Making Non-Overlapping Matters: An Unsupervised Alignment Enhanced Cross-Domain Cold-Start Recommendation

Zihan Wang [ID], Yonghui Yang [ID], Le Wu [ID], *Member, IEEE*, Richang Hong [ID], *Senior Member, IEEE*, and Meng Wang [ID], *Fellow, IEEE*

Abstract—Cold-start recommendation is a long-standing challenge when presenting potential preferred items to new users. Most empirical studies leverage side information to promote cold-start recommendation. In this work, we focus on cross-domain cold-start recommendation, which aims to provide suggestions to those non-overlapping users who have only interacted in the source domain and are viewed as new users in the target domain. Pre-training and then mapping is the common solution for the cross-domain cold-start recommendation. The former learns domain-specific user preference, and the latter transfers preference knowledge from the source to the target domain. Despite the effectiveness, we argue that current mapping-based methods still have the following limitations. First, current mapping functions fail to fully consider the similarity of user behavioral patterns, either common transfer or personalized transfer mappings. Second, sparse supervision signals from the limited overlapping users, lead to insufficient mapping function learning for recommendation. To tackle the above limitations, we propose a novel *MACDR* model for cross-domain cold-start recommendation. Specifically, *MACDR* consists of two elaborate modules: a Prototype enhanced Mixture-Of-Experts (PMOE) based mapping function and a Preference Distribution Alignment (PDA) enhanced optimization. PMOE is designed to balance the transfer patterns of common and personalized preferences, following the basis that similar users share similar preference transfer. Furthermore, to alleviate the sparse supervision issue, PDA is designed to explore the utilization of non-overlapping users in an unsupervised manner based on the prototype distribution alignment technique. Extensive experiments on three real-world datasets demonstrate the effectiveness of the proposed method.

Index Terms—Cross-domain recommendation, cold-start recommendation, transfer learning, mixture-of-experts, unsupervised learning, preference distribution alignment.

I. INTRODUCTION

PERSONALIZED recommendations [1], [4], [5], [8], [17], [38] have become an important information filtering tool
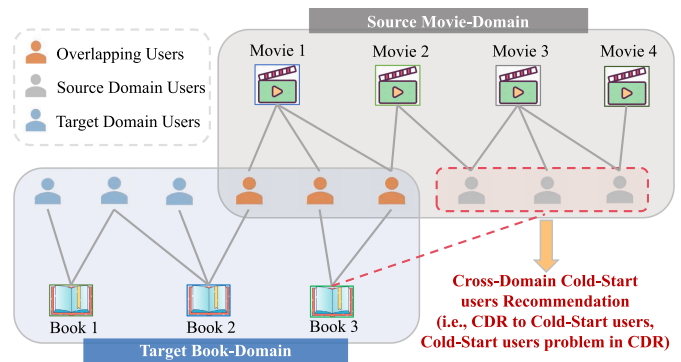
Fig. 1. Illustration of cross-domain cold-start users recommendation (i.e., cross-domain recommendation to cold-start users or cold-start users problem in CDR). There are a few overlapping users across the source domain (i.e., Movie domain) and the target domain (i.e., Book domain), the goal of our task is to recommend items in the target domain to those non-overlapping users in the source domain.

in various online applications for guiding users to discover products that appeal to them. However, they suffer from the longstanding cold-start problem, which hard to provide satisfying recommendations for cold-start users (i.e., new users without any historical interactions). Nowadays, web platforms are operating various service domains simultaneously (e.g., music streaming, game store, and eBook subscription). Users have more opportunities to interact with multiple platforms in their daily lives. Therefore, Cross-domain Recommendation (CDR) [46], [49] which aims to transfer knowledge from an informative source domain to the target domain becomes a promising solution to alleviate the cold-start problem.

Existing researchers mainly describe the task in three terms: cross-domain cold-start users recommendation, cross-domain recommendation for cold-start users, or cold-start users problem in CDR. To avoid misunderstandings or word misuse, we consistently use the phrase *cross-domain cold-start recommendation* to describe the task in the following sections. We also give a toy example as shown in Fig. 1 to illustrate cross-domain cold-start users recommendation [7], [18], [25], [41]. There are interactions from source domain (i.e., Movie domain) and target domain (i.e., Book domain). These two domains have a few overlapping users, and the others are non-overlapping users. Our goal is to recommend items in target domain to non-overlapping
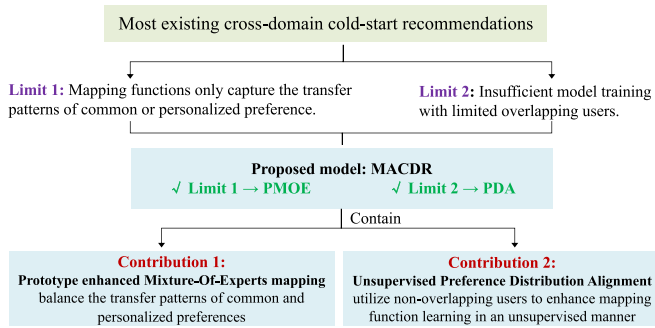
Fig. 2. Limitations of current cross-domain cold-start works and the contributions of our paper. We propose PMOE mapping module (i.e., Contribution 1) for solving Limit 1 and design unsupervised PDA optimization strategy (i.e., Contribution 2) for mitigating Limit 2.

users who have rich interactions in the source domain (which can be viewed as target domain cold-start users), as highlighted in the red dashed line. Most existing cross-domain cold-start methods follow the schema that preference pre-training and then mapping [18], [41], [50], [51]. Among them, pre-training aims to learn users' domain-specific preferences, and mapping transfer preference knowledge from the informative source to the target domain. After training the CDR model with the overlapping users, the mapping function can be used to transfer preference for those non-overlapping users, completing the recommendation process. Despite the effectiveness, we argue that current mapping-based CDR methods still have two limitations shown in Fig. 2.

Limit 1: Mapping function only captures the transfer patterns of common or personalized preference. Current mapping functions are sub-optimal either one-to-all (common transfer) [18], [22], [25] or one-to-one forms (personalized transfer) [41], [50], [51], which are difficult to adequately express interest preference transfer between domains. Considering the characteristics of recommendation, users with similar behavior patterns(interest) should share more common preference transfer [9], [24], [41]. Therefore, designing a mapping function by flexibly considering the transfer patterns of both common and personalized preferences is an emerging need.

Limit 2: Insufficient model training with limited overlapping users. Most of the current methods only use overlapping users to optimize the mapping function, which heavily relies on the number of overlapping users. Sparse supervision signals from the limited overlapping users lead to insufficient mapping function learning for recommendation. Some attempts [18], [34] incorporate all non-overlapping source user or interaction item embeddings into the training of the mapping function. However, these methods introduce some useless information, which is unstable to enhance recommendation performances [51]. Therefore, how to better utilize non-overlapping users to enhance model learning remains challenging.

To tackle the above limitations, we propose a novel *MACDR* model for the cross-domain cold-start users recommendation. Specifically, *MACDR* consists of two elaborate modules: a Prototype enhanced Mixture-Of-Experts (PMOE) based mapping function and a Preference Distribution Alignment (PDA)

enhanced optimization. PMOE function is designed to balance the transfer patterns of common and personalized preferences, following the basis that similar users share similar preference transfer. We first assign users to several prototypes by clustering. Then we utilize expert networks to process the diverse inputs from different individual users to capture the transfer pattern of personalized preference. Meanwhile, we feed the prototype embeddings as input to learn the gate logit. Thus, users with the same prototype will have the same logits, which capture the transfer pattern of common preference. Both determine the final mapping functions. To alleviate the sparse supervision issue, we combine the non-overlapping users for model learning in an unsupervised manner. Specifically, we build a preference distribution alignment based on equivariant learning. We assume that users with the same prototype have a closer representation distance after mapping, while the counterparts have a larger representation distance. Based on the above idea, we implement PDA based on generator and discriminator learning. The mapping function is just like a generator to obtain approximate target domain preference. After that, we use a discriminator to distinguish positive and negative examples, which are composed of two mapped user representations. If two users come from the same prototype, their mapped representations construct a positive sample, and vice versa. The major contributions of this work are summarized as follows:

- We propose a novel *MACDR* model for the cross-domain cold-start recommendation, which consists of the elaborate mapping function and preference distribution alignment.
- We design a Prototype enhanced Mixture-Of-Experts (PMOE) mapping function, which can flexibly combine the transfer patterns of common and personalized preferences.
- We devise an unsupervised Preference Distribution Alignment (PDA) optimization strategy to better exploit the non-overlapping users, which can effectively alleviate the sparse supervision issue from the limited overlapping users.
- Extensive experiments conducted on three real-world datasets demonstrate the effectiveness of our proposed *MACDR*, including high performances and generality of the designed PDA optimization strategy.

