A Secure and Robust Knowledge Transfer Framework via Stratified-Causality Distribution Adjustment in Intelligent Collaborative Services

Ju Jia, Siqi Ma, *Member, IEEE*, Lina Wang, *Member, IEEE*,
Yang Liu, *Senior Member, IEEE*, and Robert H. Deng, *Fellow, IEEE*

***Abstract*—The rapid development of device-edge-cloud collaborative computing techniques has actively contributed to the popularization and application of intelligent service models. The intensity of knowledge transfer plays a vital role in enhancing the performance of intelligent services. However, the existing knowledge transfer methods are mainly implemented through data fine-tuning and model distillation, which may cause the leakage of data privacy or model copyright in intelligent collaborative systems. To address this issue, we propose a secure and robust knowledge transfer framework through stratified-causality distribution adjustment (SCDA) for device-edge-cloud collaborative services. Specifically, a simple yet effective density-based estimation is first employed to obtain uncertainty scores that guide the space stratification, which is conducive to reconstructing low-density distribution regions from high-density distribution regions more adaptively and accurately. Subsequently, we devise a novel causality-aware generative model to generate synthetic features for the out-of-distribution domain by exploring the relationship between factors and variables. Ultimately, we introduce a cycle-consistent minimax optimization mechanism to ensure the effectiveness and dependability of knowledge transfer through the influence minimization and the diversity maximization. Furthermore, extensive experiments demonstrate that our scheme can protect the security of data privacy and model copyright in intelligent collaborative services through adaptive distribution adjustment.**

***Index Terms*—Intelligent collaborative service, knowledge transfer, privacy preservation, copyright protection, adaptive distribution adjustment.**

I. Introduction

THE rapid development of device-edge-cloud collaborative computing techniques has revolutionized the field of artificial intelligence (AI)[1]. Through the power of multiple device connections, edge computing infrastructures, and centralized cloud resources, this innovative paradigm enables AI models to overcome the limitations of traditional computing architectures [2]. Therefore, a large number of AI tasks arise in our daily lives, including smart healthcare, financial management, and autonomous driving. Since diverse data provide rich information for model selection to obtain good desired results, the feedback data generated by these tasks can potentially enhance the generalization performance of AI models [3]. In addition, the generalization capability of pre-trained models plays a crucial role in boosting the quality of AI services[2]. Currently, domain adaptation (DA) [4], [5], transfer learning (TL) [6], [7] and out-of-distribution (OOD) generalization [8], [9] are usually employed to improve the knowledge transferability and generalization performance of AI models.

Typical DA or TL methods are inevitably prone to discover shared or transferable representations from the source scene to fulfill the tasks in the target scene [10]. While most recent works [9], [11] of OOD generalization mainly attempt to extrapolate and optimize the empirical risks in the objective function, which allows a model to generalize to a new testing domain. However, when faced with severe distribution shifts caused by multiple bias factors (MBFs), such as domain differences [4], skewed distributions [7], and incorrect labels [5], the accuracy of the model decreases dramatically due to the lack of robustness. To be specific, since the model does not take into account the diverse scenarios presented by the MBFs during training, this leads to poor robustness and performance on unseen data. In other words, it is difficult to make robust and accurate predictions because the model is sensitive to the shifts between source and target features. Moreover, enhancing the accuracy in limited

Manuscript received 24 April 2023; revised 12 September 2023; accepted 17 September 2023. Date of publication 22 September 2023; date of current version 22 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62372334, in part by the Start-up Research Fund of Southeast University under Grant RF1028623129, in part by the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme under AISG Award AISG2-GC-2023-008, and in part by the National Research Foundation, Singapore, and the Cyber Security Agency under its National Cybersecurity R&D Programme under Grant NCRP25-P04-TAICeN. Recommended for acceptance by X. Jia. *(Corresponding author: Ju Jia.)*

Ju Jia is with the School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: jiaju@seu.edu.cn).

Siqi Ma is with the School of System and Computing, University of New South Wales, Canberra, ACT 2612, Australia (e-mail: siqi.ma@unsw.edu.au).

Lina Wang is with the Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: lnawang@163.com).

Yang Liu is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: yangliu@ntu.edu.sg).

Robert H. Deng is with the School of Information Systems, Singapore Management University, Singapore 188065 (e-mail: robertdeng@smu.edu.sg).

This article has supplementary downloadable material available at https://doi.org/10.1109/TC.2023.3318403, provided by the authors.

Digital Object Identifier 10.1109/TC.2023.3318403

[1] https://aws.amazon.com/cn/ [1]

[2] https://openai.com/research/ai-and-compute

0018-9340 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

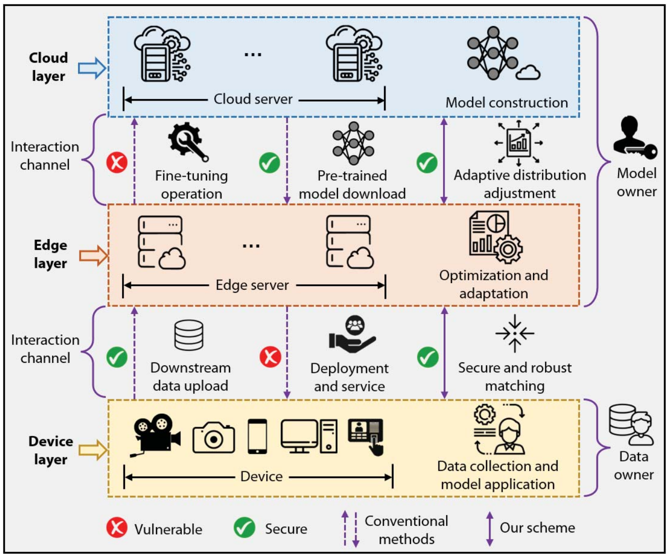


Fig. 1. The entire architecture diagram for artificial intelligence (AI) services through knowledge transfer in device-edge-cloud collaborative systems. The conventional approaches enhance the generalization capability by data upload and model fine-tuning, and then provide AI services through model download and adaptive deployment. In this way, there exist two main limitations in prior work: 1) the problem of data privacy leakage and 2) the risk of model copyright violation. However, we propose a stratified-causality distribution adjustment scheme, *i.e.*, adaptively filling low-density data regions from high-density data regions, which ensures the security and robustness of the knowledge transfer process.

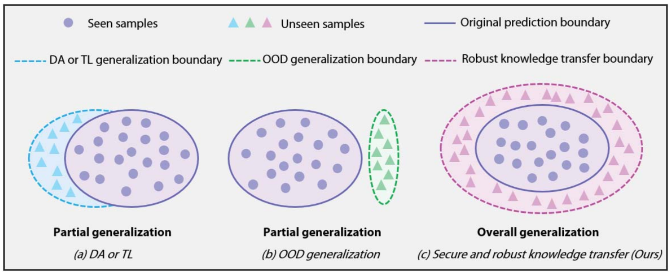

Seen samples
Unseen samples
Original prediction boundary
DA or TL generalization boundary
OOD generalization boundary
Robust knowledge transfer boundary
Partial generalization
(a) DA or TL
Partial generalization
(b) OOD generalization
Overall generalization
(c) Secure and robust knowledge transfer (Ours)

Fig. 2. The comparison of generalization performance for different methods under severe distribution shift conditions. (a) Due to the adverse impacts of MBFs, the direct application of traditional DA or TL methods will lead to the extremely limited generalization in the target domain. (b) OOD generalization can only complete the partial generalization in a certain range under the condition of severe distribution shifts. (c) Our scheme is implemented to resist the severe distribution shifts and achieve the secure and robust knowledge transfer in an overall generalization manner.

target domains or the generalization without goal can merely be regarded as a partial approximation to the real-world scenarios. Thus, a more practical problem is how to guarantee the effectiveness of knowledge transfer in a remarkable distribution shift scenario, which can make the model more pointed and powerful to achieve the overall generalization.

The existing knowledge transfer methods exhibit excellent performance in ideal or specific scenarios [7]. In other words, these approaches either focus on improving the detection accuracy or the generalization ability in target tasks. However, in intelligent collaborative systems, the direct utilization of sensitive data and the duplicate copy of pre-trained model may raise security concerns of data privacy leakage and model copyright infringement [12], [13]. Fig. 1 provides an entire architecture for implementing AI services through knowledge transfer, in which the differences between traditional methods and our scheme are compared. As a result, the direct application of existing knowledge transfer methods in intelligent collaborative services poses the risk of data privacy leakage and model copyright infringement. Moreover, we show that only taking into account the improvement of domain-specific accuracy or task-independent generalization is insufficient for constructing robust knowledge transfer framework. We give counterexamples from two aspects under the condition of MBFs in device-edge-cloud collaborative systems: 1) a domain-specific model has satisfactory accuracy but usually fails to generalize to other OOD domains and 2) a task-independent model has good generalization ability but may fail to recognize unseen samples in some particular domains.

