SAMD: An Industrial Framework for Heterogeneous Multi-Scenario Recommendation

Zhaoxin Huan
Ant Group
Hangzhou, China
zhaoxin.hzx@antgroup.com

Ang Li
Ant Group
Hangzhou, China
liang268038@antgroup.com

Xiaolu Zhang
Ant Group
Hangzhou, China
yueyin.zxl@antgroup.com

Xu Min
Ant Group
Hangzhou, China
minxu.mx@antgroup.com

Jieyu Yang
Ant Group
Hangzhou, China
jieyu.yjy@antgroup.com

Yong He
Ant Group
Hangzhou, China
heyong.h@antgroup.com

Jun Zhou*
Ant Group
Hangzhou, China
jun.zhoujun@antfin.com

ABSTRACT

Industrial recommender systems usually need to serve multiple scenarios at the same time. In practice, there are various heterogeneous scenarios, since users frequently engage in scenarios with varying intentions and the items within each scenario typically belong to diverse categories. Existing works of multi-scenario recommendation mainly focus on modeling homogeneous scenarios which have similar data distributions. They equally transfer knowledge to each scenario without considering the diversity of heterogeneous scenarios. In this paper, we argue that the heterogeneity in multi-scenario recommendations is a key problem that needs to be solved. To this end, we propose an industrial framework named **S**cenario-Aware Model-**A**gnostic **M**eta **D**istillation (SAMD) for the multi-scenario recommendation. SAMD aims to provide scenario-aware and model-agnostic knowledge sharing across heterogeneous scenarios by modeling scenarios' relationship and conducting heterogeneous knowledge distillation. Specifically, SAMD first measures the comprehensive representation of each scenario and then proposes a novel meta distillation paradigm to conduct scenario-aware knowledge sharing. The meta network first establishes the potential scenarios' relationships and generates the strategies of knowledge sharing for each scenario. Then the heterogeneous knowledge distillation utilizes scenario-aware strategies to share knowledge across heterogeneous scenarios through intermediate features distillation without the restriction of the model architecture. In this way, SAMD shares knowledge across heterogeneous scenarios in a scenario-aware and model-agnostic manner, which addresses the problem of heterogeneity. Compared with other state-of-the-art methods, extensive offline experiments, and online A/B testing demonstrate the superior performance of the proposed SAMD framework, especially in heterogeneous scenarios.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Multi-Scenario Recommendation; Meta Learning; Knowledge Distillation

ACM Reference Format:
Zhaoxin Huan, Ang Li, Xiaolu Zhang, Xu Min, Jieyu Yang, Yong He, and Jun Zhou. 2023. SAMD: An Industrial Framework for Heterogeneous Multi-Scenario Recommendation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23), August 6–10, 2023, Long Beach, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3580305.3599955

1 INTRODUCTION

In recent years, recommender systems have been widely deployed to provide users with required items in a timely and effective manner. At large commercial platforms such as Amazon, Alibaba and Alipay, the recommended items generally come from multiple scenarios to satisfy the diversified consumer demands of users. For example, in Alipay which is one of the largest online payment platforms, there are recommendations for multiple scenarios including the homepage, payment successful page, banner, etc. Figure 1 shows three typical recommendation scenarios on Alipay.

Traditional recommender systems recommend items from different scenarios through separate subsystems. Each subsystem collects the historical data from its own scenario and trains a sub-model to serve the online recommendation. However, maintaining multiple sub-models causes a tremendous amount of resource consumption and requires much human cost. Besides, different scenarios may have overlapping user groups and item categories, the sub-model can not leverage the data from all scenarios simultaneously. To

*Corresponding author.

KDD '23, August 6–10, 2023, Long Beach, CA, USA

ACM ISBN 979-8-4007-0103-0/23/08…$15.00
https://doi.org/10.1145/3580305.3599955

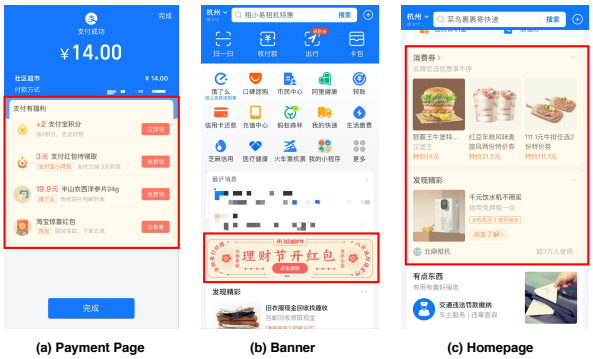

Figure 1: Three typical heterogeneous scenarios in Alipay. The homepage normally recommends discount coupons whereas the banner gives red packets to users. And users tend to actively navigate the homepage, but passively browse the payment page. The data distribution and features are heterogeneous across these scenarios.

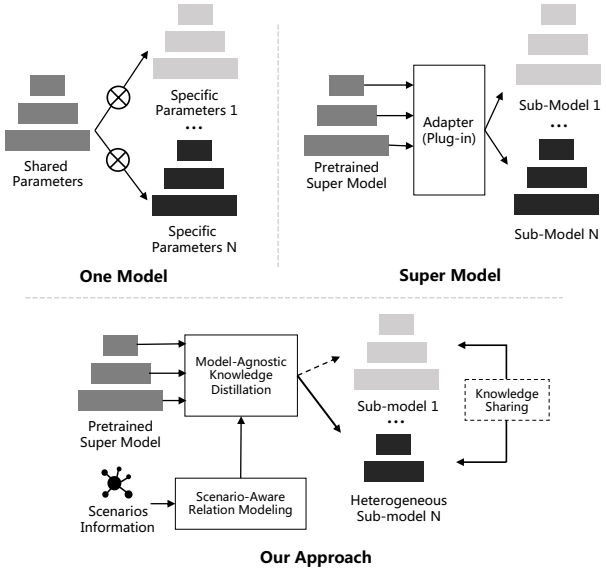

Figure 2: The different methods of multi-scenario recommendation. One Model shares parameters among scenarios and Super Model plugs knowledge into the sub-models. Our approach aims to incorporate the scenario-aware scenarios' relationship into the model-agnostic knowledge sharing.

solve this problem, the multi-scenario recommendation has been proposed to leverage the data from all scenarios and jointly train the sub-models. Based on our research, we broadly divide the existing methods into two categories: One Model and Super Model, as shown in Figure 2. The One Model [2, 12, 16, 17, 26, 32], such as Multi-Task Learning, consists of shared parameters and multiple sets of scenario-specific parameters. The final model of each sub-model is obtained by combining the shared centered parameters and the scenarios-specific parameters. The Super Model [4, 10, 11, 27] conducts the knowledge extraction from a super scenario with long-time sufficient data and further plugs the knowledge into the inner layers of sub-models.