II. RELATED WORK

A. Cold-Start Recommendation

Recommender systems [14], [23], [32], [37], [42], [43] face the persistent challenge of providing accurate recommendations for new users or items on web platforms, which is known as cold-start recommendations [2], [3], [27], [39], [40]. One common approach is to use additional contextual information to reduce dependence on interaction data. Specifically, content-based recommenders usually transform user attributes into preference representations for improving the cold-start recommendation performance. DropoutNet [31] and MTPR [11] adopt a dropout strategy that randomly inputs either preference or attribute representations in the training stage to imitate cold-start situations. Some methods attempt to use pre-trained representations to

TABLE I
MATHEMATICAL NOTATIONS

Notation	Description
n_s, n_t	The number of users in the source and target domains.
m_s, m_t	The number of items in the source and target domains.
n_o	The number of overlapping users.
K	The number of prototype centers.
$\mathcal{U}^s, \mathcal{U}^t$	Userset in the source and target domains.
$\mathcal{V}^s, \mathcal{V}^t$	Itemset in the source and target domains.
$\mathcal{R}^s, \mathcal{R}^t$	Rating matrices in the source and target domains.
$\mathcal{U}^o$	Overlapping users across two domains.
$\mathbf{U}^s, \mathbf{V}^s$	Pre-trained source user and item representations.
$\mathbf{U}^t, \mathbf{V}^t$	Pre-trained target user and item representations.
$\mathbf{C}$	Source user prototype representations.
$\mathbf{P}$	Source interacted item sequence representations.
$\mathcal{L}_r$	Cross-domain supervised rating prediction loss.
$\mathcal{L}_{dis}$	Cross-domain unsupervised discriminator loss.

better capture the correlation and reduce the difference between CF signals and content features [36], [52]. For example, Heater [52] extracts content representations and uses the sum squared error loss to align pre-trained CF representations and content representations. In addition to content attributes, knowledge graphs [12], social networks [33], and auxiliary domains [25], [48], [51] are also considered effective external preference knowledge to mitigate the cold start problem. Considering that in real-life scenarios, users visit various platforms in their daily lives, and different platforms are usually interconnected, it is relatively easy to obtain users' interaction records on other platforms. Therefore, in this paper, we focus on utilizing other domain interaction records to help target domain cold-start users for recommendations.

B. Cross-Domain Cold-Start Recommendation

Cross-domain recommendation [6], [7], [18], [25] is an emerging direction to solve cold-start recommendation. Existing methods can be broadly divided into three categories: matrix factorization methods, content-based methods, and mapping-based methods. In the initial phase, CMF [28] and CDCF [21] use shared user matrices across all domains and apply matrix factorization algorithms [20] to learn multiple rating matrices simultaneously, thus advancing cross-domain recommendation techniques. CBMF [26] first devises a cluster-level cross-domain matrix to learn the correlation between user clusters and item clusters. Subsequently, content-based methods are proposed. CATN [47] proposes an end-to-end framework to capture the aspect-level correlation between user reviews and item descriptions across domains. However, all domain interaction data and feature information are required for matrix factorization and content-based methods during training. Some researchers argue that privacy protection exists for each domain's information in real-life scenarios, making it difficult to obtain interaction

information from two domains simultaneously. Therefore, pre-training and then mapping (mapping-based) CDRs are proposed to transfer pre-trained embeddings across domains by various methods. This paradigm is based on the assumption that there is a correlation between a user's interest preferences in the source domain and their interest preferences in the target domain. The steps of the pipeline's abbreviated description are: (1) first pre-training user embedding of source/target domain, (2) learning a mapping function with the overlapped users, and (3) predicting items in target domain for source users. EMCDR [25] is first proposed as a pre-training and then mapping paradigm. it learns a mapping function to infer latent factors for new items/users in the target domain. Subsequently, many models followed the framework for further improvements. DCDCSR [48] considers the sparsity degree to construct benchmark factors with top-K similar entities first and trains a model to map latent factors into these factors. SSCDR [18] is a semi-supervised method, which utilizes records information of users and item interactions in the source domain to improve the robustness of the mapping function. LACDR [34] employs the framework of autoencoder to improve the previous problem of over-simplification of the mapping function and utilizes the reconstruction loss and alignment loss to leverage all non-overlapping users into the learning of the mapping function. Inspired by successful applications of meta-learning, the recently proposed TMCDR [50], PTUPCDR [51] and HCCDR [41] follow the MAML [13] framework to learn a meta-network that customizes the mapping function to achieve better performances instead of the traditional linear mapping. Although existing pre-training and then mapping models are effective, most of them use simple linear mapping functions and fail to fully leverage non-overlapping users' preferences, ultimately leading to suboptimal performance. Our proposed model extends the original mapping function and makes better use of non-overlapping users to generate more effective training signals for mapping function learning.

III. PRELIMINARIES

A. Problem Statement

In CDR, we have a source domain and a target domain. Each domain has a userset $\mathcal{U} = \{u_1, u_2, \ldots\}$, an itemset $\mathcal{V} = \{v_1, v_2, \ldots\}$, and a rating matrix $\mathbf{R}$. The source domain has n_s users, m_s items and the target domain has n_t users, m_t items. $r_{ij} \in \mathbf{R}$ denotes the rating between user i and item j. To distinguish two domains, we use $\mathcal{U}^s, \mathcal{V}^s, \mathcal{R}^s$ to denote the userset, itemset, and rating matrix of the source domain, while $\mathcal{U}^t, \mathcal{V}^t, \mathcal{R}^t$ denotes for the userset, itemset, and rating matrix the target domain. We define the overlapping users between the two domains as $\mathcal{U}^o = \mathcal{U}^s \cap \mathcal{U}^t$. In contrast, $\mathcal{V}^s$ and $\mathcal{V}^t$ are disjoint, which means there are no shared items between the two domains.

Given the above data, our goal is to predict the unknown non-overlapping users' preferences for the items in target domain: $\hat{r}_{ij} = f(u_i^s, v_j^t)$, where $u_i^s \in \mathcal{U}^s \& \notin \mathcal{U}^o$ denote each non-overlapping user and v_j^t denote each target domain item.
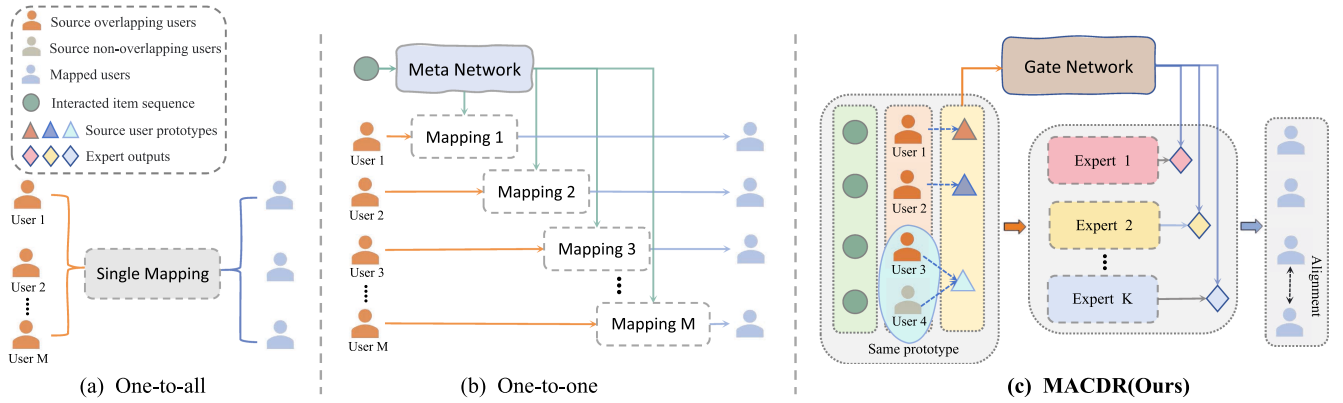
Fig. 3.　Illustration of three mapping functions. (a) One-to-all [18], [22], [25]: all users share the same mapping function; (b) One-to-one [41], [51]: each user has a personalized mapping function; (c) MACDR (Ours): we propose a Prototype enhanced Mixture-Of-Experts (PMOE) based mapping function to capture the transfer patterns of common and personalized preferences and design an unsupervised preference distribution alignment (PDA) optimization to better exploit the non-overlapping users, following the assumption that users with the same prototype have closer representation after mapping.

B. Schema of the Pre-Training & Mapping

Considering the scenario of data isolation in practical applications, that is, the interactive information between users and items in a specific domain is not observable by other domains, pre-training and then mapping is the common solution in cross-domain cold-start recommendation [18], [22], [25], [41], [51]. Among them, pre-training aims to learn the domain-specific preference representations, while mapping aims to transfer preference knowledge from the source to the target domain. As this is the basic paradigm, we first present the overall architecture.

1) Domain-Specific Preference Pre-Training: Given interaction data in both domains, we first learn domain-specific preferences with SOTA methods, i.e., Matrix Factorization [20] and LightGCN [16]. Without generality, we use $Pre()$ to denote the used pre-training model, and then we obtain the pre-trained source (target) domain preference representations:

$$\mathbf{U}^s, \mathbf{V}^s = Pre\left(R^s, \theta_1\right)$$
$$\mathbf{U}^t, \mathbf{V}^t = Pre\left(R^t, \theta_2\right), \tag{1}$$

where $\mathbf{U}^s \in \mathbb{R}^{n_s \times d}, \mathbf{V}^s \in \mathbb{R}^{m_s \times d}$ are pre-trained user and item embedding matrices in the source domain, $\mathbf{U}^t \in \mathbb{R}^{n_t \times d}, \mathbf{V}^t \in \mathbb{R}^{m_t \times d}$ are pre-trained user and item embedding matrices in the target domain. θ_1 and θ_2 denote the parameters of the used pre-training model in source and target domains, respectively.

2) Domain Preference Mapping: Given the pre-trained user preferences, the mapping function aims to transfer preference knowledge from the information-rich source domain to the target domain. There are two popular mapping functions in CDR: one-to-all mapping [18], [25] and one-to-one mapping [50], [51], which are illustrated in Fig. 3(a), (b). The first is one-to-all mapping, which assumes that all overlapping users share the same preference transfer:

$$\hat{\mathbf{U}}^t = Map(\mathbf{U}^s), \tag{2}$$

where $Map()$ denotes the generic mapping function, which is usually implemented by Multi-Layer Perceptrons (MLPs).