Fig. 2 compares the challenges faced by different solutions to improve the intensity of knowledge transfer in severe distribution shift scenarios. Specifically, when the learned source knowledge is deployed to assist target classification, the source-target distribution alignment is not enough to increase the generalization performance in the target domain. The reason is that the distribution alignment is achieved only when the distribution of stable representations is the same across different domains. However, it is difficult to satisfy this requirement under the adverse impacts of MBFs [14]. In addition, the OOD generalization boundary of the model may not fully cover a given task in real-world datasets due to the lack of robustness [15]. For instance, in digit classification, handwriting styles vary from individual to individual, which may lead to misalignment and overfitting generalization performance [4]. Our investigation reveals that simultaneously considering the improvement of the overall generalization performance and the specific prediction capability can resist the severe distribution shifts caused by MBFs in intelligent collaborative services.

To guarantee secure and robust knowledge transfer in device-edge-cloud collaborative platforms, we propose a stratified-causality distribution adjustment (SCDA) scheme based on flexible space division and latent representation inference. The motivation of leveraging adaptive distribution adjustment is to manipulate data at the distribution level to prevent privacy leakage and copyright infringement, while ensuring the security and robustness of knowledge transfer under the condition of MBFs. The requirement for the direct sharing of raw data across domains is minimized by reconstructing low-density distribution regions from high-density distribution regions in an adaptive and accurate manner. Since unique features and characteristics are protected during knowledge transfer, the adaptive distribution adjustment plays a crucial role in safeguarding the copyright of models. Specifically, we develop a density-based ratio estimation strategy that provides an uncertainty score for each sample to describe the anomaly degree and complete the partitioning of subspace. The conditional variational autoencoder (CVAE) is exploited to construct a novel causality-aware data generation module, which combines causal and non-causal attributes across different subspaces to infer the effects of the interference-free and interference-active shifts by performing counterfactual analysis. Finally, a cycle-consistent minimax optimization mechanism is introduced to minimize the influence and maximize the diversity of synthetic features, which enables a perfect distribution matching for the implementation of secure and robust knowledge transfer in intelligent collaborative scenarios. Moreover, we evaluate our proposed SCDA on Rotated MNIST, CIFAR-10 & STL-10 and Colored PACS datasets, and promising results on all datasets have demonstrated the effectiveness of the proposed scheme. In summary, the contributions of our work are as follows:

- We propose a density-based ratio estimation strategy to quantify the uncertainty degree for each sample, which can guide and adjust the partitioning of subspace by analyzing the distribution characteristics. The experimental evaluation in Section VI-D shows that the reasonable subspace partition can distinguish the anomaly degree and is more suitable for deployment under secure and robust knowledge transfer scenarios.
- We exploit the adaptive data generation to infer and expand the prior knowledge by combining basic features and state features under the guidance of counterfactual causality. In this way, low-density distribution regions can be adaptively reconstructed from high-density distribution regions, which is crucial to preserve in-distribution (ID) samples and expand beneficial OOD samples in intelligent collaborative services.
- We present a cycle-consistent minimax optimization mechanism for ensuring robustness of knowledge transfer by identifying learnable and informative synthetic samples from the given candidate pool. Large scale experiments using ablation analysis demonstrate that our scheme can achieve superior performance by simultaneously considering both the influence minimization and the diversity maximization.
- To the best of our knowledge, this is the first work to develop a stratified-causality distribution adjustment for secure and robust knowledge transfer in device-edge-cloud collaborative services. Our proposed scheme not only protects data privacy and model copyright, but also withstands distribution shifts and eliminates potential negative effects.

The rest of the article is organized as follows. Section II reviews the related work. Section III introduces the fundamental background and preliminary knowledge. Section IV analyzes and gives the threat model. Section V describes the proposed SCDA scheme for secure and robust knowledge transfer. Section VI evaluates and compares our novel SCDA with the related state-of-the-art methods in extensive experiments. Finally, the conclusions and future research directions are summarized in Section VII.

II. Related Work

A. Data-Centric Distribution Matching

A data-centric idea is mainly designed and implemented based on data augmentation to expand available datasets, control model overfitting, provide the interoperability across different distributions and obtain superior results in the classification tasks [14], [16]. Data augmentation has received attention in existing works for distribution matching [17]. For example, Hsu et al. [18] presented an augmentation-based approach to generate disentangled latent representations of speech with extra data whose distribution is more close to the desired target data. However, the disentangled image representations in the latent space are a challenging issue to achieve. Instead, we employ a conditional variational autoencoder (CVAE) where the learned transferable knowledge can be merged and fed into the decoder in order to generate the expected features. In addition, Huang et al. [17] proposed a weakly supervised generative adversarial network (GAN) based model for image-to-image translation and investigated the performance in image-translation tasks rather than classification tasks which is our focus in this work. Then, Ng et al. [19] formulated a data augmentation framework based on self-supervised manifolds to improve out-of-domain robustness by using a pair of reconstruction and corruption functions. Different from these studies, in our work, a novel distribution adjustment based on data augmentation is proposed to adaptively generate target-specific and diverse features by stratified causality learning across distributions and its effectiveness is verified through comparative experimental results in Section VI.

B. Causality-Aware Knowledge Exploration

Incorporating causal perception [20] into deep neural networks (DNNs) has recently attracted more and more attention [21]. It can not only improve the interpretability of DNNs, but also boost the exploration of meaningful knowledge and causal relationship in essence. Therefore, it has been widely used in many fields, such as image recognition [22], personalized recommendation [23], visual question answering [24] and scene graph generation [25]. Yue et al. [22] presented a counterfactual framework for generating synthetic samples to address both zero-shot learning and open-set recognition problems. Following this, Wang et al. [23] developed the recommendation models with a causal graph to reflect the cause-effect factors by performing counterfactual inference. Inspired by the great success of the causal theory, Wang et al. [24] exploited a visual commonsense region-based network for visual question answering tasks via causal intervention, which can capture sense-making knowledge to alleviate the observational bias. Then, Tang et al. [25] introduced a scene graph generation (SGG) framework to deal with the biased data distributions of SGG by counterfactual causality. Although these methods perform well in above tasks, the robustness of knowledge transfer cannot be adequately addressed and analyzed through goal-less causal reasoning in intelligent collaborative services. Therefore, our work introduces causal perception into the knowledge transfer task to reason about the effects of different shifts by enforcing counterfactual causality on a task-oriented causal representation learning.

C. Spurious Correlations and OOD Generalization

The recent literature has shown that DNNs may learn superficial representations to make the knowledge transfer, such as by relying on the background regions or other kinds of spurious rules [26]. Such case raises practical concerns because the accuracy may deteriorate under the shifts in those spurious correlations [15]. Simultaneously, it can also result in the unfair bias and worse performance on minority categories under the condition of imbalanced scenarios. Since Arjovsky et al. [9] introduced invariant predictions into a more realistic situation, a considerable amount of research has made remarkable progress in mitigating spurious correlations and capturing stable features. Krueger et al. [15] presented the risk extrapolation to enforce and constrain the variance of losses across different distributions. Similarly, Koyama and Yamaguchi [11] formulated a comprehensive set of theoretical analysis and guidance for an invariant representation to fulfill the OOD optimality. Kamani et al. [8] provided a unified data-driven regularization framework to construct a generalizable model from biased domain knowledge, which can exploit a target dataset that approximates the essence of the required test data to reinforce the learning capacity of the model. Moreover, the biases in the realistic datasets are usually employed in a spurious way, so the robustness of OOD generalization will be poor. Another line of research [27] aims to design a debiasing technique to address the bias problem and obtain better performance. Some recent works [28], [29] have started to focus on the privacy-preserving issues during knowledge transfer or OOD generalization. To our best knowledge, the proposed SCDA is the first framework for knowledge transfer to improve the overall generalization performance with the security and robustness consideration in device-edge-cloud collaborative systems. In addition, our SCDA scheme considers not only the adaptive distribution adjustment but also the matching between datasets and models (*e.g.*, stratified causal perception and cycle-consistent minimax optimization).