In real industrial settings, the **heterogeneity** is a challenging problem that needs to be solved. The heterogeneity is caused by the deviation of data distribution across scenarios, since users often engage in scenarios with diverse intentions and the items within each scenario typically belong to different categories. In addition, because scenarios with better exposure positions and hot-selling items collect more data compared with other scenarios, the amount of data is commonly unbalanced across scenarios. Therefore, heterogeneity exists widely in the industrial multi-scenario recommendations. For example, as shown in Figure 1, the homepage of Alipay normally recommends discount coupons whereas the banner often gives red packets to users. Furthermore, the homepage tends to attract more user engagement compared to the payment page, as users tend to actively browse the former, but passively navigate the latter. As a result, the feature engineering is normally not aligned across scenarios. The overlap ratio of features between different scenarios is typically below 50%, and in some cases, there is even no intersection at all. And the sub-models have diverse model architectures, which causes the joint modeling of heterogeneous scenarios to become a big challenge in the multi-scenarios recommendation.

We argue that existing methods including One Model and Super Model can hardly handle the problem of heterogeneity, because their knowledge sharing lacks diversity across scenarios, as shown in Figure 2. First, the knowledge sharing needs to be scenario-aware since the scenarios' relationship is more complicated than that in homogeneous scenarios. The One Model equally treats each scenario without diversity. As a result, the minor scenarios with little data might be overwhelmed by directly sharing knowledge with other major scenarios. The Super Model transfers the knowledge in one way, that is from the super model to the sub-model. The knowledge is transferred to sub-scenarios in the same way, which causes the scenarios that are diverse from super scenario to have limited promotion. Second, the knowledge sharing needs to be model-agnostic to flexibly support heterogeneous scenarios with diverse features and model architectures. The One Model assumes that the shared parameters are the same among scenarios and can be easily combined with scenarios-specific parameters in equal dimensions. When dealing with sub-models with different model architectures, One Model needs to manually align the feature schema and model architectures, which inevitably decreases the performance. The Super Model plugs the knowledge of super model into the inner layers of sub-models without the limitation of the same model architecture. But the Super Model has a high online delay because the online serving of sub-models depends on the intermediate output of super model, which is not suitable in low delay required scenarios.

To fully address the above problems, in this paper, we propose an industrial framework named **S**cenario-Aware Model-**A**gnostic **M**eta **D**istillation (SAMD) for the heterogeneous multi-scenario recommendations. Conceptually, as shown in Figure 2, SAMD aims to share scenario-aware and model-agnostic knowledge by modeling

the scenarios' relationship and conducting heterogeneous knowledge distillation. Specifically, as shown in Figure 3, to realize the scenario-aware knowledge sharing, SAMD first measures the scenario representation from both explicit views and implicit views and then models the complicated scenarios' relationship through meta network. The meta network automatically establishes the mapping between scenarios and knowledge sharing, which enables scenarios to have their own specific knowledge. To conduct model-agnostic knowledge sharing, SAMD proposes heterogeneous knowledge distillation to guide the heterogeneous scenarios to share knowledge through intermediate features distillation without the restriction of the model architecture. The combination of the whole framework is the novel meta distillation paradigm, which shares scenario-aware model-agnostic knowledge across heterogeneous scenarios in a diverse and flexible manner.

Since 2022, SAMD is deployed in the advertising system of Alipay, obtaining a 6.48% improvement on CVR and 1.06% on CPM. To sum up, the main contributions of this work include:

- We study an important but under-explored problem of heterogeneity in the multi-scenario recommendation, which is essential to the prosperity of large-scale industrial recommender systems. To the best of our knowledge, we are the first to attempt to propose a framework of multi-scenario recommendation that addresses the problem of heterogeneity.
- We propose SAMD, a novel multi-scenario recommendation framework, sharing scenario-aware model-agnostic knowledge across heterogeneous scenarios by incorporating the scenarios' relationship into the heterogeneous knowledge distillation.
- We evaluate SAMD on both practical and public industrial dataset and deploy it in the display advertising system of Alipay in 2022. The consistent superiority compared with state-of-the-art methods validates the effectiveness of SAMD, especially for heterogeneous scenarios.

2 RELATED WORKS

In this section, we first introduce multi-scenario recommendation which is closely related to our work. And then we give a brief introduction to meta learning and knowledge distillation.

2.1 Multi-Scenario Recommendation

In the context of reproducible industrial deployment, the multi-scenario recommendation can be categorized into two strategies: One Model and Super Model. In One Model, the basic idea is to train a unified model that serves homogeneous scenarios. STAR [16] proposes a star topology to enable knowledge transfer among scenarios through the shared centered parameters. M2M [26] simultaneously predicts multiple tasks in multiple similar advertising scenarios through the meta unit that dynamically generates scenarios' parameters. Similar to M2M, AESM [32] integrates both multi-scenario learning and multi-task learning into a unified framework with automatic structure learning. The expert selection algorithm is proposed to automatically identify scenario-/task-specific and shared experts for each input. Besides, the multi-task learning methods [2, 12, 17] can also be regarded as a special form of One Model [16]. The Super Model typically contains a knowledge extraction stage and a knowledge plugging stage. KEEP [27] extracts the knowledge from the super model and then plugs it into sub-models. The KEEP only transfers from super model to sub-model in the same way, which causes the scenarios that are diverse from super scenario to have the limited promotion. CTNet [11] designs trainable adapters to transfer the knowledge from each layer of the super model to the corresponding layer of the sub-model. But when the sub-model has different model architectures from the super model, the CTNet needs to manually specify the mapping layers.

Different from existing works of the multi-scenario recommendation, the proposed SAMD framework focuses on solving the problem of heterogeneity. The SAMD shares scenario-aware knowledge across heterogeneous scenarios without the restriction of the model architecture.

2.2 Meta Learning

Meta learning, also called learning-to-learn, aims to learn the meta function between tasks and parameters, which can be used in the multi-scenario recommendation by treating each scenario as a task. The gradient-based methods such as MAML [4], MeLU [10] and MetaEmb [30], learn a meta-optimizer that learns globally shared initial parameters among several tasks that can be quickly adapted to new tasks. To eliminate the second derivative gradient in optimization, CMML [3], MWUF [31] and M2M [26] directly establish the mapping connection between tasks and parameters through meta network.

Inspired by existing works of meta learning, the proposed SAMD leverages the scenarios' relationship into knowledge sharing through meta network. SAMD adopts meta network to dynamically generate the scenario-aware transfer strategies. Besides, instead of directly using the task embedding input to meta network, SAMD designs an automatic soft clustering module to enhance the modeling of scenarios' relationship, which can capture the commonalities and differences across scenarios.