Besides, some works argue that one-to-all mapping is not effective, and the mapping function should be personalized to each user [51]. Thus, the second kind of mapping function is one-to-one, which is defined as follows:

$$\hat{\boldsymbol{u}}_1^t = MLP_1(\boldsymbol{u}_1^s)$$
$$\hat{\boldsymbol{u}}_2^t = MLP_2(\boldsymbol{u}_2^s)$$
$$\hat{\boldsymbol{u}}_{n_o}^t = MLP_{n_o}(\boldsymbol{u}_{n_o}^s), \tag{3}$$

where n_o denotes the number of overlapping users, each user has personalized mapping MLPs. Because learning n_o MLPs is too expensive, a Meta-Network method has been proposed to learn personalized mapping functions [51].

3) Model Optimization: After obtaining the transferred user embeddings, There are two main methods for optimization. One is to minimize the distance between the transferred user embeddings and the pre-trained target user embeddings [18], [25]. This distance-oriented optimization is defined as follows:

$$\mathcal{L}_r = \sum_{u_i \in \mathcal{U}^o} ||\hat{\boldsymbol{u}}_i^t - \boldsymbol{u}_i^t||^2, \tag{4}$$

where $\hat{\boldsymbol{u}}_i^t$ denotes the transformed user embedding in the source domain. and $\boldsymbol{u}_i^t \in \mathbf{U}^t$ denotes the pre-train user embedding in the target domain. Another popular optimization is task-oriented, which constrains the transferred user embeddings to reconstruct the true preferences in the target domain [51]. This task-oriented optimization is defined as follows:

$$\mathcal{L}_r = \frac{1}{|\mathcal{R}_o^t|} \sum_{r_{ij} \in \mathcal{R}_o^t} \left(r_{ij} - \hat{\boldsymbol{u}}_i^t \boldsymbol{v}_j^t\right)^2, \tag{5}$$

where $\mathcal{R}_o^t = \{r_{ij}|u_i \in \mathcal{U}^o, v_j^t \in \mathcal{V}^t\}$ denotes the interactions of overlapping users in the target domain.

After minimizing the loss function ((4) or (5)), we have the optimized mapping function, which can be applied to non-overlapping users. Given the above, we obtain the approximated user representation in the target domain, to serve the recommendation process.
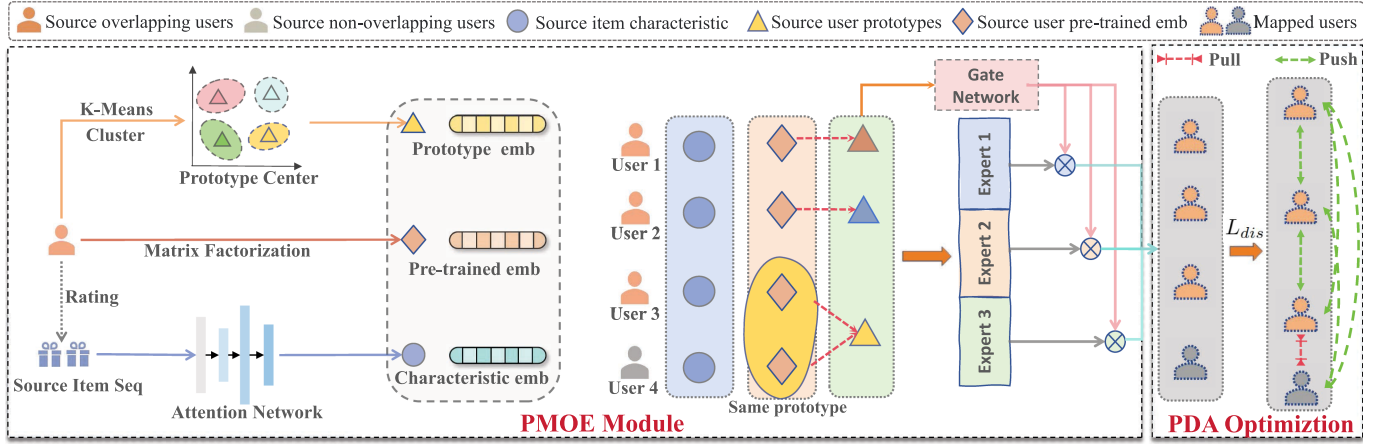
Fig. 4. An Illustration of our proposed *MACDR* framework, which consists of Prototype enhanced Mixture-Of-Experts (PMOE) mapping module and Preference Distribution Alignment (PDA) optimization strategy. The PMOE module provides a flexible mapping function that combines the patterns of common and personalized preference transfer based on user prototypes. The PDA optimization strategy utilizes non-overlapping users to boost recommendation performance in an unsupervised manner.

IV. METHODOLOGY

A. Architecture Overview of MACDR

As illustrated in Fig. 4, *MACDR* consists of two elaborate modules: a Prototype enhanced Mixture-Of-Experts (PMOE) mapping module and an unsupervised Preference Distribution Alignment (PDA) optimization. Among them, PMOE considers the transfer patterns of both common and personalized preferences for the mapping function, and PDA further exploits non-overlapping users to enhance model learning. In the following, we introduce each module in detail. Last, we give the overall learning process of *MACDR*.

B. Prototype Enhanced Mixture-of-Experts Mapping

Here, we introduce our proposed PMOE mapping function. Given the pre-trained domain-specific preference, PMOE can flexibly combine the transfer patterns of common and personalized preferences for CDR tasks.

We use prototype learning to formulate user similarity. Without loss of generality, we use $\mathbf{U}^o, \mathbf{U}^s, \mathbf{U}^t$ to denote embedding matrices of overlapping users, users in source domain, and users in target domain, respectively. Suppose there are K prototypes $\mathbf{C} \in \mathbb{R}^{K \times d}$ of the pre-trained source user embeddings $\mathbf{U}^s$, we assign these users to K prototypes by maximize the conditional probability:

$$\mathbf{c}_i = \arg \max_k p\left(\mathbf{c}^k | u_i^s\right), \qquad (6)$$

where $\mathbf{c}^k$ denotes the kth prototype, u_i^s denotes the ith source domain user, and $p(\mathbf{c}^k | u_i^s)$ denotes the probability that user u_i^s belong to prototype $\mathbf{c}^k$. We employ K-Means clustering to implement the above process.

After obtaining the prototype information, we model the user's unique characteristics, including interacted item sequence and pre-trained personalized preference. Following [51], different interacted items contribute differently to user preference transfer. We utilize the attention mechanism to formulate the

user's interaction sequence. Formally, we denote the list of user i sequential interaction items in source domain by $\mathcal{S}_i = \{v_1^s, v_2^s, \dots, v_l^s\}$, where l denotes the number of interacted items and each $v_j^s \in \mathcal{S}_i$ denotes the interacted item in the source domain. The attention network is defined as:

$$\begin{cases} \bar{a}_j^s = att(\boldsymbol{v}_j^s; \theta_a) \\ a_j^s = \dfrac{\exp(\bar{a}_j^s)}{\sum_{v_k^s \in \mathcal{S}_i} \exp(\bar{a}_k^s)}, \end{cases} \qquad (7)$$

where $att(\cdot)$ denotes the attention network which consists of a two-layer feed-forward network, and θ_a denotes the parameters of the attention network. a_j^s is the normalized attention score for item v_j^s, which can be interpreted as the importance of v_j^s in predicting the personalized mapping function. Then the user's interaction sequence is weighted and summed to obtain the user characteristic representation.

$$\boldsymbol{p}_i = \sum_{v_j^s \in \mathcal{S}_i} a_j^s \boldsymbol{v}_j^s, \qquad (8)$$

where $\boldsymbol{p}_i \in \mathbb{R}^d$ denotes the transferable characteristic embedding of user i.

Given the pre-trained user representation, user prototype representation, and characteristic representation of the interaction sequence, we fuse them as inputs to learn each expert output. This process is formulated as follows:

$$\boldsymbol{z}_i = \boldsymbol{u}_i^s \oplus \boldsymbol{c}_i \oplus \boldsymbol{p}_i, \qquad (9)$$

where $\boldsymbol{z}_i \in R^d$ denotes the input to the expert network for user i. Then the specific form of the mixture-of-expert mapping function is defined as follows:

$$\begin{cases} f_{\text{gate}}(\mathbf{Z}) = \text{Softmax}(\mathbf{W}_g \mathbf{Z} + \mathbf{b}_g) \\ h_{\text{expert}}(\mathbf{C}) = \sigma(\mathbf{W}_h \mathbf{C} + \mathbf{b}_h) \\ \hat{u}_i^t = \sum_{x=1}^{X} f_x(\boldsymbol{c}_i) \cdot h^x(\boldsymbol{z}_i), \end{cases} \qquad (10)$$

where $f_{\text{gate}}(\mathbf{C})$ and $f_{\text{expert}}(\mathbf{Z})$ represent formulas for gate and expert networks, which are both composed of multiple linear

layers and activation functions. The source user prototype representation matrix C is the input to the gate network. $\mathbf{W}_g$ and $\mathbf{b}_g$ are feature transformation matrix and bias matrix, and $\mathbf{W}_h$ and $\mathbf{b}_h$ are similar definitions in experts networks. σ is the activation function, and X is a hyper-parameter denoting the number of experts. $\hat{u}_i^t$ is the mapped target domain representation we obtained, which reflects the approximated user target domain preferences.