III. Preliminaries

A. Robust OOD Generalization

In the classical OOD generalization setting, given the features $\{x_{\epsilon_T}^1, ..., x_{\epsilon_T}^n\}$ from a testing domain ϵ_T, we aim to learn a function to identify the labels $\{y_{\epsilon_T}^1, ..., y_{\epsilon_T}^n\}$ using labeled data $\{(x_{\epsilon_A}^1, y_{\epsilon_A}^1), ..., (x_{\epsilon_A}^m, y_{\epsilon_A}^m)\}$ from an available domain ϵ_A [9]. Let X and Y represent the variable of feature and label, respectively. In contrast to the standard supervised learning scenario, the joint distributions $P_{\epsilon_A}(XY)$ and $P_{\epsilon_T}(XY)$ are mismatched. Specifically, if there are two datasets D_1 and D_2 collected from two subspaces, we assume that they are derived from $P_{\epsilon_1}(XY)$ and $P_{\epsilon_2}(XY)$ with different subspaces of ϵ_1 and ϵ_2. If $l(f(X), Y)$ denotes the error between the ground truth Y and the prediction $f(X)$, the OOD generalization problem can be formulated by finding f^* that solves

$$\min_f \max_{\epsilon \in \varepsilon} \mathbb{E}_{X,Y}[l(f(X), Y)|\epsilon], \tag{1}$$

where ε is the space of all possible testing domains. To improve the generalization performance, existing methods usually make the conditional distribution $P(\delta(X)|Y)$ invariant across domains by assuming the existence of a transformation δ. In this article, we also assume that the conditional invariant component exists. The common goal is to find a transformation δ such that $P_{\epsilon_A}(\delta(X)|Y) = P_{\epsilon_T}(\delta(X)|Y)$ and to calculate $P_{\epsilon_T}(Y)$. However, we consider a more challenging OOD generalization scenario, namely the samples drawn only from the distribution $P_{\epsilon_T}(X)$ and the biased distribution $P_{\epsilon_A}(XY)$ under the condition of interference-free and interference-active shifts, which is regarded as robust OOD generalization.

Note that the simple combination of distribution adaptation and robust classifier ignores that the learning of invariant representations can be influenced by complex domain shifts, which may lead to significantly biased results. Learning δ becomes even more challenging in the setting where only mismatched training data and insufficient unlabeled testing data are available. The reason is that without unbiased label Y in both available and testing domains, there is no critical information to ensure the consistency of conditional distribution $P(\delta(X)|Y)$. In addition, it is difficult to learn δ and to calculate $P_{\epsilon_T}(Y)$ as discussed above. Therefore, we propose a novel stratified-causality-based adaptive distribution adjustment scheme to identify $P_{\epsilon_T}(Y)$ and capture the conditional invariant component $\delta(X)$ for robust OOD generalization.

B. Invertible Mapping via Normalizing Flows

The deep invertible generation models, or normalizing flows, are a set of likelihood-based mappings that reflect the bidirectional transformation relationship between a simple and a complex continuous probability density through the change of variables formula. The flows are parameterized by deep neural networks $g_\tau : \mathcal{X} \to \mathcal{Z}$ with well-designed architectures so that the entire transformation consists of diverse two-way mappings with the tractable inverse and the Jacobian determinant [30]. Similarly, $g_\tau^{-1} : \mathcal{Z} \to \mathcal{X}$ represents the inverse of the mapping function g_τ. As a consequence, the probability density p of random variable $X = g_\tau^{-1}(Z)$ can be accurately calculated:

$$p(x) = q(g_\tau(x)) \left| \det \frac{\partial g_\tau(x)}{\partial x} \right|, \tag{2}$$

where q denotes the probability density of the variable Z. The basic (prior) distribution $q(z)$ is generally selected as an isotropic Gaussian $\mathcal{N}(0, I)$, and the simplicity of estimating this prior density, together with the tractability of g_τ^{-1} and its Jacobian, enables us to obtain the normalizing flow through maximum likelihood. We exploit the key invertibility of normalizing flows to compute the uncertainty based on density ratio estimation (DRE). That is, if the design dimensions of $\mathcal{X}$ and $\mathcal{Z}$ are the same, and data points can be mapped mutually and losslessly between the two latent spaces. As we will discuss in Section V, this will be vital for converting the density ratios learned in the latent space back to that obtained in the input space. The improved density ratio estimation makes the mismatched distributions closer to each other in the latent space, which is beneficial to calculate the uncertainty scores more accurately.

Due to the irreversibility, the non-invertible neural networks may lead to imprecise space division, which in turn affects the overall performance. However, in this article, the invertible normalizing flow is trained on a mixture of datasets collected from the different distributions and then exploited to map the

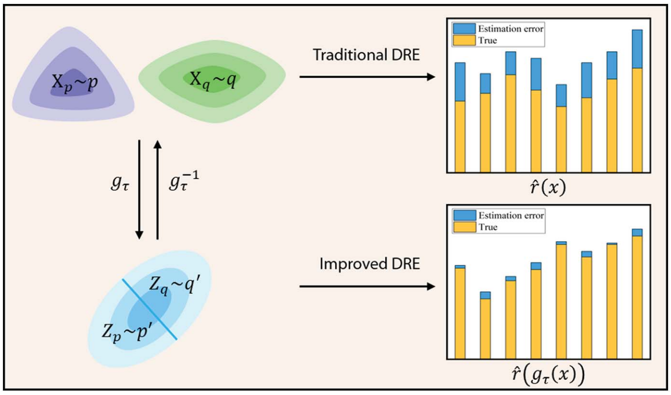


Fig. 3. Flowchart of traditional DRE and DRE based on invertible normalizing flows. Traditional DRE directly applying a black-box algorithm on samples will result in an inaccurate ratio estimation $\hat{r}(x)$ when p and q are drawn from dramatically different distributions. The normalizing flow g_τ is trained on samples from diverse densities, which can be used to obtain more accurate ratio estimation $\hat{r}(g_\tau(x))$ by encoding biased data in a shared feature space.

samples into a shared feature space. In this way, the data are encoded by leveraging the invertibility of normalizing flow, and the observed samples of different densities are converted to be located in a unit Gaussian sphere. Moreover, we observe that this contraction contributes to mitigating the distribution shift to improve the learned ratio estimation. In other words, the invertibility of feature map ensures that the ratios computed in the latent space are the same as those in the input space. Therefore, the key component of DRE is an invertible normalizing flow to complete the calculation of the uncertainty scores. A flowchart of DRE based on invertible mapping via normalizing flows can be seen in Fig. 3.

C. Causal Intervention via Interaction Adjustment

The causal graph [31] is a directed acyclic graph (DAG), expressed as $\mathcal{G} = \{\mathcal{N}, \mathcal{R}\}$, which describes how a set of nodes (variables) $\mathcal{N}$ interact with each other by a causal relationship $\mathcal{R}$. Since the basic features (*e.g.*, "class attributes") and the state features (*e.g.*, "transformation attributes") mutually affect the prediction results from different aspects, we construct a node-link causal graph to investigate potential causal relationships among the basic features, the state features and the class labels, as illustrated in Fig. 4. The node U denotes the basic feature x_u, the node V represents the state feature x_v, and the node Y is the class label. The edge $U \rightarrow Y$ indicates that basic feature U is used to predict class label Y, and the edge $V \rightarrow Y$ denotes that the state feature V is used to predict class label Y. We learn these edges by a classification network, which exploits the conventional cross-entropy loss between the class labels y_{ϵ_A} and the feature mapped to $\mathcal{X}_{\epsilon_T}$.

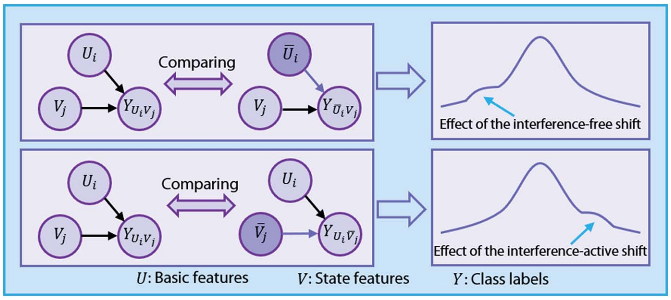


Fig. 4. A node-link causal graph to analyze the effect of the interference-free and interference-active shifts.