2.3 Knowledge Distillation

Knowledge distillation [1, 23–25] is proposed to transfer the knowledge from the well-trained teacher model to the student model. It has been widely used in model compression tasks. According to the difference of the transferred knowledge, the distillation can be classified as logits-based methods [6] and feature-based methods [8, 15, 25]. In logits-based distillation, the knowledge is transferred through the soft probability generated by the teacher model. Compared with the logits-based approaches, the intermediate features of the teacher models contain more knowledge than logits and enable student models to learn richer information. However, the transfer of intermediate features needs to design the transfer weight between each layer in the teacher model and student model. Some methods [7, 19] have been proposed to automatically decide the transfer weights based on the semantic similarity between layers.

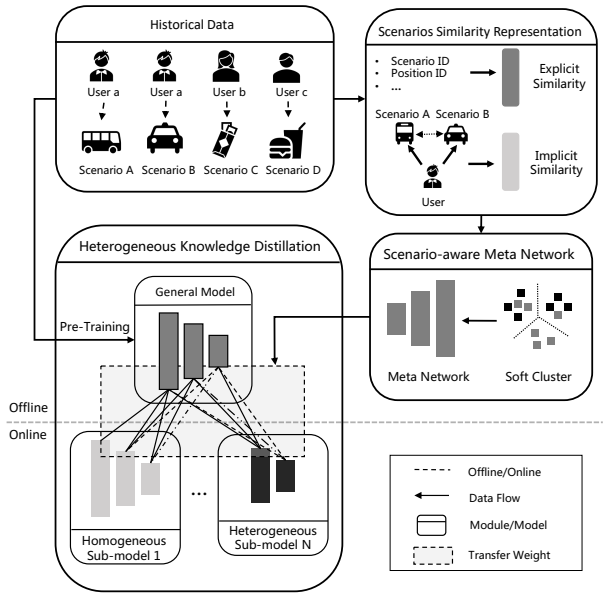

Figure 3: An illustration of the overall architecture of the proposed SAMD framework.

SAMD adopts knowledge distillation to conduct heterogeneous knowledge transfer. Compared with the existing feature-based distillation, SAMD dynamically generates the transfer weights according to the scenarios' relationship rather than the layer-level semantic similarity.

3 FRAMEWORK OVERVIEW

In this section, we will describe the system overview of the proposed SAMD framework, as well as some implementation details of the core components. As shown in Figure 3, SAMD contains several components, including scenario similarity representation, scenario-aware meta network and heterogeneous knowledge distillation.

SAMD first collects long-time feedback data and trains the general model based on the super scenario which contains sufficient users and item categories. At the same time, the scenarios similarity representation module is trained to generate comprehensive representation by considering both scenario-related attributes and user-scenario relationships, which can be regarded as explicit similarity and implicit similarity respectively.

In order to share scenario-aware knowledge across scenarios, the modeling of complicated scenarios' relationship is essential. Therefore, the scenario-aware meta network first models the relationship by soft clustering which assigns each scenario with a multiple cluster probability. Then the meta network takes the scenarios representation and cluster representation as input and establishes the mapping between scenarios and the weights of knowledge sharing which are used in heterogeneous knowledge distillation. In this way, the separated scenarios are connected by the meta network. And the knowledge sharing is scenario-aware since the scenarios with close representation share similar knowledge.

Finally, the heterogeneous knowledge distillation utilizes the scenario-aware transfer weights to conduct model-agnostic knowledge sharing between each layer in the general model and sub-model. The distillation of intermediate features makes full use of information in the general model. Heterogeneous sub-models can also learn from the general model without the manual layer mapping. Besides, the online serving of sub-models only depends on their own specific parameters. The inference delay and model storage are not affected by the knowledge distillation.

Compared with single scenario modeling, SAMD only needs one training process to get multiple sub-models, which reduces maintenance costs. Compared with multi-scenario modeling, SAMD enables knowledge to be shared across homogeneous and heterogeneous scenarios simultaneously, which effectively addresses the problem of heterogeneity in multi-scenario recommendations.

4 ALGORITHMIC DESIGN

In this section, we elaborate the algorithmic design of SAMD, mainly including the knowledge extraction, scenario similarity representation and meta distillation, as well as their connection.

4.1 Knowledge Extraction

In order to obtain the general knowledge which contains diverse user behaviors and item categories, SAMD first trains the general model $\mathcal{F}_g$ based on the long-time data collected from large-scale scenarios $\mathcal{D}_g$. The general model adopts the simple yet effective embedding-based Deep Neural Network (DNN) to support more sub-models. The input of the general model contains both item features and user features. The embedding layer transforms them into low-dimensional dense representations. For each user-item pair (u, i), the general model optimizes the cross-entropy loss between the predicted probability $\hat{y}$ and the binary label y (e.g. click or conversion), as shown in Equation 1.

$$\mathcal{L}_{ce} = \sum_{(u,i)\in\mathcal{D}_g} -y log\ \hat{y} - (1-y) log\ (1-\hat{y}) \tag{1}$$

Note that the general model is not the main concern of SAMD, the pre-training objective is not limited to the above single task and the backbone of the general model can be changed to other powerful recommendation models.

4.2 Scenario Similarity Representation

To model the complicated relationships among scenarios, the modeling of scenario representation is one of the key problems. SAMD proposes to learn a comprehensive scenario representation from both explicit and implicit views. As shown in Figure 3, the explicit representation reflects the inherent properties of the scenario which can be defined as the static attributes of scenarios. Moreover, there exist potential relations among scenarios that are connected by users. This complementary latent relationship of *scenario-user-scenario* can be regarded as an implicit similarity.

4.2.1 Explicit Similarity. Given the scenario information $\mathcal{S}_i$ of i-th scenario with m attributes, the explicit representation (ER) can be defined as:

$$ER_i = \tau(W_{exp}[e_{s_1} \oplus e_{s_2} \oplus \cdots \oplus e_{s_m}]^\top), \quad ER_i \in \mathbb{R}^d \tag{2}$$

where s_m is the m-th attribute and e_{s_m} is the corresponding embedding. W_{exp} is the weight matrix that extracts the explicit representation from the concatenation $\oplus$ of embedding. $\tau(\cdot)$ is the activation function. The explicit similarity between scenario i and j can be measured using the ER_i and ER_j.