The design benefits of PMOE inputs are as follows: First, our designed PMOE mapping function escapes from the previous simple linear function form. Most importantly, we feed the prototype embeddings as input to learn the gate logit. Thus, users with the same prototype will get the same gate logit, which can capture the transfer pattern of common preference based on prototype guidance. while MOE networks are suitable for processing different types of inputs with different experts. We input different types of information from the source domain into the expert network to capture a personalized preference relationship from the source domain to the target domain. Based on the above, our designed PMOE module can not be fully common or personalized, but flexibly combine the preferences of both information to obtain better cold-start mapped representations in CDR.

C. Preference Distribution Alignment

After obtaining the approximated user preference in the target domain, most of the current methods optimize the model by aligning the users' real feedback in the target domain. Despite the effectiveness, their performances heavily rely on the limited overlapping users. Here, we introduce our proposed PDA, an unsupervised preference distribution alignment optimization, which can utilize non-overlapping users in a more rational way to facilitate mapping functions in learning. Specifically, PDA is based on equivariant learning, that users with the same prototype are also more similar in the mapped representation space. We implement PDA with generator and discriminator learning. First, We sample a user u_b^s of the same prototype as u_a^s and N_s non-overlapping users of different prototypes (i.e., u_c^s) in the source domain. In the following description, the value of N_s is set to 1. Then we view the PMOE as a generator that produces approximated user representation $\hat{\mathbf{U}}^t$ in the target domain, and use a discriminator to distinguish positive or negative user pairs based on their prototypes. In the following, we detail present how to optimize the discriminator in an unsupervised manner.

Generator: For each user tuple $< u_a^s, u_b^s, u_c^s >$, we obtain their approximated preference embedding by PMOE:

$$\begin{cases} \mathbf{z} = \mathbf{u}^s \oplus \mathbf{p} \oplus \mathbf{c}, \\ \hat{\boldsymbol{u}}_a^t = \text{PMOE}(\mathbf{z}_a, \mathbf{c}_a) \\ \hat{\boldsymbol{u}}_b^t = \text{PMOE}(\mathbf{z}_b, \mathbf{c}_b), \quad c_b = c_a \\ \hat{\boldsymbol{u}}_c^t = \text{PMOE}(\mathbf{z}_c, \mathbf{c}_c), \quad c_c \neq c_a, \end{cases} \quad (11)$$

where $\mathbf{u}^s$, $\mathbf{p}$, $\mathbf{c}$ denote the pre-trained, interacted item sequence and prototype representation. Through (11), we can use PMOE module to get the overlapping users u_a^s, source domain users of the same prototype u_b^s, and different prototype users u_c^s mapped to the target domain embedding $\hat{\boldsymbol{u}}_a^t$, $\hat{\boldsymbol{u}}_b^t$ and $\hat{\boldsymbol{u}}_c^t$, respectively.

Discriminator: Given the approximated user representation in the target domain, we concatenate each user pair's approximated representation as the input to the discriminator $\mathcal{D}$. For two users with the same prototype, the discriminator outputs a "real" value, otherwise, it outputs a "fake" value. The optimization is defined as follows:

$$L_{dis} = \min_{\theta_d} \mathbb{E}_{u_a^s \sim U^o, u_c^s \notin c_a} \left[\log \left(\mathcal{D} \left(\hat{u}_a^t || \hat{u}_c^t \right) \right) \right]$$
$$- \mathbb{E}_{u_a^s \sim U^o, u_b^s \in c_a} \left[\log \left(\mathcal{D} \left(\hat{u}_a^t || \hat{u}_b^t \right) \right) \right]. \quad (12)$$

In practice, we use 3-layer MLPs to implement the discriminator $\mathcal{D}$, where θ_d is the parameters of the discriminator $\mathcal{D}$. Based on the discriminator, we involve non-overlapping users in model training, which can effectively mitigate the sparse supervision issue. Finally, we present the overall learning process of *MACDR*.

D. Overall Learning of MACDR

1) Domain-Specific Preference Pre-Training: We follow previous work [18], [25] in adopting the widely used Matrix Factorization(MF) [20] as the pre-training model to obtain the user and item embedding matrices in each domain. Specifically, we use $\mathbf{U}^s \in \mathbb{R}^{n_s \times d}$, $\mathbf{V}^s \in \mathbb{R}^{m_s \times d}$ to denote pre-trained user and item embedding matrices in the source domain, and $\mathbf{U}^t \in \mathbb{R}^{n_t \times d}$, $\mathbf{V}^t \in \mathbb{R}^{m_t \times d}$ to denote pre-trained user and item embedding matrices in the target domain. Please note that, with the development of recommendation techniques, better CF methods can obtain better user and item preference representations, such as graph-based model (LightGCN [16] and self-supervised model (VGCL [44]). Here, for a fair comparison, we employ MF as the pre-training model.

2) Domain Preference Mapping: After obtaining the pre-trained source domain user embeddings $\mathbf{U}^s$, we deploy our proposed PMOE mapping function to learn the approximate user embeddings. The brief process of PMOE includes user prototype learning via K-Means clustering, user-interacted sequence modeling with attention network, and prototype enhanced MOE learning. Thus, we obtain the approximated user embeddings in the target domain $\hat{\mathbf{U}}^t$.

3) Model Optimization: Given the approximated user embeddings, we adopt task-oriented optimization [51], which is supervised by overlapping users' ratings in the target domain. The supervised loss is defined as follows:

$$\mathcal{L}_r = \frac{1}{|\mathcal{R}_o^t|} \sum_{r_{ij} \in \mathcal{R}_o^t} (r_{ij} - \hat{\boldsymbol{u}}_i^t \boldsymbol{v}_j^t)^2, \quad (13)$$

where $\mathcal{R}_o^t = \{r_{ij} | u_i \in \mathcal{U}^o, v_j^t \in \mathcal{V}^t\}$ denotes the interactions of overlapping users in the target domain. In addition to the supervised rating loss, we combine our proposed PDA loss (12) to involve the non-overlapping users in model training. Therefore, we employ a multi-task learning framework to optimize our proposed *MACDR*:

$$\mathcal{L}(\Theta) = \mathcal{L}_r + \alpha \mathcal{L}_{dis} + \lambda ||\Theta||_2^2, \quad (14)$$

where L_r is supervised rating prediction loss, $\mathcal{L}_{dis}$ is unsupervised discriminator loss (defined as (12)). α is the balance

parameter between two losses, λ is the regularization coefficient, Θ is all model parameters of *MACDR*.

4) Inference: For the non-overlapping cold-start users in the source domain, we first obtain their pre-trained embeddings, and learn their approximated preference in the target domain based on our proposed PMOE mapping. Finally, we employ the widely used inner product to compute the score of ith cold-start user to j^{th} target item:

$$\hat{r}_{ij} = \hat{\mathbf{u}}_i^t \mathbf{v}_j^t. \tag{15}$$

The overall implement of *MACDR* is illustrated in the Algorithm 1.

Algorithm 1: The Algorithm of *MACDR*.

Input: Source domain interactions $\mathcal{R}^s$, Target domain interactions $\mathcal{R}^t$, Overlapping users $\mathcal{U}^o$;

Pre-training Stage (based on MF):

1: Learn pre-trained user and item embedding matrices $\boldsymbol{U}^s$ and $\boldsymbol{V}^s$ in the source domain;

2: Learn pre-trained item embedding matrices $\boldsymbol{V}^t$ in the target domain;

Mapping Stage:

3: Compute user prototypes via K-Means (6);

4: Calculate interacted item sequences via (7)–(8);

While not converged

5: Compute the user's target domain preference representation using PMOE mapping via (9)–(11);

7: Calculate unsupervised discriminator loss via (12);

8: Calculate supervised rating prediction loss via (13);

9: Update all parameters according to (14).

End While

Inference Stage:

10: For a non-overlapping cold-start user u^s in the source domain, we obtain the approximated preference in the target domain (10), then compute the rating score via (15).

E. Model Analysis

In this section, we discuss the complexity of our proposed model in terms of space complexity and time complexity. We use n, m to denote the number of users and items, subscripts s and t to denote specific domains, s is the source domain, t is the target domain, d denotes the dimension of embeddings.

1) Space Complexity: As illustrated in Algorithm 1, the trainable parameters of *MACDR* are composed of two parts: embedding parameters and network parameters. Embedding parameters include user embeddings and item embeddings on the source domain as well as the target domain, all embedding parameters can be denoted as $\Theta_e = \{\mathbf{E}_u^s, \mathbf{E}_i^s, \mathbf{E}_u^t, \mathbf{E}_i^t\}$. Network parameters include attention network parameter Θ_a, prototype enhanced mixture-of-experts (PMOE) mapping network parameters $\Theta_p = \{\mathbf{W}_g, \mathbf{W}_h\}$ and domain discriminator network parameter Θ_D. In fact, our model follows the pre-training-then-mapping paradigm, while most of the pre-trained networks such as NGCF [35] and GAT [30] require additional model parameters

to be trained. This means that the bottleneck for training space complexity is usually in the pre-training module, and there is not much increase in storage space when compared to current mainstream models. Meanwhile, the pre-trained model only needs to be trained once in both the source and the target domains before it can be reused, which is convenient and affordable for recommender systems. To sum up, our model needs to learn embeddings of $(n_s + m_s + n_t + m_t)d$ size and a few parameters for multi-layer linear perceptron (MLP) networks. As $d \ll min(n_s, m_s, n_t, m_t)$, our proposed model can be considered fairly light.