We perform counterfactual intervention [32] via interaction adjustment on the node-link causal graph to analyze the effect of the interference-free and interference-active shifts from the basic and state features. Specifically, $do(U = u)$ means that we assign a certain value u to the node U through different combinations of basic features. Given a sample from the subspace ϵ, denoted as x_ϵ, we have $U = x_{\epsilon,u}$, $V = x_{\epsilon,v}$, and the output Y is represented as $Y_{UV} = y_{x_{\epsilon,u}x_{\epsilon,v}}$. A counterfactual reasoning scenario is defined by performing the intervention on the basic feature to evaluate the effect, where the basic feature is erased by $do(U = \overline{x}_{\epsilon,u})$ and the node V is retained as the original state feature $x_{\epsilon,v}$. The erased basic feature $\overline{x}_{\epsilon,u}$ is set as the zero vector of the same dimension as $x_{\epsilon,u}$. The output variable Y after intervention is represented as a counterfactual $Y_{\overline{U}V} = y_{\overline{x}_{\epsilon,u}x_{\epsilon,v}}$, which is distinguished from Y_{UV}. It is easy to infer the effect of the interference-free shift on x_ϵ by comparing the deviation between Y_{UV} and $Y_{\overline{U}V}$, which is formulated as

$$FE(x_\epsilon) = Y_{UV} - Y_{\overline{U}V} = y_{x_{\epsilon,u}x_{\epsilon,v}} - y_{\overline{x}_{\epsilon,u}x_{\epsilon,v}}. \quad (3)$$

Similarly, we erase the state feature by $do(V = \overline{x}_{\epsilon,v})$ while maintaining the node U as the original basic feature $x_{\epsilon,u}$, and obtain the counterfactual $Y_{U\overline{V}} = y_{x_{\epsilon,u}\overline{x}_{\epsilon,v}}$. Then, the effect of the interference-active shift on x_ϵ is formulated by

$$AE(x_\epsilon) = Y_{UV} - Y_{U\overline{V}} = y_{x_{\epsilon,u}x_{\epsilon,v}} - y_{x_{\epsilon,u}\overline{x}_{\epsilon,v}}. \quad (4)$$

We evaluate the influence of the two mode shifts for each class in the subspace ϵ by averaging the estimated effects $FE(x_\epsilon)$ and $AE(x_\epsilon)$, which are formulated as follows

$$FE_\epsilon^k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{I}(\hat{y}_\epsilon^i = k) FE(x_\epsilon^i), \quad (5)$$

and

$$AE_\epsilon^k = \frac{1}{N_k} \sum_{i=1}^{N_k} \mathbb{I}(\hat{y}_\epsilon^i = k) AE(x_\epsilon^i), \quad (6)$$

where FE_ϵ^k and AE_ϵ^k denote the influence of the interference-free and interference-active shifts of the k-th class, respectively, $\hat{y}_\epsilon^i$ is the observed label of x_ϵ^i, and N_k represents the number of samples classified into the k-th class. $\mathbb{I}(\cdot)$ denotes an indicator function, indicating that if $\hat{y}_\epsilon^i = k$, the value of $\mathbb{I}(\hat{y}_\epsilon^i = k)$ is 1, otherwise 0.

Based on the above analysis, in the absence of stratified causal reasoning, it is hard to figure out the influence of different variables on the distribution shifts, and thus unable to learn generalizable and discriminative representations. Motivated by that, we deploy a stratified-causality distribution adjustment scheme, which can adaptively improve the quality of data distribution to ensure the security and robustness of knowledge transfer in intelligent collaborative applications.

IV. Threat Model

We discuss the objective, capability and knowledge of the adversary for stealing attacks. We consider a scenario in which a victim uploads downstream data to update the pre-trained model by adopting transfer learning techniques (*e.g.*, fine-tuning operation or knowledge distillation), and then seeks to improve the performance for better AI services by deploying the updated model.

The objective of adversary. The adversary aims to steal the private data or violate the model copyright during intelligent collaborative scenarios in one of the following ways:

- *Privacy leakage attack.* An adversary tries to perform illegal access in data transmission or storage on various service resources (*e.g.*, device, edge, cloud). This includes potential attacks, such as malicious data tampering, unauthorized data access, and private data leakage.
- *Copyright infringement attack.* The knowledge transfer involves the sharing of model parameters and the alignment of intermediate representations, which can lead to the risk of intellectual property theft. Therefore, an adversary may exploit vulnerabilities in deployment and application to extract or clone well-trained models used under intelligent collaborative services.

The capability of adversary. During the transfer learning process, an adversary can adopt malicious samples and then induce the model to output sensitive data by executing queries. Since this attack does not make any manipulation on the specific samples, we mainly consider the performance of knowledge transfer under the condition of protecting data privacy. Moreover, during the deployment and application of a well-trained model, an adversary may steal the intellectual property through direct replication operations or model surrogate attacks.

The knowledge of adversary. The adversary can access the pre-trained model by using elaborated samples and then record the correct predictions (*i.e.*, the real decision-making behavior of model) through input-output pairs from the accessible test batch. However, the adversary cannot access the training data or the training process. In this article, the proposed SCDA scheme focuses on the effectiveness of handling severe distribution shifts while mitigating the risk of data leakage and copyright infringement.

V. The Proposed Scheme

A. Detailed Problem Statement

We formulate the secure and robust knowledge transfer as a new setting since the severe distribution shifts caused by multiple bias factors are more realistic but challenging problems, where 1) we only have access to mismatched data in the available domain ϵ_A and unlabeled data in the testing domain ϵ_T; 2) both P_Y and $P_{X|Y}$ change across different bias conditions. We find that if the information of interference is obtained from the observation data, a specific relationship between $P_{\epsilon_A}(\delta(X)|Y)$ and $P_{\epsilon_T}\delta(X)$ can be established, which is, in turn, a cue for us to learn invariant components from $\delta(X)$. Moreover, we observe that the slight label error does not impact the distribution of $\delta(X)$. So, intuitively, if we remove the variable $\hat{Y}$ from the inaccurate labels, $P_{\epsilon_A}(\delta(X)|Y) = P_{\epsilon_T}(\delta(X)|Y)$ can be achieved by aligning the marginal distribution $P_{\delta(X)}$. With the above analysis, in this article, we try to solve the secure and robust knowledge transfer problem by adaptive distribution adjustment scheme, which combines the advantage of uncertainty quantification and causal perception and has feasibility of handling highly biased feature distributions across domains to improve the performance under intelligent collaborative scenarios. Consequently, data privacy leakage and model copyright infringement can be avoided through adaptive distribution adjustment. The pipeline of SCDA is presented in Fig. 5. The details of each stage will be elaborated in the subsequent sections.

B. Adjustable Space Division With Uncertainty Quantification

The challenge of uncertainty quantification is to guarantee that the distribution discrepancy between p and q is closer together to make the DRE problem feasible and tractable. In our article, the degree of uncertainty for data points is described based on DRE, which can eliminate outlier samples and achieve subspace division by setting thresholds. To accomplish this goal, we leverage an invertible mapping to perform density estimation and uncertainty calculation for each sample in the latent space. Specifically, we train a model g_τ on a mixture density of $p(x)$ and $q(x)$ using an invertible deep generative network, such that both $g_\tau(X_p)$ and $g_\tau(X_q)$ are mapped to the common feature space $\mathcal{Z}$. By mapping the low density region in $\mathcal{X}$ to the high density region in $\mathcal{Z}$, and training our probability classifier c_ϕ on $g_\tau(x) \in \mathcal{Z}$ instead of $x \in \mathcal{X}$ directly, this contraction enables a more accurate density ratio to be learned. Let $X_p \sim p$ denote a random variable with density p, and $X_q \sim q$ represent a random variable with density q. If there exists an invertible mapping g_τ, so that p' and q' are the densities of $Z_p = g_\tau(X_p)$ and $Z_q = g_\tau(X_q)$ respectively, then the following equation can be obtained for any x:

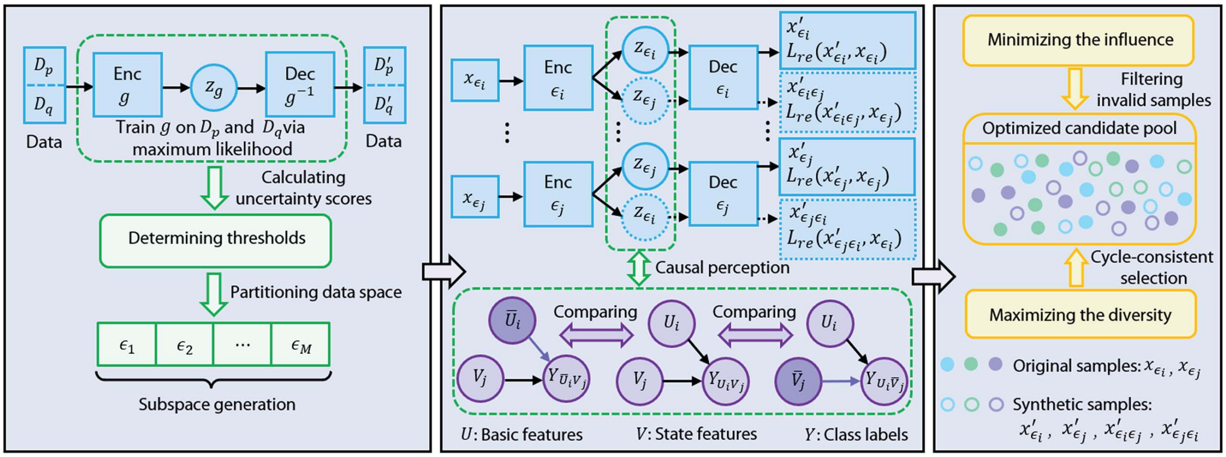
$$\frac{p(x)}{q(x)} = \frac{p'(g_\tau(x))}{q'(g_\tau(x))}. \tag{7}$$

In practice, we employ a pre-trained flow g_τ as an invertible encoder to map the inputs into the common feature space with the separate training strategy, which is able to handle all parametric and non-parametric models working directly in the input space. For example, in the probabilistic classification case, the DRE algorithm needs to realize the binary classifier c_ϕ to identify D_p and D_q. To facilitate this process, we adapt the normalizing flow g_τ to learn the known structure of c_ϕ. As a consequence, the normalizing flow g_τ and the discriminant classifier c_ϕ can be jointly trained according to the following objective function:

$$\mathcal{L}_{joint}(\tau, \phi) = \lambda \mathcal{L}_{sup}(\tau, \phi) + (1 - \lambda)\mathcal{L}_{flow}(\tau), \tag{8}$$

where $\mathcal{L}_{sup}$ represents the standard binary cross entropy (BCE) loss, $\mathcal{L}_{flow}$ is the maximum likelihood estimator of the flow g_τ, and $\lambda \in [0, 1]$ denotes a tuning parameter which balances the importance between the $\mathcal{L}_{sup}$ and $\mathcal{L}_{flow}$ in the loss function.