4.2.2 Implicit Similarity. To fully exploit the implicit similarity, we construct *scenario2user* relation through swing algorithm [21, 22]. The *scenario2user* relation is proposed to measure the implicit similarity according to the number of co-occurrence users in different scenarios. If scenarios i and j often interacted with the same user, there may exist a latent connection. To alleviate the bias brought by active users, we also utilize the weighting factor of users to relieve the intensity of the links built by active users. Specifically, the similarity between scenario i and j can be defined as:

$$Sim(i,j) = \sum_{u \in U_i \cap U_j} \sum_{v \in U_i \cap U_j} w_u * w_v \frac{1}{\alpha + |\mathcal{S}_u \cap \mathcal{S}_v|} \tag{3}$$

where U_i and U_j are the set of users who have interacted with scenario i and j, $\mathcal{S}_u$ and $\mathcal{S}_v$ are the set of scenarios that were visited by user u and v. $w_u = 1/\sqrt{|S_u|}, w_v = 1/\sqrt{|S_v|}$ represents the penalized factor of user u and v which measures the user activity. α is a smoothing coefficient. By ranking the $Sim(i,j)$, it gives the most similar G scenarios for i. The implicit representation (IR) can be calculated as:

$$IR_i = \tau(W_{imp}[MeanPooling(\{ER\}_G)]), \quad IR_i \in \mathbb{R}^d \tag{4}$$

where $\{ER\}_G$ is the explicit representation of top G similar scenarios and W_{imp} is the weight matrix which extracts the implicit representation. The combination of the IR_i and ER_i formulates the final similarity representation:

$$T_i = \tau(W_{sim}[ER_i \oplus IR_i]), \quad T_i \in \mathbb{R}^d \tag{5}$$

where W_{sim} is the weight matrix which extracts the final scenario representation. The T_i can be offline pre-trained and updated periodically (i.e., weekly in our implementation). The meta distillation (see in Section 4.3.1) will take T_i as input to generate the transfer strategies.

4.3 Scenario-Aware Model-Agnostic Knowledge Sharing

After getting the offline trained general model and scenario representation, the SAMD framework proposes a novel meta distillation paradigm to share scenario-aware model-agnostic knowledge across scenarios. As shown in Figure 3, the meta distillation consists of two components, including scenario-aware meta network and heterogeneous knowledge distillation.

4.3.1 Scenario-aware Meta Network. As shown in Figure 4, the meta network first mines the potential relationship among scenarios using soft clustering based on the scenario representation T_i. Then the meta network establishes the mapping connection between scenarios and transfer strategies that guide the heterogeneous knowledge distillation (see in Section 4.3.2).

Soft Cluster Module. The differences and commonalities coexist across heterogeneous scenarios. For example, the homepage and payment page may recommend the same type of advertisement

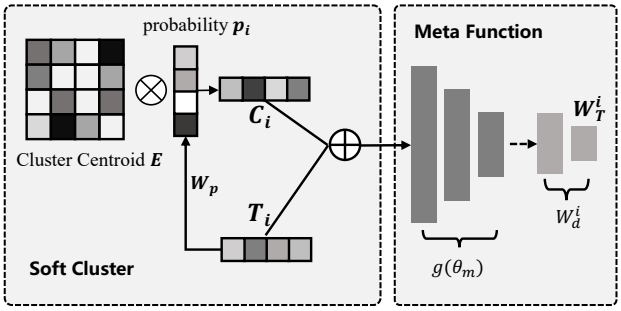

Figure 4: An illustration of scenario-aware meta network. In the soft cluster module, the scenarios are clustered into different clusters based on the offline trained scenarios representation. And then the meta function takes both scenario and cluster embedding as input and generates the scenario-aware strategies of knowledge sharing.

but in different exposure positions. Directly using the scenario representation T_i can not comprehensively express such differences and commonalities. Therefore, the soft cluster module is proposed to automatically discover the potential relationship among scenarios. Specifically, given K clusters, a trainable cluster center vector $E \in \mathbb{R}^{K \times d}$ is initialized, which has the equal second dimension with scenario representation $T_i \in \mathbb{R}^d$. Then a projection matrix $W_p \in \mathbb{R}^{K \times d}$ is leveraged to assign each scenario T_i to the K clusters by:

$$q_i = \tau(W_p T_i), \quad q_i \in \mathbb{R}^K \tag{6}$$

where $\tau(\cdot)$ is an activation function and q_i is the projection output of i-th scenario embedding T_i on the k-th cluster. The probability that the i-th scenario assigned to the k-th cluster p_i^k can be defined as:

$$p_i^k = \frac{e^{q_i^k}}{\sum_{k=1}^{K} e^{q_i^k}} \tag{7}$$

Finally, a weighted sum aggregation function is used to get the cluster representation C_i of i-th scenario, which contains shared characteristics of different clusters, defined as:

$$C_i = \sum_{k=1}^{K} p_i^k \cdot E_k, \quad C_i \in \mathbb{R}^d \tag{8}$$

Through automatic soft clustering operation, each scenario can obtain a cluster enhanced embedding which contains the latent relationship among scenarios.

Meta Function. After getting the scenario embedding T_i and cluster embedding C_i, the meta function takes T_i and C_i as input and generates the corresponding transfer strategies W_T^i, defined as follows:

$$W_T^i = g(T_i \oplus C_i; \theta_m) W_d^i, \tag{9}$$

where $g(\cdot) : \mathbb{R}^{2d} \Rightarrow \mathbb{R}^{M_g * M_g}$ is a fully connected meta network with parameters θ_m. $2d$ is the dimension of $T_i \oplus C_i$. M_g is the number of layers in the general model (see in Section 4.1). M_i is the number of layers in sub-model i. The output of the $g(\cdot)$ is the general transfer strategy among all scenarios. $W_d^i \in \mathbb{R}^{M_g * M_g \times M_g * M_i}$ is the

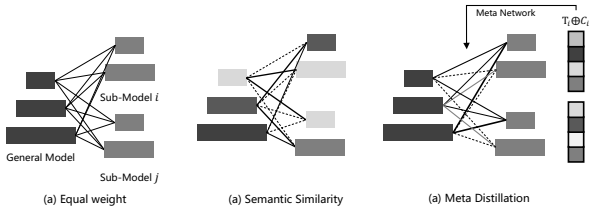

Figure 5: An illustration of different distillation patterns. Darker lines represent higher weights. (a) The all2all assigns each layer with equal weight. (b) The layers with close semantic similarity are assigned high weight. (c) Meta distillation generates weights based on the scenarios' relationship, similar scenarios have similar strategies of knowledge sharing.

plug-in matrix that converts the general transfer strategy to specific strategies for heterogeneous sub-models. $W_T^i \in \mathbb{R}^{M_g * M_i}$ is the transfer strategy for scenario i, which will be used in heterogeneous knowledge distillation to guide the knowledge sharing between the layers in the general model and the layers in sub-model.

In this way, since the final transfer strategies are generated based on scenarios' relationship, the scenarios (i, j) with close $(T_i \oplus C_i, T_j \oplus C_j)$ could share similar (W_T^i, W_T^j). The knowledge is not only transferred from the general model to sub-models but also shared among sub-models, which enhances the diversity of knowledge sharing.