2) Time Complexity: The total time cost consists of two parts, pre-training and our method cost. For the pre-training phase, training time is the same as the general CF methods. The exact time consumption depends on the pre-trained model used(e.g., MF, LightGCN). For the mapping phase of our method, Our time is consumed in three main aspects. 1) we first need to cluster the users in the source domain and use the K-means algorithm to get the K prototype centers of the users with a method time complexity of $O(n_s dK)$. 2) Second, we need to train the PMOE mapping function for overlapping users of the two domains and sample non-overlapping users of the same prototype and different prototypes. 3) Finally, we utilize the score prediction loss function and discriminator loss to predict the scores of all existing interaction records.

We implement the K-means clustering algorithm with Faiss-GPU,[1] and this operation is performed only once in the whole training process, it cannot cause excessive time consumption. Meanwhile, prototype sampling, mapping function learning, and computation of losses are done on the mini-batch, considering the sparse feedback on the target domain, the time consumption of this step is negligible.

To further illustrate the operational efficiency of our model, we recorded the time of a particular experiment for Task 3 in Table III. The format shows the total time on the left and the average time per epoch on the right. Our model has more user and item interaction data to learn the more complex mapping functions due to the need to sample both the same and different prototype users. Thus it is observed that our model has a similar runtime and convergence speed as models that need to utilize non-overlapping entities (LACDR [34], SSCDR [18]) to improve mapping function learning. However, compared to the learning of mapping functions, the pre-training process of the source and target domains tends to have more epochs and longer time, so the running time of our method is acceptable.

V. EXPERIMENTS

A. Experiments Settings

1) Datasets: To evaluate the effectiveness of the proposed *MACDR*, we conduct experiments on the large-scale Amazon datasets [15]. Following the existing works [41], [45], [51], we select the Amazon-5cores datasets,[2] which filter users/items with less than 5 ratings. Then, we define three CDR tasks: (1)

---

[1][Online]. Available: https://faiss.ai/

[2][Online]. Available: https://jmcauley.ucsd.edu/data/amazon/links.html/

TABLE II
STATISTICS OF THE CROSS-DOMAIN TASKS ("OVERLAP" DENOTES THE NUMBER OF OVERLAPPING USERS)

CDR Tasks	Domain		User			Item		Rating		Density	
	Source	Target	Overlap	Source	Target	Source	Target	Source	Target	Source	Target
Task 1	Sports	Clothing	3,908	35,598	39,387	18,357	23,033	296,337	278,677	0.045%	0.030%
Task 2	Movie	Music	18,031	123,960	75,258	50,052	64,443	1,697,533	1,097,592	0.027%	0.023%
Task 3	Book	Movie	37,388	603,668	123,960	367,982	50,052	8,898,041	1,697,533	0.004%	0.027%

TABLE III
THE COMPARISON OF METHODS AND RUNTIME EFFICIENCY

Methods	CPT	PPT	NOUI	Train time(s)
EMCDR	✓	✗	✗	116s/4.14s
SSCDR	✓	✗	✗	3894s/97.35s
LACDR	✓	✗	✓	2040s/68.00s
PTUPCDR	✗	✓	✗	159s/12.23s
HCCDR	✓	✓	✗	266s/16.63s
MACDR	✓	✓	✓	3313s/73.62s

"CPT" represents the common preference transfer, "PPT" represents personalized preference transfer, "NOUI" represents non-overlapping users' information.

Task 1: Sports → Clothing; (2) Task 2: Movie → Music; (3) Task 3: Book → Movie. There are only a certain percentage of users overlap in both domains. The statistics of three cross-domain datasets are summarized in Table II.

2) Evaluation Metrics and Baselines: All used datasets have explicit feedback. Users rate each item with a 0-5 score. To evaluate the performances of the proposed *MACDR* with other baselines, we employ two widely used evaluation metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). We compare our proposed *MACDR* with the following state-of-the-art baselines:

- *TGT* denotes a simple MF model, which is only trained on target domain data.
- *CMF* [28] is an extension of MF. In CMF, the embeddings of users are shared across the source and target domains.
- *EMCDR* [25] is a classic CDR model that adopts a network as the mapping function to transfer the user embeddings from the source domain into the target domain.
- *SSCDR* [18] considers that the proportion of common users or items between two domains is usually rare in real life, leading to poor mapping functions. Thus, SSCDR learns mapping functions via items in source domain in a semi-supervised method to enhance the robustness of the learned functions.
- *LACDR* [34] employs an encoder-decoder structure to learn the mapping function by inputting all source domain user representations. Then it aligns the low-dimension embedding spaces of different domains, leading to better generalization.
- *PTUPCDR* [51] is a state-of-the-art single-target CDR method. Differing from EMCDR, PTUPCDR learns a meta-network fed with users' characteristic embeddings to generate personalized mapping functions to achieve personalized transfer of preferences for each user.

- *HCCDR* [41] generates high-quality user and item representations through heterogeneous latent factor modeling, relying on diverse semantic relations. Following the PTUPCDR framework, it employs a meta-network to learn domain-specific transfer functions, taking into account both individual and shared user characteristics.

3) Implement Details: We implement our *MACDR* model and all baselines with Pytorch.[3] We initialize all model parameters with a Gaussian distribution with a mean value of 0 and a standard variance of 0.01, embedding size is fixed to 32. For each method, We use Adam [19] as the optimizer for model optimization, and the learning rate is tuned by grid searches within {0.0005, 0.001, 0.005, 0.01, 0.1}. The batch size is 512 for all tasks. We adjust the number of expert networks among {1, 2, 4, 8, 16, 32}. For discriminator learning of the proposed PDA, we tune the number of prototypes among {0, 50, 100, 150, 200, 250} and sample different prototype users with different ratios for three tasks. Specifically, we adopt a 1:5 ratio for task1 and 1:1 for task2 and task3 due to consideration of time efficiency and the number of overlapping users. For a fair comparison, we use Matrix Factorization (MF) as the pre-training approach for all models. For all parameters, we give more detailed setup explanations and experiments in the parameter analysis.

Following [51], we randomly select a proportion of overlapping users who are regarded as the bridge between the two domains (training the mapping function). The other overlapping users remove all their interactions in the target domain as the cold-start users for evaluation. In our experiments, we set the proportions of training overlapping users β as 20%, 50%, and 80% of the total overlapping users, respectively. For all baselines, we search the parameters carefully for fair comparisons. We repeat all experiments 5 times and report the average results.

B. Performance Comparisons

We compare our model with seven baselines in three CDR tasks. Table IV reports the results of MAE and RMSE [29] on the three CDR scenarios. From Table IV, we have the following observations:

- TGT only uses data from the trainset of target domain, which means the test user embeddings are untrained due to data sparsity. As a result, it can not accurately obtain the representations of the cold-start users, resulting in poor performance compared to other cross-domain recommendation models.

---

[3][Online]. Available: https://www.pytorch.org/

TABLE IV
RECOMMENDATION PERFORMANCE OF OUR PROPOSED *MACDR* AND OTHER BASELINES

	β	*Metric*	TGT	CMF	SSCDR	EMCDR	LACDR	PTUPCDR	HCCDR	MACDR	Improve
Task 1	20%	MAE	4.3874	1.5572	1.8001	1.5372	1.9178	1.3276	<u>1.3154</u>	**1.1548***	12.21%
		RMSE	5.0311	1.8533	2.0839	1.8254	2.1618	1.7068	<u>1.6970</u>	**1.5303***	9.82%
	50%	MAE	4.4119	1.1255	1.1733	1.1461	1.2128	1.0242	<u>1.0145</u>	**0.9281***	8.51%
		RMSE	5.0556	1.4483	1.4305	1.4104	1.4352	1.3266	<u>1.3130</u>	**1.2248***	6.72%
	80%	MAE	4.3797	1.0832	1.1360	1.0833	1.1487	0.9773	<u>0.9644</u>	**0.8760***	9.17%
		RMSE	5.0667	1.4041	1.3870	1.3439	1.3647	1.2769	<u>1.2683</u>	**1.1768***	7.21%
Task 2	20%	MAE	4.5821	1.6490	1.2492	1.1997	1.2692	1.1313	<u>1.1062</u>	**0.9597***	13.24%
		RMSE	5.3703	2.1677	1.5616	1.5183	1.5531	1.4922	<u>1.4655</u>	**1.3065***	10.85%
	50%	MAE	4.5756	1.6470	1.0300	1.0499	1.0133	0.9613	<u>0.8969</u>	**0.8151***	9.12%
		RMSE	5.3594	2.1662	1.3025	1.3153	1.2545	1.2618	<u>1.2058</u>	**1.0928***	9.37%
	80%	MAE	4.4873	1.6267	1.0189	1.0051	0.9822	0.9299	<u>0.8365</u>	**0.7734***	7.54%
		RMSE	5.2890	2.1559	1.2812	1.2675	1.2113	1.2128	<u>1.1247</u>	**1.0409***	7.45%
Task 3	20%	MAE	4.3086	1.7789	1.1247	1.1361	1.0887	1.1277	<u>1.0387</u>	**0.8997***	13.38%
		RMSE	4.9952	2.3458	1.4336	1.4380	1.3763	1.4561	<u>1.3680</u>	**1.2011***	12.20%
	50%	MAE	4.5821	1.7526	1.1621	1.1744	1.0468	1.0841	<u>0.9389</u>	**0.8645***	7.92%
		RMSE	4.4642	2.3193	1.4621	1.4666	1.2918	1.4100	<u>1.2582</u>	**1.1579***	7.97%
	80%	MAE	4.2939	1.6973	1.0932	1.1001	1.0319	1.0158	<u>0.8743</u>	**0.8409***	3.82%
		RMSE	4.9603	2.2526	1.3678	1.3763	1.2778	1.3130	<u>1.1651</u>	**1.1094***	4.78%

For MAE and RMSE metrics, the lower value denotes better performance. The best results are in boldface and the second best results are in underline. * indicates 0.05 level, paired t-test of MACDR (Ours) vs. the best baselines. Improve denotes relative improvement over the best baseline.