Inspired by the excellent performance of invertible neural networks (INNs) [30], [33], we modify the architecture of


D_p
D_q
Data
Enc g
z_g
Dec g^{-1}
D'_p
D'_q
Data
Train g on D_p and D_q via maximum likelihood
Calculating uncertainty scores
Determining thresholds
Partitioning data space
ϵ_1 ϵ_2 ⋯ ϵ_M
Subspace generation
x_{ϵ_i}
x_{ϵ_j}
Enc ϵ_i
Enc ϵ_j
z_{ϵ_i}
z_{ϵ_j}
Dec ϵ_i
Dec ϵ_j
x'_{ϵ_i} $L_{re}(x'_{\epsilon_i}, x_{\epsilon_i})$
$x'_{\epsilon_i\epsilon_j}$ $L_{re}(x'_{\epsilon_i\epsilon_j}, x_{\epsilon_j})$
x'_{ϵ_j} $L_{re}(x'_{\epsilon_j}, x_{\epsilon_j})$
$x'_{\epsilon_j\epsilon_i}$ $L_{re}(x'_{\epsilon_j\epsilon_i}, x_{\epsilon_i})$
Causal perception
$\overline{U}_i$
V_j
$Y_{\overline{U}_iV_j}$
Comparing
U_i
V_j
$Y_{U_iV_j}$
Comparing
U_i
$\overline{V}_j$
$Y_{U_i\overline{V}_j}$
U: Basic features
V: State features
Y: Class labels
Minimizing the influence
Filtering invalid samples
Optimized candidate pool
Cycle-consistent selection
Maximizing the diversity
Original samples: x_{ϵ_i}, x_{ϵ_j}
Synthetic samples: x'_{ϵ_i}, x'_{ϵ_j}, $x'_{\epsilon_i\epsilon_j}$, $x'_{\epsilon_j\epsilon_i}$

(a) Adjustable space division *(b) Stratified causality-aware sample generation* *(c) Cycle-consistent minimax optimization*

Fig. 5. Overview of our proposed scheme. (a) Adjustable space division. The uncertainty scores are exploited to map the samples to different subspaces to determine the degree of importance for the original data transformed to an adjacent augmented space. (b) Stratified causality-aware sample generation. The encoder and decoder are trained and conditioned on the different subspaces ϵ_i or ϵ_j ($i, j \in [1, n]$). We consider interventions $do(U_i = \overline{U}_i)$ and $do(V_j = \overline{V}_j)$ to perform counterfactual analysis to infer the effects of the interference-free and interference-active shifts. Given a hidden representation z_{ϵ_i} or z_{ϵ_j} as the input, the model generates reconstructed samples x'_{ϵ_i} or $x'_{\epsilon_i\epsilon_j}$ from the subspace ϵ_i with the guidance of counterfactual causality. (c) Cycle-consistent minimax optimization. Minimizing the influence aims at improving the quality of individual samples by removing invalid ones, while maximizing the diversity is designed to obtain a diverse dataset from candidate pool without considering quality.

classifier c_ϕ by discriminative training and combine it with the architecture of g_τ to construct a probabilistic classifier $c_{\phi,\tau} : \mathcal{X} \to [0, 1]$, which is trained only by BCE loss $\mathcal{L}_{sup}(\tau, \phi)$. Moreover, it is assumed that the density of reliable samples is roughly similar to that of adjacent samples, while the density of unreliable samples is significantly different from that of its neighbors. Therefore, the greater the relative density ratio, the higher the degree of uncertainty for samples. Based on this consideration, for the given data point, the ratio of density to its neighboring density is calculated as the uncertainty score, which can be formulated mathematically as follows:

$$US(x) = (\hat{r} \circ g_\tau)(x) = DRE(p'(g_\tau(x)), q'(g_\tau(x))), \quad (9)$$

where $(\hat{r} \circ g_\tau)(x)$ denotes the composite function $\hat{r}(g_\tau(x))$, and $DRE(\cdot)$ is density ratio estimation function that can be obtained by directly utilizing or slightly modifying the existing methods. Ultimately, adjustable space division is achieved to generate diverse subspaces by exploring and setting different thresholds α based on the computed uncertainty scores.

The complete procedure of adjustable space division using uncertainty quantification is summarized in Algorithm 1 in Appendix B. It can be observed that, except for the threshold determination, the uncertainty partitioning process exploits the pre-trained normalizing flow to encode samples into a common feature space and then inserts them into the base $DRE(\cdot)$ to obtain more accurate uncertainty scores, which are implicitly coupled with the trained flow.

C. Adaptive Data Augmentation Using Stratified Causality

After inferring the effects of the interference-free and interference-active shifts, our method uses the stratified causal perception to promote the knowledge transfer for the purpose of generating synthetic features and reducing domain disparity, which does not aim to directly tackle the distribution mismatch problem. Instead, it learns prior knowledge from dense sample regions (low uncertainty scores) to adaptively assist the synthesis in sparse sample regions (high uncertainty scores), so as to focus on the scarcity of available data in a goal-directed way.

Our proposed causality-aware adaptive generation model is based on the conditional variational autoencoder (CVAE) and the causal-effect relationships of domain shifts. Given input samples x_{ϵ_i} from different subspaces ϵ_i ($i \in [1, n]$), the encoder tries to hierarchically learn diverse distributions $p_\psi(z_{\epsilon_i})$ from which the latent encoding variables z_{ϵ_i} can be causally selected and then fed into the decoder to obtain the reconstructed input samples, where ψ denotes the underlying parameter of the model. The decoder can be parameterized with $p_\psi(z_{\epsilon_i}|z_{\epsilon_j})$, so that the model can generate the synthetic samples in the subspace ϵ_j with the prior knowledge of the samples in the subspace ϵ_i, and vice versa. In basic CVAE, the loss function consists of two terms: the reconstruction loss (the first term) and the KL-divergence between the learned and standard normal distribution (the second term), formulated as

$$\mathcal{L}_{cvae}(x_{\epsilon_i}; \psi) = \mathcal{L}_{re}(x'_{\epsilon_i}, x_{\epsilon_i}) + D_{kl}(\mathcal{N}(\mu_{\epsilon_i}, \sigma_{\epsilon_i})||\mathcal{N}(0, I)), \quad (10)$$

where x'_{ϵ_i} is the reconstructed sample of x_{ϵ_i} from the subspace ϵ_i, and μ_{ϵ_i} and σ_{ϵ_i} denote the mean and standard deviation in the subspace ϵ_i, respectively. The KL-divergence is a regularization term that forces the learned latent representation z_{ϵ_i} to obey the standard normal distribution. This regularization ensures the learned model obtain the ability of generating beneficial data based on a random latent variable from a standard normal distribution.

However, our goal is to generate synthetic features by causal perception from latent variable layers rather than by random sampling. To this end, the following two aspects need to be considered. On the one hand, with the guidance of the shift effects, we are able to achieve the causal perceptive sampling of meaningful variables from the latent encoding layer for feature synthesis. On the other hand, the latent codes require to be more discriminative across subspaces than the original, so we relax the constraint of D_{kl} in Eq. (10) by replacing it with L_2 normalization. Moreover, L_2 normalization can force the mean and standard deviation vectors located on the hypersphere rather than around the origin to facilitate the distinguishability of latent representation z_{ϵ_i}.