4.3.2 Model-Agnostic Heterogeneous Distillation. To realize the model-agnostic knowledge sharing, the heterogeneous knowledge distillation is proposed to automatically explore the sharing patterns without the restriction of the model architecture. As mentioned in Section 2.3, compared with the distillation of the logits layer, intermediate features of deep neural networks contain more information. Therefore, SAMD adopts feature-based distillation to leverage information from all layers in the general model. The loss function of feature-based distillation can be formulated as:

$$\mathcal{L}_{fd} = \sum_{m=1}^{M_i} \sum_{n=1}^{M_g} w_{mn} \mathcal{L}(\Phi_i(\mathcal{F}_i^m(x)), \Phi_g(\mathcal{F}_g^n(x))) \tag{10}$$

where $\mathcal{F}_i$ is the sub-model of scenario i with M_i layers and $\mathcal{F}_g(x)$ is the general model with M_g layers. m and n denote the layer in $\mathcal{F}_i(x)$ and $\mathcal{F}_g(x)$. w_{mn} denotes the transfer weight from m-th layer of $\mathcal{F}_i(x)$ to n-th layer of $\mathcal{F}_g(x)$ and w_{mn} is subjected to $\sum_{m=1}^{M_i} w_{mn} = 1$. Φ_i, Φ_g are designed to convert $\mathcal{F}_i^m(x)$ and $\mathcal{F}_g^n(x)$ into easy-to-transfer forms [5], such as attention maps [25], FSP matrix [23]. $\mathcal{L}$ is usually L1/L2 norm or Maximum Mean Discrepancy (MMD) Loss.

As illustrated in Figure 5, to leverage information of all layers in general model, w_{mn} can be defined as three formulations. All2all distillation is a simple way that assigns the same weight to each layer, which neglects the information gap between layers. Semantic similarity has been used to generate the transfer weight automatically. The semantically related layers (layers with close color) should be given a high transfer weight (dark line). These distillation patterns ignore the scenarios' relationship and treat each sub-model equally.

Algorithm 1: SAMD.

Input: Multi-scenarios dataset $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_N\}$;
Scenario attributes $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N\}$;
Output: Sub-models for heterogeneous scenarios
$\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N\}$

;
Train the general model $\mathcal{F}_g$ in Equation 1
Obtain the representation of i-th scenarios T_i in Equation 5
while not converged **do**
  for $i \leftarrow 1$ **to** N **do**
    Sample a batch from i-th scenario $\mathcal{D}_i$;
    Get the scenario cluster embedding C_i in Equation 8;
    Generate the transfer strategy W_T^i from meta network $g(\theta_{meta})$ in Equation 9;
    Transfer the knowledge from $\mathcal{F}_g$ to sub-model $\mathcal{F}_i$ based on W_T^i;
    Get the output $\hat{y}$ of $\mathcal{F}_i$; Compute the loss function $\mathcal{L}$ in Equation 11;
    Optimize the parameters: sub-model $\mathcal{F}_i$; meta network $g(\theta_m)$, W_d^i, E, W_p;
  end
end

The proposed heterogeneous distillation utilizes the transfer strategies $W_T^i \in \mathbb{R}^{M_g * M_i}$ learned by the meta network in Equation 9. In this way, the transfer strategy could guide the distillation module to transfer scenario-aware layer-level knowledge from general model to sub-models. In turn, the distillation module would update the meta network to generate a better transfer strategy. The loss function of the whole meta distillation framework can be formulated as Equation 11 and the whole algorithm of SAMD is demonstrated in Algorithm 1.

$$\mathcal{L} = \sum_{i=1}^{N} \sum_{x \in \mathcal{D}_i} \mathcal{L}_{fd}(\mathcal{F}_i(x), \mathcal{F}_g(x)) + \sum_{i=1}^{N} \sum_{x,y \in \mathcal{D}_i} \mathcal{L}_{ce}(\mathcal{F}_i(x), y) \tag{11}$$

5 EXPERIMENTS

In this section, we conduct extensive offline and online experiments on the proposed SAMD framework, to demonstrate its effectiveness and efficiency [1].

5.1 Experiment Settings

5.1.1 Datasets. Since our work mainly focuses on the industrial problem of heterogeneous multi-scenario recommendation, we first employ the real-world industrial dataset, which collects Click-through rate (CTR) data of advertisements from different scenarios on Alipay. Specifically, the dataset contains 25 scenarios with diverse positions of exposure, businesses and item categories. We choose a scenario that covers more users and item categories as the super scenario to train the general model. The rest of the scenarios contain 10 homogeneous scenarios, 8 heterogeneous scenarios,

[1] See more details in https://github.com/yangli268038/SAMD

Table 1: Performance on AliExpress Dataset and Alipay Ad Dataset.

Category	Methods	AliExpress Dataset			Alipay Ad Dataset					
		AUC_{homo}	AUC_{heter}	AUC_{minor}	AUC_{homo}	AUC_{heter}	AUC_{minor}	$GAUC_{homo}$	$GAUC_{heter}$	$GAUC_{minor}$
MTL	MMOE	0.7169	0.6735	0.7019	0.8614	0.7272	0.7267	0.7343	0.6835	0.6732
	PLE	0.7180	0.6708	0.7056	0.8607	0.7300	0.7287	0.7351	0.6834	0.6741
	MSSM	0.7185	0.6668	0.7045	0.8598	0.7275	0.7259	0.7332	0.6829	0.6730
Meta Learning	MAML	0.7181	0.6772	0.7031	0.8606	0.7271	0.7297	0.7312	0.6821	0.6747
	MeLU	0.7164	0.6782	0.7035	0.8605	0.7297	0.7303	0.7310	0.6842	0.6744
	s^2Meta	0.7195	0.6786	0.7063	0.8600	0.7303	0.7340	0.7331	0.6814	0.6738
One Model	STAR	0.7209	0.6706	0.7034	0.8616	0.7287	0.7260	0.7347	0.6822	0.6740
	M2M	0.7214	0.6716	0.7030	0.8644	0.7248	0.7260	0.7351	0.6837	0.6739
	AESM	0.7211	0.6705	0.7043	0.8620	0.7300	0.7200	0.7350	0.6828	0.6737
Super Model	PF	0.7186	0.6751	0.7032	0.8574	0.7230	0.7258	0.7314	0.6810	0.6722
	KEEP	0.7193	0.6762	0.7056	0.8627	0.7341	0.7273	0.7342	0.6851	0.6734
	CtNet	0.7244	0.6815	0.7059	0.8635	0.7350	0.7324	0.7441	0.6847	0.6735
Our Approach	SAMD	**0.7257**	**0.6879**	**0.7123**	**0.8667**	**0.7459**	**0.7393**	**0.7461**	**0.6863**	**0.6764**

and 6 minor scenarios. The homogeneous scenarios share similar data distribution, while the heterogeneous scenario differs from the super scenario in structures and features. The minor scenarios are homogeneous and have a 1/10 amount of data compared with other scenarios.

