant 61972270.

REFERENCES

[1] Vito Walter Anelli, Yashar Deldjoo, Tommaso DiNoia, and Felice Antonio Merra. 2021. Adversarial recommender systems: Attack, defense, and advances. In *Recommender systems handbook*. Springer, 335–379.

[2] Tuan Anh Bui, Trung Le, Quan Tran, He Zhao, and Dinh Phung. 2022. A unified wasserstein distributional robustness framework for adversarial training. *arXiv preprint arXiv:2202.13437* (2022).

[3] Jiangxia Cao, Xixun Lin, Xin Cong, Jing Ya, Tingwen Liu, and Bin Wang. 2022. Disencdr: Learning disentangled representations for cross-domain recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 267–277.

[4] Yizhou Dang, Enneng Yang, Guibing Guo, Linying Jiang, Xingwei Wang, Xiaoxiao Xu, Qinghui Sun, and Hong Liu. 2023. Uniform Sequence Better: Time Interval Aware Data Augmentation for Sequential Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 4225–4232.

[5] Yashar Deldjoo, Tommaso Di Noia, and Felice Antonio Merra. 2021. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks. *ACM Computing Surveys (CSUR)* 54, 2 (2021), 1–38.

[6] Honghui Du, Leandro L Minku, and Huiyu Zhou. 2020. Marline: Multi-source mapping transfer learning for non-stationary environments. In *2020 IEEE International Conference on Data Mining (ICDM)*. IEEE, 122–131.

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research* 17, 1 (2016), 2096–2030.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[10] Xiangnan He, Zhankui He, Xiaoyu Du, and Tat-Seng Chua. 2018. Adversarial personalized ranking for recommendation. In *The 41st International ACM SIGIR conference on research & development in information retrieval*. 355–364.

[11] Guangneng Hu, Yu Zhang, and Qiang Yang. 2018. Conet: Collaborative cross networks for cross-domain recommendation. In *Proceedings of the 27th ACM international conference on information and knowledge management*. 667–676.

[12] Liang Hu, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu, and Can Zhu. 2013. Personalized recommendation via cross-domain triadic factorization. In *Proceedings of the 22nd international conference on World Wide Web*. 595–606.

[13] Chenglin Li, Yuanzhen Xie, Chenyun Yu, Bo Hu, Zang Li, Guoqiang Shu, Xiaohu Qie, and Di Niu. 2023. One for All, All for One: Learning and Transferring User Embeddings for Cross-Domain Recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 366–374.

[14] Pan Li, Brian Brost, and Alexander Tuzhilin. 2022. Adversarial Learning for Cross Domain Recommendations. *ACM Transactions on Intelligent Systems and Technology* 14, 1 (2022), 1–25.

[15] Pan Li and Alexander Tuzhilin. 2020. Ddtcdr: Deep dual transfer cross domain recommendation. In *Proceedings of the 13th International Conference on Web Search and Data Mining*. 331–339.

[16] Jianxun Lian, Fuzheng Zhang, Xing Xie, and Guangzhong Sun. 2017. CCCFNet: A content-boosted collaborative filtering neural network for cross domain recommender systems. In *Proceedings of the 26th international conference on World Wide Web companion*. 817–818.

[17] Jae Hyun Lim and Jong Chul Ye. 2017. Geometric gan. *arXiv preprint arXiv:1705.02894* (2017).

[18] Weiming Liu, Xiaolin Zheng, Mengling Hu, and Chaochao Chen. 2022. Collaborative filtering with attribution alignment for review-based non-overlapped cross domain recommendation. In *Proceedings of the ACM Web Conference 2022*. 1181–1190.

[19] Jianxin Ma, Chang Zhou, Peng Cui, Hongxia Yang, and Wenwu Zhu. 2019. Learning disentangled representations for recommendation. *Advances in neural information processing systems* 32 (2019).

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[21] Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. 188–197.

[22] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. BPR: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618* (2012).

[23] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. 2018. Adversarially robust generalization requires more data. *Advances in neural information processing systems* 31 (2018).

[24] Ajit P Singh and Geoffrey J Gordon. 2008. Relational learning via collective matrix factorization. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. 650–658.

[25] Matthew Staib and Stefanie Jegelka. 2017. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, Vol. 3. 4.

[26] Caiqi Sun, Jiewei Gu, Binbin Hu, Xin Dong, Hai Li, Lei Cheng, and Linjian Mo. 2023. REMIT: Reinforced Multi-Interest Transfer for Cross-Domain Recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 9900–9908.

[27] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[28] Jinhui Tang, Xiaoyu Du, Xiangnan He, Fajie Yuan, Qi Tian, and Tat-Seng Chua. 2019. Adversarial training towards robust multimedia recommender system. *IEEE Transactions on Knowledge and Data Engineering* 32, 5 (2019), 855–867.

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

[30] Xiang Wang, Hongye Jin, An Zhang, Xiangnan He, Tong Xu, and Tat-Seng Chua. 2020. Disentangled graph collaborative filtering. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 1001–1010.

[31] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. 2018. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610* (2018).

[32] Ruobing Xie, Qi Liu, Liangdong Wang, Shukai Liu, Bo Zhang, and Leyu Lin. 2022. Contrastive cross-domain recommendation in matching. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 4226–4236.

[33] Yangjun Xu, Liang Chen, Fenfang Xie, Weibo Hu, Jieming Zhu, Chuan Chen, and Zibin Zheng. 2020. Directional Adversarial Training for Recommender Systems.. In *ECAI*. 553–560.

[34] Haoran Yan, Pengpeng Zhao, Fuzhen Zhuang, Deqing Wang, Yanchi Liu, and Victor S Sheng. 2020. Cross-domain recommendation with adversarial examples. In *Database Systems for Advanced Applications: 25th International Conference, DASFAA 2020, Jeju, South Korea, September 24–27, 2020, Proceedings, Part III 25*. Springer, 573–589.

[35] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. Adversarial collaborative auto-encoder for top-n recommendation. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[36] Feng Yuan, Lina Yao, and Boualem Benatallah. 2019. Adversarial collaborative neural network for robust recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1065–1068.

[37] Tianzi Zang, Yanmin Zhu, Haobing Liu, Ruohan Zhang, and Jiadi Yu. 2022. A survey on cross-domain recommendation: taxonomies, methods, and future directions. *ACM Transactions on Information Systems* 41, 2 (2022), 1–39.

[38] Xiaoyun Zhao, Ning Yang, and Philip S Yu. 2022. Multi-sparse-domain collaborative recommendation via enhanced comprehensive aspect preference learning. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. 1452–1460.

[39] Feng Zhu, Yan Wang, Chaochao Chen, Guanfeng Liu, and Xiaolin Zheng. 2020. A graphical and attentional framework for dual-target cross-domain recommendation.. In *IJCAI*. 3001–3008.

[40] Feng Zhu, Yan Wang, Chaochao Chen, Jun Zhou, Longfei Li, and Guanfeng Liu. 2021. Cross-domain recommendation: challenges, progress, and prospects. *arXiv preprint arXiv:2103.01696* (2021).

[41] Zhiwen Zuo, Lei Zhao, Ailin Li, Zhizhong Wang, Zhanjie Zhang, Jiafu Chen, Wei Xing, and Dongming Lu. 2023. Generative image inpainting with segmentation confusion adversarial training and contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 3888–3896.