As previously described in Section III-C, the causal perception compares the influence of two different shifts in the subspace ϵ_i and the subspace ϵ_j, denoted as $\mathbf{e}^k_{\epsilon_i} = (FE^k_{\epsilon_i}, AE^k_{\epsilon_i})$ and $\mathbf{e}^k_{\epsilon_j} = (FE^k_{\epsilon_j}, AE^k_{\epsilon_j})$, where k denotes the class label. We introduce a novel inference loss term, $\mathcal{L}_{inf}$, to reduce the difference via the weighted angular similarity, which can be mathematically written as

$$\mathcal{L}_{inf}(\mathbf{e}) = \sum_{i,j=1(i\neq j)}^{M} \sum_{k=1}^{N} \theta(\mathbf{e}^k_{\epsilon_i}, \mathbf{e}^k_{\epsilon_j}) \cdot s(\mathbf{e}^k_{\epsilon_i}, \mathbf{e}^k_{\epsilon_j}), \tag{11}$$

where M denotes the total number of subspaces, and N represents the total number of categories. In addition, $\theta(\mathbf{e}^k_{\epsilon_i}, \mathbf{e}^k_{\epsilon_j})$ denotes the weight coefficients of the corresponding angle, and $s(\mathbf{e}^k_{\epsilon_i}, \mathbf{e}^k_{\epsilon_j})$ is a function to describe the angle correlation. We exploit the radial basis function (RBF) kernel to obtain the weight value between $\mathbf{e}^k_{\epsilon_i}$ and $\mathbf{e}^k_{\epsilon_j}$, which is formulated as follows:

$$\theta(\mathbf{e}^k_{\epsilon_i}, \mathbf{e}^k_{\epsilon_j}) = \exp\left(-\frac{\left\|\mathbf{e}^k_{\epsilon_i} - \mathbf{e}^k_{\epsilon_j}\right\|_2^2}{2\beta^2}\right), \tag{12}$$

where β denotes a kernel width that controls the impact of radial range. Then, the angular-based correlation calculation can be defined as:

$$s(\mathbf{e}^k_{\epsilon_i}, \mathbf{e}^k_{\epsilon_j}) = \frac{(\mathbf{e}^k_{\epsilon_i})^T \cdot \mathbf{e}^k_{\epsilon_j}}{\left\|\mathbf{e}^k_{\epsilon_i}\right\|_2 \left\|\mathbf{e}^k_{\epsilon_j}\right\|_2}. \tag{13}$$

To enable the capability of causally generating synthetic features across subspaces, we train the CVAE in a novel and causal way, denoted as CAU-CVAE. Specifically, the paired data $\{x_{\epsilon_i}, x_{\epsilon_j}\}$ belonging to the same class from different subspaces are fed into the CAU-CVAE to generate a group of reconstructed data such as $\{x'_{\epsilon_i}, x'_{\epsilon_j}, x'_{\epsilon_i\epsilon_j}, x'_{\epsilon_j\epsilon_i}\}$. The expression of the loss function is defined as:

$$\begin{aligned}\mathcal{L}_{cau-cvae}(x_\epsilon, \mathbf{e}; \psi) = {} & \mathcal{L}_{re}(x'_{\epsilon_i}, x_{\epsilon_i}) + \mathcal{L}_{re}(x'_{\epsilon_j}, x_{\epsilon_j}) \\ & + \mathcal{L}_{re}(x'_{\epsilon_i\epsilon_j}, x_{\epsilon_j}) + \mathcal{L}_{re}(x'_{\epsilon_j\epsilon_i}, x_{\epsilon_i}) \\ & + \gamma\mathcal{L}_{inf}(\mathbf{e}),\end{aligned} \tag{14}$$

where γ is a trade-off parameter to control the relative importance of causal inference. The first two terms describe the intra-domain reconstruction errors for the samples from the same subspace. The middle two terms measure the inter-domain reconstruction errors across different subspaces. The last one term calculates inference loss in the process of causal perception. Although the sample pairs $\{x'_{\epsilon_i\epsilon_j}, x_{\epsilon_j}\}$ or $\{x'_{\epsilon_j\epsilon_i}, x_{\epsilon_i}\}$ come from the same class, they do not necessarily belong to two views of the same instance. To reduce the intra-domain and inter-domain reconstruction errors, the stratified causal perception is adopted to infer and preserve the useful information in the latent representation space. As a result, the utilization of reconstruction loss $\mathcal{L}_{re}$ and inference loss $\mathcal{L}_{inf}$ facilitate the model to adaptively and reliably generate the diverse and discriminative features across domains. The adaptive data augmentation based on CAU-CVAE is summarized in Algorithm 2 in Appendix B.

D. Cycle-Consistent Minimax Optimization Mechanism

The generation method proposed above can obtain a large number of samples, but training on all of them will consume a large amount of computation and may degrade the performance due to the existence of noise samples. Here, we propose a cycle-consistent minimax optimization mechanism aimed at selecting more effective training examples from the candidate pool.

Minimizing the influence. We minimize the influence by filtering out detrimental synthetic samples to boost downstream performance. A given training sample x_i is considered harmful if the inclusion of x_i in the training set results in a larger generalization error. Therefore, we can construct the following expression by the validation loss in an approximate way:

$$\mathcal{L}(\mathcal{X}, \psi) = \frac{1}{|\mathcal{X}|} \sum_{x_i \in \mathcal{X}} l(x_i, \psi), \tag{15}$$

$$\mathcal{L}(\mathcal{X}_{val}, \hat{\psi}(\mathcal{X}_{tr} \cup \{x_i\})) - \mathcal{L}(\mathcal{X}_{val}, \hat{\psi}(\mathcal{X}_{tr})) > 0, \tag{16}$$

where $\mathcal{X}_{val}$ and $\mathcal{X}_{tr}$ denote the validation data and training data respectively.

Hence, it is necessary to retrain the model with sample x_i, which would lead to a large amount of additional runtime. While previous works have focused on removing or perturbing the existing training samples [34], we utilize the influence function to evaluate the performance after adding new synthetic samples. In this way, the change in the validation loss can be efficiently approximated by using the influence function [34]. For instance, the effect of weighting a training sample x_i with a small disturbance η on the model parameters $\hat{\psi}$ in the corresponding parameter space Ψ is described as:

$$\hat{\psi}_{\eta,x} = \arg\min_{\psi\in\Psi} \eta l(x, \psi) + \frac{1}{\sum_{i=1}^{N_t} \omega_i} \sum_{i=1}^{N_t} \omega_i l(x_i, \psi), \tag{17}$$

$$\mathcal{I}_{para}(x) := \left.\frac{d\hat{\psi}_{\eta,x}}{d\eta}\right|_{\eta=0} = -H_{\hat{\psi}}^{-1} \nabla_\psi l(x, \hat{\psi}), \tag{18}$$

where N_t is the total number of samples, ω_i is the weight for the training sample x_i, and $H_{\hat{\psi}}$ denotes the Hessian evaluated at $\hat{\psi}$. Subsequently, we utilize the chain rule to obtain the influence of upweighting the sample x_i on validation loss:

$$\mathcal{I}_{loss}(x) := \left.\frac{d\mathcal{L}(\mathcal{X}_{val}, \hat{\psi}_{\eta,x})}{d\eta}\right|_{\eta=0} = \nabla_\psi \mathcal{L}(\mathcal{X}_{val}, \hat{\psi})\mathcal{I}_{para}(x). \tag{19}$$

Note that $\mathcal{L}(\mathcal{X}_{tr}, \psi)$ can be equivalent to the weighted average form below to incorporate a new training sample x_{new}:

$$\mathcal{L}(\mathcal{X}_{tr}, \psi) = \frac{1}{\sum_{i=1}^{N_t+1} \omega_i} \sum_{i=1}^{N_t+1} \omega_i l(x_i, \psi), \tag{20}$$

where $\omega_i = 1\ (\forall i \neq N_t + 1), \omega_i = 0\ (i = N_t + 1)$ and $x_{N_t+1} = x_{new}$. When adding the training sample x_{new}, we obtain the following linear combination approximation of the validation loss change according to the influence function $\mathcal{I}_{loss}(x)$:

$$\mathcal{L}(\mathcal{X}_{val}, \hat{\psi}(\mathcal{X}_{tr} \cup \{x_{new}\})) - \mathcal{L}(\mathcal{X}_{val}, \hat{\psi}(\mathcal{X}_{tr})) \approx \frac{1}{N_t} \mathcal{I}_{loss}(x_{new}). \tag{21}$$

We efficiently calculate $\mathcal{I}_{loss}$ by minimizing the influence, where the detrimental synthetic data will have $\frac{1}{N_t}\mathcal{I}_{loss} > 0$. In this way, detrimental synthetic data, *i.e.*, the samples that produce a poor estimation on the validation losses, can be filtered out.