Moreover, we also conduct experiments on a public industrial dataset AliExpress to demonstrate the effectiveness of our framework. The AliExpress dataset includes CTR settings and contains five scenarios split by user nationality, including NL, FR, ES, RU, and US. Note that the public datasets commonly include fewer scenarios than the industrial dataset. On the AliExpress dataset, we train the general model based on the RU scenario and artificially construct two heterogeneous scenarios (NL, ES). Specifically, we randomly mask some features and design sub-models with different structures. The FR scenario is considered a minor scenario since the small amount of data. Because of data desensitization, the semantic meaning of each feature is hardly defined in AliExpress dataset. The explicit similarity (see in Section 4.2.1) can not be calculated. Therefore, we only use the implicit similarity (see in Section 4.2.2) to generate the scenario representation. Specifically, We use *mu*3 (col17-col26) as user features to build *scenario2user* relation that generates the implicit similarity.

5.1.2 Compared Methods. State-of-the-art methods are employed in the experiments. They fall into four categories: Multi-Task Learning, One Model, Super Model, and Meta learning (the definition of One Model and Super Model is in Figure 2).

- **Multi-task learning (MTL)** We choose shared bottom based multi-task methods as baseline models, including MMOE [13], PLE [18] and MSSM [2]. To deal with heterogeneous scenarios, MTL shares the homogeneous features in one embedding table and builds the scenario-specific embedding tables for scenario-specific features. Misaligned features are filled with zeros, and the tower layers have different numbers of layers.
- **Meta learning** To compare the performance of minor scenarios, we also compare our framework with meta learning methods, including MAML [4], MeLU [10], and s^2Meta [28]. We treat each scenario as a task and split the corresponding support set and query set. The meta-test is conducted on all scenarios.
- **One Model** We choose STAR [16], M2M [26] and AESM [32] as baselines. As mentioned in Section 1, we manually align the schema of features and model architectures on heterogeneous scenarios.
- **Super model** The Pretrain-Fintune (PF) paradigms such as KEEP [27] and CtNet [11] are regarded as baseline models for Super Model. For KEEP, we plug the knowledge of the general model into the output layer of sub-models. For CtNet, we manually build the one-to-one mapping adapters in the last three layers of sub-models.

5.1.3 Implementation Details. The embedding sizes of user and scenario features are set to 16 for all models. The backbone of the general model is DNN with [256, 128, 64, 16] hidden units. The meta network has a three-layer MLP structure, with [64, 32, 16] hidden units in each layer. On the AliExpress dataset, for the convenience of the experiment, the backbones of heterogeneous sub-models are manually designed with [64, 8] hidden units, and the features are randomly masked to construct the heterogeneity of the features. On Alipay Ad dataset, according to the practical application, heterogeneous scenarios contain diverse model architectures, including DNN, DCN [20], MMOE, etc. And the features are spontaneously not aligned across scenarios. The minor scenarios are selected according to the amount of data on the homogeneous scenarios. For all the neural network-based methods, the learning optimizer is Adam [9], with a learning rate of 1e-4, and the batch size is set to 512. For hyperparameters, the number of similar scenarios for G in Equation 4 is set to 2 in AliExpress dataset and 10 in Alipay Ad dataset. The number of clusters in meta networks is set to 3 in AliExpress dataset and 15 in Alipay Ad dataset after tuning the hyperparameters.

5.1.4 Evaluation Metrics. The area under the ROC curve (AUC) is the common metric used to evaluate the performance of CTR prediction. Thus, we employ the AUC in our experiments. Moreover, we also report the user-weighted GAUC [29] on Alipay Ad dataset by averaging AUC over users to measure the ability to rank users internally. To compare the performance of different kinds of scenarios in the multi-scenario recommendation, we also report

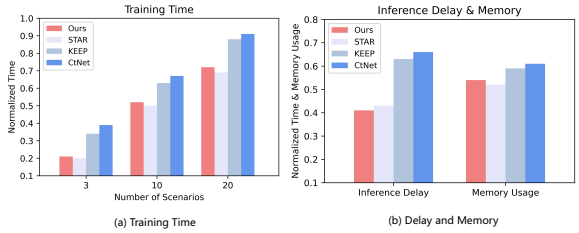

Figure 6: The comparison of deployment efficiency. (a) The comparison of training time. The horizontal axis represents the number of scenarios and the vertical axis is normalized time. (b) The comparison of inference delay and memory usage.

the averaged AUC and GAUC across scenarios of different types, including homogeneous, heterogeneous, and minor scenarios.

5.2 Model Performance

5.2.1 Offline Evaluation. We evaluate all baseline models and proposed SAMD framework on the AliExpress dataset and Alipay Ad dataset. To give a convincing comparison, we conducted five experiments with different random seeds and reported the average results. As illustrated in Table 1, the consistent improvement over all kinds of scenarios validates the efficiency of SAMD. We also added t-test, which indicates the statistical significance of SAMD for $p < 0.01$ compared with the best baseline. Specifically, the SAMD reaches competitive results in homogeneous scenarios compared with other baseline models, which demonstrates that homogeneous scenarios can also benefit from scenario-aware knowledge sharing. In heterogeneous scenarios, since the SAMD conducts model-agnostic knowledge sharing by automatically leveraging the information of each layer in the general model, the proposed SAMD significantly outperforms the Super Models. The One Model and MTL methods suffer from the inevitable performance reduction in heterogeneous scenarios due to the manual alignment of features and model architectures. In minor scenarios, SAMD significantly outperforms baseline methods because SAMD models the scenario-aware scenarios' relationship, minor scenarios can share knowledge with similar major scenarios.

5.2.2 Online Evaluation. We deployed the SAMD method in Alipay's advertising system and conducted an A/B test over two weeks. Due to industrial constraints, comparing all baseline models in the online recommendation system is difficult. Therefore, we choose KEEP as the baseline model. The online experiment shows that the SAMD increased the Conversion Rate (CVR) and Cost Per Mile (CPM) by 6.48% and 1.06%, respectively. This result demonstrates the practical effectiveness of the SAMD framework in industrial multi-scenario recommendations.

5.2.3 Deployment Efficiency. To demonstrate the effectiveness of SAMD in practical deployment, SAMD is compared with other industrial frameworks in terms of training time, inference delay and memory usage. As shown in Figure 6, in comparison to Keep and CtNet, SAMD exhibits an advantage in training time due to its ability to train multiple sub-models simultaneously. The general model and scenarios representation are trained offline and can be updated weekly or monthly, which has little impact on the overall training time. Although the meta distillation module introduces more parameters, these parameters can be calculated through matrix manipulation and updated by gradient descent. The memory usage of the SAMD is not significantly increased compared to other models. For online serving, SAMD no longer relies on the meta distillation and general model and can produce predictions based on the parameters of the sub-model. Compared to KEEP and CtNet which need the intermediate features of the general model for inference, SAMD's inference delay is significantly lower while remaining almost the same as that of One Model methods. In conclusion, combining deployment efficiency and offline/online evaluation, SAMD has a superior performance in the industrial multi-scenario recommendation.