- Compared with CMF, mapping-based CDR methods have better performances in most tasks. This demonstrates that pre-training & mapping is an effective schema than embedding sharing, either one-to-all mapping (SSCDR, EMCDR, LACDR) or one-to-one mapping (PTUPCDR, HCCDR).
- On mapping-based methods, one-to-one mapping methods achieve better performances compared with one-to-all in most cases. This shows that the common preference transfer is not a good choice for cross-domain recommendation. However, single one-to-one mapping methods do not always present better performances, such as PTUPCDR worse than LACDR under some settings. The above observations indicate that a single common or personalized mapping is insufficient to capture the pattern of user preference transfer.
- Although SSCDR and LACDR are proposed to mitigate the sparse supervision issue by incorporating information of non-overlapping users or items to model training, extensive experiments show that these methods do not always perform better than EMCDR. The probable reason is that SSCDR considers user and interaction items to be input to the same mapping function, whereas the items in the two domains do not overlap, making it difficult to guarantee the quality of the mapped item representation. LACDR utilizes all users in the source domain indiscriminately, which may result in the transfer of useless information and sub-optimal performances.
- Our proposed *MACDR* significantly outperforms all baselines in all experimental settings, which strongly

demonstrates the effectiveness of the proposed *MACDR* for cross-domain cold-start recommendation. The reasons are twofold: Compared with single one-to-all or one-to-one mapping-based methods, *MACDR* designs a PMOE mapping module to flexibly combine the transfer patterns of user common and personalized preference. Besides, *MACDR* proposed an unsupervised PDA optimization strategy to fully exploit non-overlapping users to boost performances. In addition to significant performance improvement, we observe that *MACDR* has more performance gain under sparser overlapping scenarios. It strongly verifies that our proposed PDA optimization strategy can effectively alleviate the sparse supervision issue.

C. Ablation Study

To investigate the effectiveness of each component of our model components, we conduct three variants of *MACDR*. MACDR-w/o PDA denotes that remove the Preference Distribution Alignment optimization of *MACDR* and only retain the Prototype enhanced Mixture-Of-Experts Mapping (PMOE) Module. MACDR-w/o PMOE denotes that only use linear mapping function which maintains the same input as the PMOE module and Preference Distribution Alignment optimization. MACDR-w/o PMOE+PDA denotes that only use linear mapping function which maintains the same input as PMOE module and rating prediction loss. Due to space constraints, we report the results for the three division cases in the Sports-Clothing dataset from Table V and ablation study of *MACDR* on the

TABLE V
ABLATION STUDY OF *MACDR* ON THE SPORTS-CLOTHING DATASET(TASK1)

Models	$\beta = 20\%$		$\beta = 50\%$		$\beta = 80\%$	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
MACDR	**1.1548(-)**	**1.5303(-)**	**0.9281(-)**	**1.2248(-)**	**0.8760(-)**	**1.1768(-)**
MACDR-w/o PDA	1.1675(-1.10%)	1.5486(-1.20%)	0.9341(-0.65%)	1.2333(-0.69%)	0.8846(-0.98%)	1.1846(-0.66%)
MACDR-w/o PMOE	1.1817(-2.33%)	1.5550(-1.61%)	0.9484(-2.19%)	1.2405(-1.28%)	0.8937(-2.02%)	1.1837(-0.59%)
MACDR-w/o PMOE+PDA	1.2781(-10.68%)	1.6645(-8.77%)	1.0190(-9.79%)	1.3315(-8.71%)	0.9492(-8.36%)	1.2577(-6.87%)

"MACDR-w/o PDA" denotes removing the PDA optimization of MACDR. "MACDR-w/o PMOE" denotes using a linear mapping function instead of PMOE for MACDR. "MACDR-w/o PMOE+PDA" denotes removing PDA and PMOE modules.

TABLE VI
ABLATION STUDY OF MACDR ON THE THREE TASKS($\beta = 20\%$)

Models	Task 1		Task 2		Task 3	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
MACDR	**1.1548(-)**	**1.5303(-)**	**0.9597(-)**	**1.3065(-)**	**0.8997(-)**	**1.2011(-)**
MACDR-w/o PDA	1.1675(-1.10%)	1.5486(-1.20%)	0.9682(-0.89%)	1.3251(-1.42%)	0.9088(-1.01%)	1.2205(-1.62%)
MACDR-w/o PMOE	1.1817(-2.33%)	1.5550(-1.61%)	0.9813(-2.25%)	1.3267(-1.55%)	0.9175(-1.98%)	1.2212(-1.67%)
MACDR-w/o PMOE+PDA	1.2781(-10.68%)	1.6645(-8.77%)	1.0424(-7.93%)	1.4029(-7.38%)	0.9635(-7.09%)	1.2756(-6.20%)

three tasks($\beta = 20\%$) from Table VI. From Table V, we observe that MACDR-w/o PDA consistently improves MACDR-w/o PMOE+PDA under the three modes of data division, This verifies that the proposed PMOE mapping module can better combine the advantages of the common and personalized mapping functions to transfer preference knowledge. Besides, MACDR-w/o PMOE also shows better performances than MACDR-w/o PMOE+PDA, demonstrating the effectiveness of the unsupervised Preference Distribution Alignment optimization. Further, we find that the smaller the number of overlapping users in the training set, the greater the boosting of the PDA optimization strategy. This validates our view that users of the same prototype in the source domain are also closer after mapping. PDA utilizes non-overlapping users more rationally in the learning of the mapping function, which complements the sparse supervised signals of the mapping function and mitigates the difficulty that the mapping function heavily relies on the number of overlapping users. Finally, *MACDR* consistently outperforms both variants, proving the effectiveness of combining the two components.

From Table VI, we find consistent conclusions with Table V that our proposed components are effective on all datasets, with greater relative improvement for datasets with fewer overlapping users.

Based on the above analysis, we can conclude that the prototype enhanced mixture-of-experts mapping module can provide a better mapping function than a single common or personalized function, and unsupervised preference distribution alignment optimization makes better use of non-overlapping users in the source domain to facilitate the learning of the mapping function. All of our proposed modules are beneficial to cross-domain cold-start recommendation.

D. Generality of PDA Optimization Strategy

Our proposed PDA optimization strategy is an unsupervised optimization objective, that can easily coupled with other cross-domain cold-start baselines. Here, we conduct experiments to
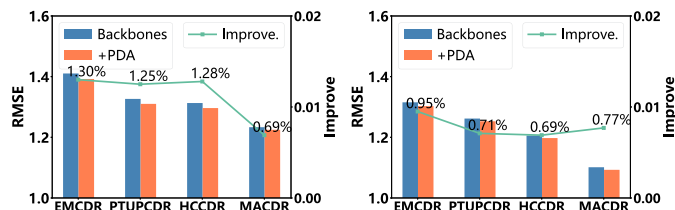

Fig. 5.   Generality of PDA combined with various backbones.

exploit the generality of the proposed PDA with various backbones. As shown in Fig. 5, we compare four SOTA baselines and their joint with PDA. Specifically, we select one-to-one mapping method (EMCDR), one-to-one mapping methods (PTUPCDR, HCCDR), and our proposed PMOE mapping method as backbones. Due to the space limit, we only report comparisons on two datasets. From this figure, we can find that all CDR backbones achieve performance improvements when joined with the PDA optimization strategy. This indicates that our proposed PDA optimization strategy is general to other cross-domain cold-start models, not just our proposed PMOE mapping method. Besides, we find that PDA contributes most in the worst backbone, i.e., over 1.3% RMSE improvement on the EMCDR backbone. Overall, PDA is a general component, that is flexible and easily equipped to current mapping-based methods and boosts their performances significantly.

E. Parameter Sensitivity

In this part, we investigate the impact of hyper-parameters in *MACDR*. We first analyze the number of experts X, which plays an important role in the PMOE mapping module. Next, we study the impact of clustering prototype numbers K. Besides, we explore the effect of the number of negative samples N_s for discriminator training and the effect of loss weights α, Finally, we explore the effect of different gate network inputs on model performance.
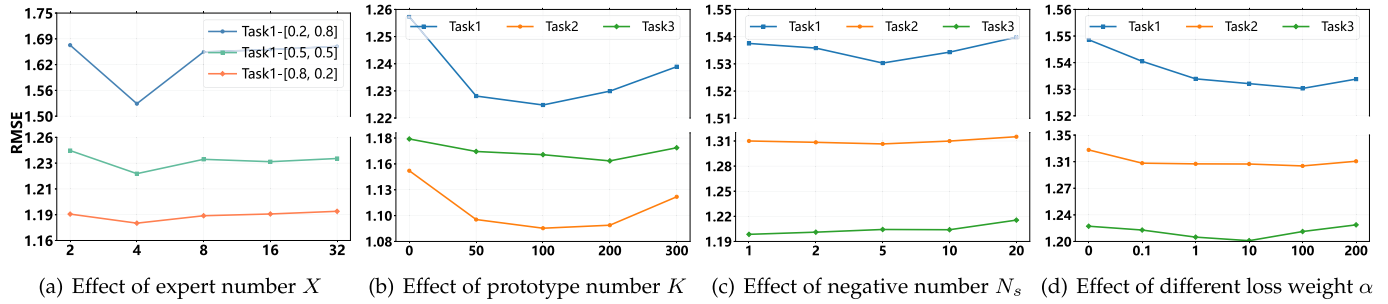
Fig. 6. Performance comparisons under different experts number X, prototype number K, negative number N_s and loss weight α.