Maximizing the diversity. While minimizing the influence improves training data quality, it ignores the diversity that can provide a more powerful and reliable training knowledge. To measure the diversity, we calculate the sum of Mahalanobis distance between each pair of samples (x_i and x_j) for the selected datasets, which can be expressed as:

$$D_M(x_i, x_j) = \sqrt{(x_i - x_j)^T C^{-1} (x_i - x_j)}, \tag{22}$$

$$S_{div} = \sum_{r=1}^{N_r} D_{M_r}(x_i, x_j), \tag{23}$$

where C means the covariance matrix of random variables, D_{M_r} denotes the r-th Mahalanobis distance, S_{div} is the diversity measure, and N_r represents the number of combinations of different sample pairs in the extracted samples. We present a simple greedy algorithm that progressively and iteratively selects training samples from the candidate pool to maximize the diversity. Ultimately, the candidate sample set that maximizes the diversity measure is selected.

The optimized candidate pool is mainly employed for the stable model training, which provides high quality and diversity data from original and synthetic samples to capture the characteristics of distributions across different domains. The candidate pool is a fundamental component in the cycle-consistent minimax optimization mechanism, which provides the flexibility to reduce the distribution discrepancy between the source (*i.e.*, seen available domain) and target (*i.e.*, unseen testing domain) data in knowledge transfer tasks. During the training iterations, a random batch of samples is drawn from the candidate pool. In this way, the update of model parameters contributes to the reduction of distribution discrepancy between source and target domains, which motivates the model to effectively generalize on unseen target domains through the learning of stable and transferable representations. In the practical implementation, the required parameters are first determined based on the knowledge transfer tasks, and then the candidate pool is optimized to realize the desired results.

TABLE I
THE NUMBER OF DIFFERENT CATEGORIES IN THE IMBALANCED CIFAR-10 (IM-C) AND IMBALANCED STL-10 (IM-S) DATASETS

Datasets	0	1	2	3	4	5	6	7	8	Total
Im-C	800	1200	1000	800	700	1200	1000	1200	1100	9000
Im-S	900	700	1200	900	800	1100	1100	1300	1000	9000

The cycle-consistent minimax optimization through considering both the influence minimization and the diversity maximization is described in Algorithm 3 in Appendix B, which have complementary benefits. Specifically, the former is designed to improve the quality of individual samples by eliminating detrimental ones, while the latter aims to construct a diverse training set without paying attention to the quality of data. To reap both benefits, our proposed cycle-consistent minimax optimization mechanism utilizes a hybrid selection technique that first eliminates the detrimental data by minimizing the influence, and then selects diverse samples by maximizing the diversity. Later, in the evaluation, the ablation studies of the proposed SCDA are shown in Section VI-E, which demonstrate the superiority of cycle-consistent minimax optimization mechanism in capturing complementary information and verify the effectiveness of our model for secure and robust knowledge transfer.

VI. Evaluation

A. Datasets and Model Selection

Rotated MNIST. A modified version of MNIST [35] is generated by rotating the original grayscale MNIST digits from 0° to 90° with the interval of 15°, which is denoted as Rotated MNIST. Each rotation angle is treated as a domain, and our task is to complete the prediction of the class label. Since the images of different domains are obtained from the same basic images, there may exist diverse causal matches across domains. Following the setting of [36], the samples with 0° and 90° are used as the testing domain that are extremely difficult to detect, and the rest are adopted as training domains.

CIFAR-10 & STL-10. CIFAR-10 and STL-10 are both the 10-class image datasets, which consist of 60,000 32×32 pixel images and 13,000 96×96 pixel images, respectively. The two datasets are similar but one class is different. Therefore, we select the shared nine classes in the experiments. In the data preprocessing stage, we resize the images in STL-10 to 32×32 pixels. We revise the number of different classes of images in CIFAR-10 and STL-10 to construct two imbalanced datasets, namely imbalanced CIFAR-10 (Im-C), and imbalanced STL-10 (Im-S) which have different class distributions. The details are given in Table I.

Colored PACS. PACS [37] is a public generalization performance evaluation dataset with the remarkable distribution shift across different domains. It includes seven classes (dog, elephant, house, giraffe, horse, guitar, person) and four domains, such as, Photo (P), Art paintings (A), Cartoon (C) and Sketch (S). We drew inspiration from [38] to construct the biased dataset using color differences in the background. Primarily, seven classes of objects are extracted from PACS dataset. For the in-domain, we put onto seven different kinds of color backgrounds and define a one-to-one background-color relationship with seven categories (*e.g.*, dog↔red, elephant↔blue). For the out-domain, the color filling is performed randomly to evaluate the generalization performance influenced by the background factors.

Model Selection. While using validation sets from the testing domain may increase classification accuracy, it violates the motivation of generalization to unseen domains. Therefore, we only use data from the source domain to form validation sets, unless explicitly mentioned in Table V to explore how model selection affects the results.

B. *Implementation Details*

Our robust generalization network is implemented based on a cascaded stratified architecture. Prior to density ratio estimation, a pre-trained flow g_τ is leveraged as an invertible encoder to transform the inputs into a discriminative feature space. An invertible classifier c_ϕ is built by modifying the architecture to incorporate that of the flow g_τ, which is trained merely through the BCE loss $\mathcal{L}_{sup}(\tau, \phi)$. We then train an invertible deep generative model based on i-RevNet [33] in a mixture of samples to map feature distributions and obtain uncertainty scores. Following [39], we employ the ResNet-50 [40] features to perceive causal and non-causal variables. Our data generation module is performed by exploiting the networks in [41]. Since it does not take causal and non-causal attributes into account, here we modify the encoder to construct them. Our networks are trained for 50 epochs with early stopping rules using the Adam [42] optimizer with a momentum of 0.9 and a decay of 0.0001. The initial learning rate is set as 0.001 in Rotated MNIST, 0.005 in CIFAR-10, and 0.0001 in Colored PACS.

C. *Compared Methods*

In intelligent collaborative scenarios, we test the performance of the proposed scheme, including the security of data privacy and model copyright, and the robustness against distribution shift attacks [43]. Specifically, our scheme addresses the new challenges of knowledge transfer in severe distribution shifts where the interference-free shift and the interference-active shift coexist, namely $\mathcal{X}^{if}_{\epsilon_A} \neq \mathcal{X}^{if}_{\epsilon_T}$ and $\mathcal{X}^{ia}_{\epsilon_A} \neq \mathcal{X}^{ia}_{\epsilon_T}$. We compare our method with domain generalization or domain adaptation methods that mainly focus on the interference-free shift assumption (*i.e.*, $\mathcal{X}^{if}_{\epsilon_A} \neq \mathcal{X}^{if}_{\epsilon_T}$ and $\mathcal{X}^{ia}_{\epsilon_A} = \mathcal{X}^{ia}_{\epsilon_T}$): Meta-Learning Domain Generalization (MLDG) [4], Universal Adaptation Network (UAN) [5], Common Specific Decomposition (CSD) [35], and Federated Simple Representation (FedSR) [44]. Our scheme is also compared with OOD generalization methods that pay more attention to the influence of distribution shift caused by interference-active factors (*i.e.*, $\mathcal{X}^{if}_{\epsilon_A} = \mathcal{X}^{if}_{\epsilon_T}$ and $\mathcal{X}^{ia}_{\epsilon_A} \neq \mathcal{X}^{ia}_{\epsilon_T}$): Invariant Risk Minimization (IRM) [9], Targeted Data-driven Regularization (TDR) [8], and Risk Extrapolation (REx) [15]. For all the compared methods, we adopt ResNet-50 as the backbone network which is pre-trained on the ImageNet dataset [45]. Note that MLDG, UAN, and CSD require reference information from target domains as input, and we give some target knowledge as input to guide the model training. In the testing phase, we input samples to be tested into the corresponding model and take the average results to determine which categories of test samples belong to.