Table 2: Ablation studies on AliExpress Dataset and Alipay Ad Dataset.

Ablation Methods	AliExpress Dataset			Alipay Ad Dataset		
	AUC_{homo}	AUC_{heter}	AUC_{minor}	AUC_{homo}	AUC_{heter}	AUC_{minor}
w/o explicit	0.7257	0.6879	0.7123	0.8647	0.7413	0.7364
w/o implicit	/	/	/	0.8613	0.7367	0.7310
w/o cluster	0.7223	0.6856	0.7098	0.8592	0.7332	0.7284
all2all	0.7189	0.6749	0.7039	0.8593	0.7331	0.7294
similarity	0.7201	0.6783	0.7051	0.8599	0.7358	0.7347
SAMD	0.7257	0.6879	0.7123	0.8667	0.7459	0.7393

5.3 Ablation Study

5.3.1 Scenario-Aware Knowledge Sharing. We analyze the impact of different components of scenario-aware knowledge sharing, including explicit similarity (*w/o explicit*), implicit similarity (*w/o implicit*), and scenario clustering (*w/o cluster*). As mentioned in Section 5.1.1, we do not report the *w/o implicit* on the AliExpress dataset, because the explicit similarity can not be obtained. As shown in Table 2, the results demonstrate the effectiveness of each module. Specifically, we find that *w/o cluster* and *w/o implicit* perform significantly worse than *w/o explicit* on Alipay Ad dataset that contains more scenarios. This suggests that the modeling of scenarios' relationship is more crucial when there are more diverse scenarios. Additionally, *w/o cluster* has the worst performance on minor scenarios, demonstrating that modeling scenarios' relationship allows scenarios with fewer data to benefit from the knowledge of scenarios with more data.

5.3.2 Model-Agnostic Knowledge Sharing. To prove the effectiveness of model-agnostic knowledge sharing, we evaluate different distillation forms, including *all2all distillation*, *semantic similarity distillation* and *SAMD*, as shown in Table 2. *All2all distillation* has the worst performance because it sets the same transfer weights between the general model and the sub-model. *Semantic similarity distillation* is better than *all2all* but relying on pairwise layer-level semantic similarity cannot achieve the best performance in knowledge sharing. *SAMD* generates scenario-aware transfer strategies without manual mapping and outperforms the others in both heterogeneous and minor scenarios, which further emphasizes the importance of flexible model-agnostic knowledge sharing in SAMD.

5.4 Visualization of Knowledge Sharing

To give some in-depth analyses of why SAMD framework works, we investigate the scenarios representation and transfer strategies of different scenarios on Alipay Ad dataset.

5.4.1 Visualization of Scenarios' Relationship. As illustrated in Figure 7, we use t-SNE [14] to present a visual representation of the scenarios representation T_i and cluster representation C_i for different scenarios. Each dot in the figure represents a sample from scenario #N. Initially, the distribution of T_i is uniform in the latent space without significant discrimination. The samples from the same scenario have similar T_i values, but the relationships among scenarios are not clear. After the soft clustering, C_i of similar scenarios become closer, and some scenarios share the latent space with several scenarios. For instance, the advertisements recommended by scenarios #12 and #9 both belong to the e-commerce category, and their cluster embeddings are close to each other. Meanwhile, scenario #13 is far away from scenarios #12 and #9 as it recommends mini-program advertisements. The scenarios in the middle area have overlapping attributes from different scenarios. Taking scenario #7 as an example, its exposure position is closer to scenario #4, and its advertisements have the same categories as scenario #21. Hence, the cluster embedding of scenario #7 is closely related to scenarios #4 and #21. To summarize, the visualization of scenarios highlights the insightful modeling of scenario relationships, which provides high-quality scenario-aware representations for knowledge sharing.

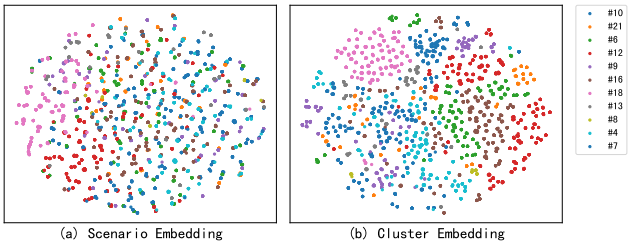

Figure 7: The visualization of (a) scenario embedding T_i and (b) cluster embedding C_i.

5.4.2 Visualization of Scenario-aware Knowledge Sharing. Figure 8 shows the Interpretability of the proposed SAMD approach, which visualizes the scenario-aware of knowledge strategy W_T^i. To highlight the diversity of knowledge sharing, we reshape W_T^i to $\mathbb{R}^{4\times4}$ that is consistent with the number of layers between the general model and sub-model and visualize the intensity of W_T^i in homogeneous scenarios, where darker blue represents higher weight. Overall, there are both similarities and differences among the transfer strategies across different scenarios. The commonality is that the bottom layer typically has lower weights (columns 0), indicating that the bottom layers of the general model contain less general knowledge to be shared. The middle layers (columns 1 and 2) are generally given high weights across all scenarios, reflecting the importance of the knowledge in the middle layer. At the same time, the knowledge sharing also reflects the diversity of scenarios. For example, scenarios #21, #12 and #9 have similar T_i and C_i, and their weights are also similar. On the other hand, the weights of dissimilar scenarios #12 and #10 are obviously different. Additionally, scenario #19 is a minor scenario with limited data. Through meta distillation, scenario #19 can learn to automatically share knowledge with similar scenarios #9 and #13 without manual assignment. As a result, the knowledge sharing of scenario #19 is the combination of scenarios #9 and #13. In conclusion, the visualization of scenario-aware knowledge sharing effectively demonstrates the interpretability of SAMD, allowing us to better understand what and where knowledge should be shared across scenarios.

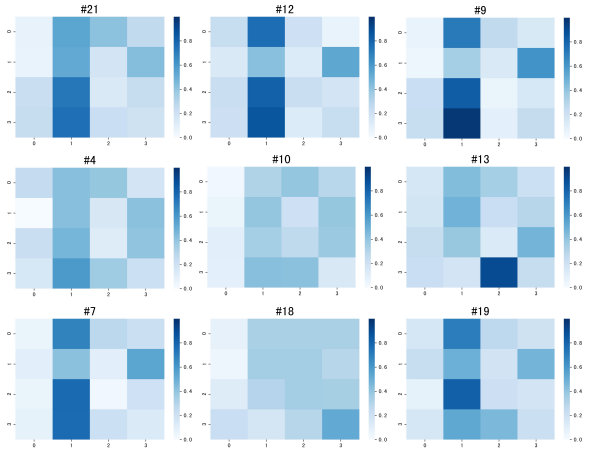

Figure 8: The visualization of scenario-aware knowledge sharing $W_T^i \in \mathbb{R}^{4\times4}$. Each row in the first two rows is W_T^i of similar scenarios. The third row is scenarios that are learned from both scenarios in the first two rows.