(a) Effect of expert number X (b) Effect of prototype number K (c) Effect of negative number N_s (d) Effect of different loss weight α

A - Effect of Experts Numbers X. To exploit the effect of different expert numbers, we search the parameter X in the range of $\{2, 4, 8, 16, 32\}$. Due to space constraints, we only compare experimental results of different expert numbers on task1 As shown in Fig. 6(a), and other datasets with similar findings. We observe that the best experts number is 4 in task1. That means a small number of expert settings can achieve optimal performances, and the size of the number of experts does not have a particularly large impact on the final results, which we attribute to the fact that the number of overlapping users training the mapping function is sparse and does not require a complex mapping network. Meanwhile, we add an L2 regularity term for the model parameters so that they do not vary significantly even with a larger number of model parameters. However, we find that the performance is sensitive to the number of experts when the data is divided in a 2:8 ratio for task1. We argue the possible reason is that the number of overlapping users in the training set is limited, which can easily cause overfitting of the model.

B - Effect of Prototype Numbers K. To investigate the effect of prototype numbers, we set the prototype numbers from zero to three hundred. Since we only cluster for all users in the source domain one time, we don't need to for parametric analysis of all possible ways of dividing data for the three tasks, so we focus on determining the optimal number of clustering centers for one specific way of dividing the data for all three tasks. We illustrate the experimental results in Fig. 6(b). Please note that when $K = 0$, *MACDR* degenerates to MACDR-w/o PDA and the PMOE input without the user prototype embeddings. Thus, It has the worst results. From this figure, We find that the optimal number of prototypes varies from task to task, and the optimal number of prototypes is larger for tasks with a larger number of users in the source domain such as the optimal number of prototypes is 100in task1, and 200in task2 and task3.

C - Effect of Negative Number N_s for Discriminator. As can be observed from Fig. 6(c), we conduct experiments to analyze the impact of the number of negative samples for the discriminator training. We can find that the optimal values of N_s are different for three tasks, and we summarize the reasons as follows. First, tasks with fewer overlapping users tend to sample slightly larger optimal values than tasks with more overlapping users. For example, the optimal value for task1 is 5, and the optimal values for task2 and task3 are 1 or 2. Second, smaller N_s samples can already achieve the optimal performances, and

TABLE VII
PERFORMANCE OF GATE NETWORK INPUTS ON TASK 1

Variant	User	Item Sequence	Prototype	MAE	RMSE
PMOE			✓	**0.9281**	**1.2248**
PMOE$_a$		✓		0.9421	1.2391
PMOE$_b$	✓		✓	0.9405	1.2379
PMOE$_c$		✓	✓	<u>0.9286</u>	<u>1.2250</u>
PMOE$_d$	✓	✓	✓	0.9340	1.2291

too many N_s can lead to performance degradation, which we argue may be since too many non-prototypical samples may lead to some relevant prototype distributions tending to be different, resulting in suboptimal performances.

D - Effect of Loss Weights α. As illustrated in Fig. 6(d), we carefully tune the loss weights α on the three tasks. We observe that *MACDR* achieves the best performances when $\alpha = 100$ on the task1, $\alpha = 100$ on the task2 and $\alpha = 10$ on the task3. Besides, when $\alpha = 0$, the PDA optimization strategy is not used and the model has the worst result. The performance increases first and then drops quickly while α increases. It indicates that proper unsupervised loss weights can make reasonable use of a large number of non-overlapping users and effectively improve the problem of sparse supervision of the mapping function, but a too-strong unsupervised loss will lead to model optimization neglecting the rating prediction loss.

E - Input of Gate Network. We chose to test the effect of PMOE module variants with multiple types of gate network inputs on the final performance of Task 1 as shown in Table VII. We observe that PMOE$_d$ performs much lower than PMOE$_c$, which indicates that the inputs to the gate network do not require pre-trained users and item sequence representations in the source domain. PMOE$_c$ and PMOE have competitive results and are far better than other variants, which indicates that good results are achieved as long as the gate network contains the input features of the user prototype representations, whereas, without the prototype inputs, the performance drops drastically. Similar conclusions are found for other tasks as well as for divided dataset ratios. This validates our hypothesis that users with the same prototypes have the same gate logits, so that the mapping function can capture the transfer patterns of both common and personalized user preferences, achieving the best performance.
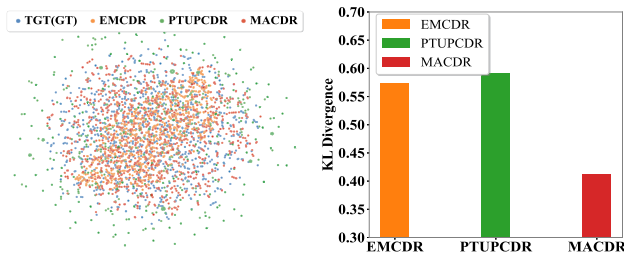
Fig. 7. T-SNE visualization and KL divergence values of 1000 randomly sampled user embeddings in target-domain feature space.

F. Visualization

In this section, we analyze the user embeddings generated by our model *MACDR* and two common cross-domain cold-start recommendation models. We reveal the relationship between ground-truth embeddings in the target domain and the transfer embeddings obtained through various mapping functions to help us better understand the advantages of *MACDR*.

We use t-SNE [10] implemented in scikit-learn package to visualize 1000 randomly sampled test users embeddings in task1 with $\beta = 0.2$. To obtain the best user latent factors as ground truth, we add the test users into the train set and train an additional TGT model (i.e, ground truth). Meanwhile, we also calculate the KL divergence values between the various baseline methods and the target domain ground-truth representations shown in Fig. 7, where yellow and green color denote EMCDR and PTUPCDR models, respectively, red color denote our model, and blue color denotes ground-truth. Ideally, the distributions of the transformed embeddings are the same as the target embeddings. From Fig. 7, we can observe that the distribution of EMCDR representations is relatively centralized, but does not fit the true distribution of users very well, which may be because a single and shared mapping function between the source and target domains cannot learn the relatively complex relationship between the source and target domains well. PTUPCDR employs a meta-network to allow different users to have personalized transfer functions, which alleviates the situation that the common mapping function cannot fit the relationship well. Since the mapping function is personalized for different users, it is difficult to capture the correlation across users, resulting in a diffuse distribution of representations and sub-optimal performance. Compared with other methods, TGT and *MACDR* have the smallest t-SNE visualization representation distance and KL divergence value, which implies that our model has the best result in transferring the knowledge of the target domain. The most important reason why *MACDR* achieves competitive results is that our model combines the commonalities and characteristics between domains and designs an unsupervised loss to exploit user prototypes to constrain similar users to have similar behaviors, which gives more supervised signals to the mapping function.

VI. CONCLUSION

In this paper, we propose a novel *MACDR* model for cross-domain cold-start recommendation. Specifically, we argue that current mapping functions are sub-optimal with strict assumptions, and propose a novel prototype-enhanced MOE (PMOE) module to combine the transfer patterns of common and personalized preferences. Besides, to alleviate the sparse supervision from the limited overlapping users, we propose an unsupervised preference distribution alignment (PDA) optimization based on equivariant learning. The core idea is that preference-similar users share a closer representation distribution after mapping. To this end, the proposed PDA optimization strategy can involve non-overlapping users in model training, which effectively alleviates the data sparsity issue. We conduct extensive experiments on three real-world datasets to demonstrate the effectiveness of the proposed *MACDR*.

Although our framework emphasizes the importance of incorporating non-overlapping users for mapping function learning, it still fails to get rid of the reliance on overlapping users. Additionally, extending the mapping function-based learning paradigm to multiple domains poses challenges. In the future, we consider eliminating the limitation of overlapping users from technical dependencies and exploring cross-domain cold-start recommendations in scenarios without overlapping users. Meanwhile, we would like to broaden the application scenarios of our approach to adapt multiple domains, not just dual domains.

REFERENCES

[1] G. Adomavicius and T. Alexander, "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 6, pp. 734–749, Jun. 2005.

[2] H. Bai et al., "GoRec: A generative cold-start recommendation framework," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 1004–1012.

[3] H. Bai et al., "Multimodality invariant learning for multimedia-based new item recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2024, pp. 677–686.

[4] M. Cai et al., "Popularity-aware alignment and contrast for mitigating popularity bias," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2024, pp. 187–198.

[5] M. Cai et al., "Mitigating recommendation biases via group-alignment and global-uniformity in representation learning," *ACM Trans. Intell. Syst. Technol.*, vol. 15, 2024, Art. no. 101.

[6] J. Cao, S. Li, B. Yu, X. Guo, T. Liu, and B. Wang, "Towards universal cross-domain recommendation," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2023, pp. 78–86.

[7] J. Bi, L. Song, M. Yao, Z. Wu, J. Wang, and J. Xiao, "DCDIR: A deep cross-domain recommendation system for cold start users in insurance domain," *SIGIR*, 2020.