6 CONCLUSION

In this paper, we study an important but under-explored problem of heterogeneity in the multi-scenario recommendation and propose an industrial framework named Scenario-Aware Model-Agnostic Meta Distillation (SAMD) for the heterogeneous multi-scenario recommendation. SAMD aims to share scenario-aware and model-agnostic knowledge by modeling the scenarios' relationship and conducting heterogeneous knowledge distillation. Specifically, SAMD first measures the scenario representation from both explicit and implicit views. Then SAMD proposes a novel meta distillation paradigm to incorporate the complicated scenarios' relationship into the heterogeneous knowledge sharing. The meta network automatically establishes the mapping between scenarios and knowledge sharing, which enables similar scenarios to share knowledge with each other. And heterogeneous knowledge distillation shares the model-agnostic knowledge through intermediate features without the restriction of the model architecture. Compelling results from both offline evaluation and online A/B tests demonstrate the superiority of SAMD over state-of-the-art methods, especially in heterogeneous scenarios. And the SAMD has superior deployment efficiency compared with other industrial frameworks. Since 2022, SAMD is deployed in the advertising system of Alipay, obtaining a 6.48% improvement on CVR and 1.06% on CPM.

REFERENCES

[1] Jang Hyun Cho and Bharath Hariharan. 2019. On the efficacy of knowledge distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*.

[2] Ke Ding, Xin Dong, Yong He, Lei Cheng, Chilin Fu, Zhaoxin Huan, Hai Li, Tan Yan, Liang Zhang, Xiaolu Zhang, and Linjian Mo. 2021. MSSM: A Multiple-level Sparse Sharing Model for Efficient Multi-Task Learning. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (2021).

[3] Xidong Feng, Chen Chen, Dong Li, Mengchen Zhao, Jianye Hao, and Jun Wang. 2021. Cmml: Contextual modulation meta learning for cold-start recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*. PMLR.

[5] Byeongho Heo, Jeesoo Kim, Sangdoo Yun, Hyojin Park, Nojun Kwak, and Jin Young Choi. 2019. A comprehensive overhaul of feature distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1921–1930.

[6] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[7] Yunhun Jang, Hankook Lee, Sung Ju Hwang, and Jinwoo Shin. 2019. Learning What and Where to Transfer. In *ICML*.

[8] Jangho Kim, SeongUk Park, and Nojun Kwak. 2018. Paraphrasing complex network: Network compression via factor transfer. *Advances in neural information processing systems* 31 (2018).

[9] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014).

[10] Hoyeop Lee, Jinbae Im, Seongwon Jang, Hyunsouk Cho, and Sehee Chung. 2019. Melu: Meta-learned user preference estimator for cold-start recommendation. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.

[11] Lixin Liu, Yanling Wang, Tianming Wang, Dong Guan, Jiawei Wu, Jingxu Chen, Rong Xiao, Wenxiang Zhu, and Fei Fang. 2022. Continual Transfer Learning for Cross-Domain Click-Through Rate Prediction at Taobao. (2022).

[12] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*.

[13] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H. Chi. 2018. Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2018).

[14] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[15] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550* (2014).

[16] Xiang-Rong Sheng, Liqin Zhao, Guorui Zhou, Xinyao Ding, Binding Dai, Qiang Luo, Siran Yang, Jingshan Lv, Chi Zhang, Hongbo Deng, et al. 2021. One model to serve all: Star topology adaptive recommender for multi-domain ctr prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*.

[17] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Proceedings of the 14th ACM Conference on Recommender Systems*.

[18] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive Layered Extraction (PLE): A Novel Multi-Task Learning (MTL) Model for Personalized Recommendations. *Proceedings of the 14th ACM Conference on Recommender Systems* (2020).

[19] Can Wang, Defang Chen, Jian-Ping Mei, Yuan Zhang, Yan Feng, and Chun Chen. 2022. SemCKD: Semantic Calibration for Cross-Layer Knowledge Distillation. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[20] Ruoxi Wang, Bin Fu, Gang Fu, and Mingliang Wang. 2017. Deep & Cross Network for Ad Click Predictions. In *Proceedings of the ADKDD'17, Halifax, NS, Canada, August 13 - 17, 2017*.

[21] Jieyu Yang, Zhaoxin Huan, Yong He, Ke Ding, Liang Zhang, Xiaolu Zhang, Jun Zhou, and Linjian Mo. 2022. Task Similarity Aware Meta Learning for Cold-Start Recommendation. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4630–4634.

[22] Xiaoyong Yang, Yadong Zhu, Yi Zhang, Xiaobo Wang, and Quan Yuan. 2020. Large scale product graph construction for recommendation in e-commerce. *arXiv preprint arXiv:2010.05525* (2020).

[23] Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4133–4141.

[24] Kaiyu Yue, Jiangfan Deng, and Feng Zhou. 2020. Matching guided distillation. In *European Conference on Computer Vision*. Springer, 312–328.

[25] Sergey Zagoruyko and Nikos Komodakis. 2016. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928* (2016).

[26] Qianqian Zhang, Xinru Liao, Quan Liu, Jian Xu, and Bo Zheng. 2022. Leaving no one behind: A multi-scenario multi-task meta learning approach for advertiser modeling. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*.

[27] Yujing Zhang, Zhangming Chan, Shuhao Xu, Weijie Bian, Shuguang Han, Hongbo Deng, and Bo Zheng. 2022. KEEP: An Industrial Pre-Training Framework for Online Recommendation via Knowledge Extraction and Plugging. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

[28] Yujia Zheng, Siyi Liu, Zekun Li, and Shu Wu. 2021. Cold-start sequential recommendation via meta learner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 4706–4713.

[29] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1059–1068.

[30] Yongchun Zhu, Ruobing Xie, Fuzhen Zhuang, Kaikai Ge, Ying Sun, Xu Zhang, Leyu Lin, and Juan Cao. 2021. Learning to warm up cold item embeddings for cold-start recommendation with meta scaling and shifting networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[31] Yongchun Zhu, Ruobing Xie, Fuzhen Zhuang, Kaikai Ge, Ying Sun, Xu Zhang, Leyu Lin, and Juan Cao. 2021. Learning to warm up cold item embeddings for cold-start recommendation with meta scaling and shifting networks. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

[32] Xinyu Zou, Zhi Hu, Yiming Zhao, Xuchu Ding, Zhongyi Liu, Chenliang Li, and Aixin Sun. 2022. Automatic Expert Selection for Multi-Scenario and Multi-Task Search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.