[8] L. Chen et al., "Improving recommendation fairness via data augmentation," in *Proc. Int. Conf. World Wide Web*, 2023, pp. 1012–1020.

[9] Y. Chen, Z. Liu, J. Li, J. McAuley, and C. Xiong, "Intent contrastive learning for sequential recommendation, in *Proc. Int. Conf. World Wide Web*, 2022, pp. 2172–2182.

[10] V. der Maaten, Laurens, and H. Geoffrey, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[11] X. Du, X. Wang, X. He, Z. Li, J. Tang, and C. Tat-Seng, "How to learn item representation for cold-start multimedia recommendation?," in *Proc. ACM Int. Conf. Multimedia*, 2020.

[12] Y. Du, X. Zhu, L. Chen, Z. Fang, and Y. Gao, "MetaKG: Meta-learning on knowledge graph for cold-start recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 10, pp. 9850–9863, Oct. 2023.

[13] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 1126–1135.

[14] C. Gao et al., "Bursting filter bubbles by counterfactual interactive recommender system," *ACM Trans. Intell. Syst. Technol.*, vol. 42, no. 1, pp. 1–27, 2023.

[15] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. Int. Conf. World Wide Web*, 2016, pp. 507–517.

[16] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "LightGCN: Simplifying and powering graph convolution network for recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 639–648.

[17] Z. He et al., "Double correction framework for denoising recommendation," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2024, pp. 1062–1072.

[18] S. Kang, J. Hwang, D. Lee, and H. Yu, "Semi-supervised learning for cross-domain recommendation to cold-start users," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2019, pp. 1563–1572.

[19] P. D. Kingma and L. J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[20] Y. Koren, B. Robert, and V. Chris, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, pp. 30–37, 2009.

[21] B. Li, Q. Yang, and X. Xue, "Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction," in *Proc. Int. Joint Conf. Artif. Intell.*, 2009, pp. 2052–2057.

[22] P. Li and A. Tuzhilin, "DDTCDR: Deep dual transfer cross domain recommendation," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2020, pp. 331–339.

[23] D. Lian, Y. Wu, Y. Ge, X. Xie, and E. Chen, "Geography-aware sequential location recommendation," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 2009–2019.

[24] Z. Lin, C. Tian, Y. Hou, and W. X. Zhao, "Improving graph collaborative filtering with neighborhood-enriched contrastive learning," in *Proc. Int. Conf. World Wide Web*, 2022, pp. 2320–2329.

[25] T. Man, H. Shen, X. Jin, and X. Cheng, "Cross-domain recommendation: An embedding and mapping approach," in *Proc. Int. Joint Conf. Artif. Intell.*, 2017, pp. 2464–2470.

[26] N. Mirbakhsh and C. X. Ling, "Improving top-N recommendation for cold-start users via cross-domain information," *ACM Trans. Knowl. Discov. Data*, vol. 9, 2015, Art. no. 33.

[27] T. Qian, Y. Liang, Q. Li, and H. Xiong, "Attribute graph neural networks for strict cold start recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 8, pp. 3597–3610, Aug. 2022.

[28] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2008, pp. 650–658.

[29] H. Steck, "Evaluation of recommendations: Rating-prediction and ranking," in *Proc. 7th ACM Conf. Recommender Syst.*, 2013, pp. 213–220.

[30] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Representations*, 2018.

[31] M. Volkovs, G. Yu, and T. Poutanen, "DropoutNet: Addressing cold start in recommender systems," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4957–4966.

[32] H. Wang et al., "MCNE: An end-to-end framework for learning multiple conditional network representations of social network," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2019, pp. 1064–1072.

[33] L. Wang et al., "Preference-adaptive meta-learning for cold-start recommendation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 1607–1614.

[34] T. Wang, F. Zhuang, Z. Zhang, D. Wang, J. Zhou, and Q. He, "Low-dimensional alignment for cross-domain recommendation," in *Proc. ACM Int. Conf. Inf. Knowl. Manage.*, 2021, pp. 1064–1072.

[35] X. Wang, X. He, M. Wang, F. Feng, and T.-S. Chua, "Neural graph collaborative filtering," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2019, pp. 165–174.

[36] Y. Wei et al., "Contrastive learning for cold-start recommendation," in *Proc. ACM Int. Conf. Multimedia*, 2021, pp. 5382–5390.

[37] C. Wu, X. Wang, D. Lian, X. Xie, and E. Chen, "A causality inspired framework for model interpretation," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2023, pp. 2731–2741.

[38] L. Wu, X. He, X. Wang, K. Zhang, and M. Wang, "A survey on accuracy-oriented neural recommendation: From collaborative filtering to information-rich recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 5, pp. 4425–4445, May 2023.

[39] L. Wu, Y. Yang, L. Chen, D. Lian, R. Hong, and M. Wang, "Learning to transfer graph embeddings for inductive graph based recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1211–1220.

[40] L. Wu, Y. Yang, K. Zhang, R. Hong, Y. Fu, and M. Wang, "Joint item recommendation and attribute inference: An adaptive graph convolutional network approach," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 679–688.

[41] J. Xu, X. Wang, H. Zhang, and P. Lv, "Heterogeneous and clustering-enhanced personalized preference transfer for cross-domain recommendation," *Inf. Fusion*, vol. 99, 2023, Art. no. 101892.

[42] Y. Yang, L. Wu, R. Hong, K. Zhang, and M. Wang, "Enhanced graph learning for collaborative filtering via mutual information maximization," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 71–80.

[43] Y. Yang, L. Wu, Z. Wang, Z. He, R. Hong, and M. Wang, "Graph bottlenecked social recommendation," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2024, pp. 3853–3862.

[44] Y. Yang et al., "Generative-contrastive graph learning for recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2023, pp. 1117–1126.

[45] F. Yuan, L. Yao, and B. Benatallah, "DARec: Deep domain adaptation for cross-domain recommendation via transferring rating patterns," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 4227–4233.

[46] T. Zang, Y. Zhu, H. Liu, R. Zhang, and J. Yu, "A survey on cross-domain recommendation: Taxonomies, methods, and future directions," *ACM Trans. Inf. Syst.*, vol. 41, 2022, Art. no. 42.

[47] C. Zhao, C. Li, R. Xiao, H. Deng, and A. Sun, "CATN: Cross-domain recommendation for cold-start users via aspect transfer network," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 229–238.

[48] F. Zhu, Y. Wang, C. Chen, G. Liu, M. Orgun, and J. Wu, "A deep framework for cross-domain and cross-system recommendations," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 3711–3717.

[49] F. Zhu, Y. Wang, C. Chen, J. Zhou, L. Li, and G. Liu, "Cross-domain recommendation: Challenges, progress, and prospects," in *Proc. Int. Joint Conf. Artif. Intell.*, 2021, pp. 4721–4728.

[50] Y. Zhu et al., "Transfer-meta framework for cross-domain recommendation to cold-start users," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1813–1817.

[51] Y. Zhu et al., "Personalized transfer of user preferences for cross-domain recommendation," in *Proc. ACM Int. Conf. Web Search Data Mining*, 2022, pp. 1507–1515.

[52] Z. Zhu, S. Sefati, P. Saadatpanah, and J. Caverlee, "Contrastive learning for cold-start recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 5382–5390.

Zihan Wang is currently working towards the master's degree with the Hefei University of Technology, China. He has published papers in referred conferences, such as KDD and IJCAI. His research interests are cross-domain recommender systems and graph learning.

Yonghui Yang received the master's degree from the Hefei University of Technology, in 2021. He is currently working toward the PhD degree with the Hefei University of Technology, China. He has published several papers in referred conferences and journals, such as KDD, SIGIR, IJCAI, ACM Multimedia, *IEEE Transactions on Knowledge and Data Engineering*, and *IEEE Transactions on Big Data*. His research interests include graph learning and recommender systems.

Le Wu (Member, IEEE) received the PhD degree from the University of Science and Technology of China (USTC). She is currently a professor with the Hefei University of Technology (HFUT), China. Her general area of research interests are data mining and knowledge engineering, personalized recommendation, trustworthy user modeling and applications. She has published more than 70 papers in leading journals and conferences, such as *IEEE Transactions on Knowledge and Data Engineering*, *ACM Transactions on Information Systems*, WWW, SIGIR, KDD, NeurIPS and so on. She is an associate editor of *IEEE Trans. on Big Data*, *AI Open and Frontieres of Computer Science*.

Richang Hong (Senior Member, IEEE) received the PhD degree from the University of Science and Technology of China, in 2008. He is currently a professor with the Hefei University of Technology. He has published more than 100 publications in the areas of his research interests, which include multimedia question answering, video content analysis, and pattern recognition. He is a member of the Association for Computing Machinery. He was a recipient of the Best Paper award in the ACM Multimedia 2010.

Meng Wang (Fellow, IEEE) received the BE and PhD degree in the Special Class for the Gifted Young and the Department of Electronic Engineering and Information Science from the University of Science and Technology of China (USTC), Hefei, China, in 2003 and 2008, respectively. is a professor with the Hefei University of Technology, China. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He has authored more than 200 book chapters, journal and conference papers in these areas. He is the recipient of the ACM SIGMM Rising Star Award 2014. He is an associate editor of *IEEE Transactions on Knowledge and Data Engineering (IEEE TKDE)*, *IEEE Transactions on Circuits and Systems for Video Technology (IEEE TCSVT)*, *IEEE Transactions on Multimedia (IEEE TMM)*, and *IEEE Transactions on Neural Networks and Learning Systems (IEEE TNNLS)*.