TABLE II
CLASSIFICATION ACCURACY (%) ON ROTATED MNIST DATASET USING TARGET DOMAINS OF 0° AND 90°

Source	MLDG	UAN	IRM	CSD	FedSR	TDR	REx	SCDA
15°, 30°, 45°, 60°, 75°	92.6	92.1	93.3	94.5	93.6	92.7	93.9	**95.3**
30°, 45°, 60°, 75°	80.3	84.2	85.6	87.9	85.1	85.3	87.1	**90.7**
15°, 45°, 75°	70.4	65.3	73.1	75.6	73.9	74.8	77.2	**80.4**
30°, 60°	60.5	58.1	62.7	61.3	66.4	63.9	65.2	**68.5**
Average	76.0	74.9	81.2	79.8	79.8	79.2	80.9	**83.7**

TABLE III
CLASSIFICATION ACCURACY (%) ON ROTATED MNIST DATASET OF DIFFERENT THRESHOLD SELECTION USING TARGET DATASETS OF 0° AND 90°. (α_1, α_2, α_3) REPRESENTS THREE DIFFERENT THRESHOLDS

Source	Threshold settings (α_1, α_2, α_3)			Average
	(0.3,0.6,1)	(0.5,0.8,1.1)	(0.6,0.9,1.2)	
15°, 30°, 45°, 60°, 75°	93.4	**95.6**	94.7	94.6
30°, 45°, 60°, 75°	86.5	**89.7**	87.6	87.9
15°, 45°, 75°	76.5	**81.3**	77.1	78.3
30°, 60°	64.1	**67.8**	63.7	65.2

D. *Experimental Results*

Results on the Rotated MNIST. Table II gives classification accuracy on Rotated MNIST using testing domains with 0° and 90° rotation. From the results, it is obvious that our proposed SCDA outperforms other methods. With the reduction of the number of training domains, the accuracy gradually decreases and the gap between SCDA and baselines becomes significant. In three source domain experiments, SCDA achieves an accuracy of 80.4% while the second best REx achieves 77.2%. It is worth highlighting that the adaptive distribution adjustment through stratified-causality data augmentation can avoid both the data privacy leakage caused by sample updating and the model copyright infringement caused by the organization deploying under intelligent collaborative scenarios. In addition, we observe that the improvement of knowledge transfer ability benefits from the space division based on uncertainty scores. In the following experiments, we explore the effect of space division generated by different threshold selection on the performance improvement. We explore the impact of space division on performance using different threshold settings, as shown in Table III. It can be seen from the experimental results that different threshold settings have an impact on performance. Appropriate threshold selection is conducive to the improvement of knowledge transfer capability. The reason is that the samples with similar uncertainty scores in the same data space are more effective to causal perception, which can make the changes among samples progressive and continuous.

Results on the CIFAR-10 & STL-10. As shown in Table IV, our proposed SCDA achieves the best accuracy in all tasks. Compared with the domain generalization methods, CSD and FedSR, the performance of SCDA has improved obviously. The average accuracy of SCDA is 12.6% higher than that of CSD. Among all compared methods, REx obtains the best

TABLE IV
CLASSIFICATION ACCURACY (%) ON CIFAR AND STL DATASETS WITH BALANCED AND IMBALANCED DISTRIBUTIONS COMPARED WITH BASELINE METHODS

Tasks	MLDG	UAN	IRM	CSD	FedSR	TDR	REx	SCDA
CIFAR→ STL	56.9	59.5	60.2	63.4	64.8	60.9	67.5	**72.4**
STL→ CIFAR	59.7	57.1	64.8	65.2	61.9	63.2	71.6	**76.3**
Im-C→Im-S	56.1	53.8	53.1	56.3	55.4	57.3	60.7	**69.7**
Im-S→Im-C	55.2	52.4	54.6	54.7	52.0	58.8	62.1	**71.6**
Average	57.0	55.7	58.2	59.9	58.5	60.1	65.5	**72.5**

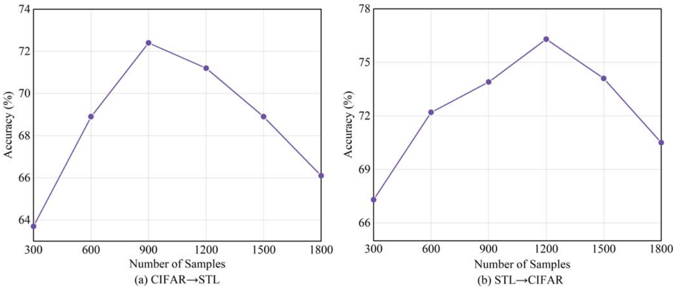


Fig. 6. Influence of candidate pool size on model performance under CIFAR→STL and STL→CIFAR tasks.

performance, while the average accuracy of SCDA is 7% higher than REx. For imbalanced datasets, corresponding results show that SCDA can still significantly enhance the performance of knowledge transfer in the crafted biased settings, where other related methods bring very limited benefit. The reasons can be attributed to the following two aspects. On the one hand, the minority class instances can obtain more complementary information from the majority class instances to boost the generation of new instances by partitioning the data space with uncertainty scores. On the other hand, the causal perception in a stratified manner facilitates the capture of imbalanced distribution knowledge and is more stable for imbalanced datasets.

In addition, we conduct experiments to analyze the impact of candidate pool size on the model performance. In fact, the optimal candidate pool is determined based on the cycle-consistent minimax optimization. We aim to investigate the impact of candidate pool size on performance by deleting or adding samples during experiments. Then, the size of the candidate pool is measured by the number of samples in the candidate pool. Specifically, based on the optimal candidate pool, we reduce the size of the candidate pool by randomly deleting samples, and expand the size of the candidate pool by adding unselected remaining samples from the original and synthetic datasets. Experiments are carried out on CIFAR→STL and STL→CIFAR tasks, and the results are shown in Fig. 6. From the experimental results, it can be observed that the optimal candidate pool with an appropriate size through cycle-consistent minimax optimization can force the model to achieve superior performance. The reason for the degradation of model performance is that the small-size candidate pools generally lack diversity, while large-size candidate pools with many unscreened samples may lead to poor data quality.

Results on the Colored PACS. These color backgrounds make this task more complex and biased, and therefore more difficult than the previous task. As shown in Table V, our proposed SCDA on the Colored PACS dataset is competitive to state-of-the-art results averaged over all domains. The SCDA (val_s) shows the highest performance across different domains, except compared to CSD in the "C" domain. In addition, the average accuracy of CSD is similar to that of SCDA in this article. The reason is that the common component and the domain specific component are jointly learnt in a decomposed way, which may promote positive transfer and combat negative transfer to some extent. We can also observe that the models gradually perform better when the test or mixed domain validation is used as a model selection strategy. Therefore, in some scenarios where a little test domain data is available, the model selection strategy can be considered as an appropriate alternative solution for the improvement of knowledge transfer performance.

TABLE V
CLASSIFICATION ACCURACY (%) ON COLORED PACS DATASET COMPARISON WITH THE STATE-OF-THE-ART METHODS. THE P, A, C AND S IN THE FIRST ROW RESPECTIVELY INDICATE THE TARGET DOMAIN WITH THE REMAINING THREE DOMAINS USED AS THE SOURCE DOMAINS. THE val_s, val_t AND val_m DENOTE THREE MODEL SELECTION STRATEGIES, NAMELY USING SOURCE DATA, TARGET DATA AND MIXED SOURCE-TARGET DATA AS VALIDATION SETS TO OBTAIN EXPERIMENTAL RESULTS

Methods	P	A	C	S	Average
MLDG	91.4	75.3	74.6	71.8	78.3
UAN	92.3	76.9	75.1	72.3	79.2
IRM	91.8	76.7	73.9	73.5	79.0
CSD	93.7	78.4	**77.5**	76.1	81.4
FedSR	92.0	78.8	76.2	75.7	80.7
TDR	94.2	77.5	76.4	75.3	80.9
REx	93.1	77.0	75.8	74.6	80.1
SCDA (val_s)	**94.6**	**79.2**	77.1	**76.4**	**81.8**
SCDA (val_t)	**95.8**	81.3	78.6	**78.3**	83.5
SCDA (val_m)	95.3	**82.9**	**80.3**	77.9	**84.1**

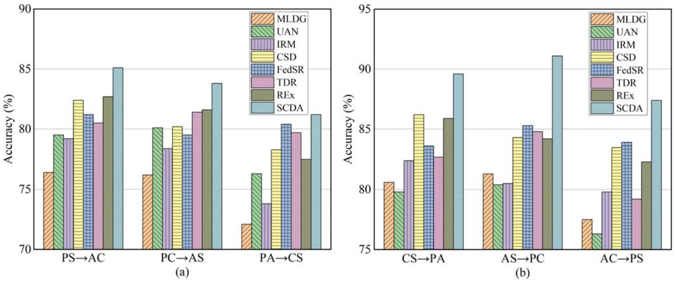


Fig. 7. Comparison of state-of-the-art methods on the Colored PACS dataset under the condition of double source and target domains.

To further analyze the performances of the proposed SCDA, we conduct a series of experiments on the Colored PACS dataset under the condition of double source and target domains. As can be seen from Fig. 7, the performance of MLDG is relatively poor, so the model agnostic training strategy is susceptible to the interference-active shifts. We can also observe that our SCDA is superior to other state-of-the-art approaches in each case, which further demonstrates the effectiveness of our proposed method in the Colored PACS dataset. Moreover, Fig. 7(b) exhibits better performance than Fig. 7(a). The reason may be that the interference-free shift of "P" domain is small

TABLE VI
RESULTS OF ABLATION EXPERIMENTS ON THREE REPRESENTATIVE TASKS BY REMOVING ONE COMPONENT WHILE FIXING THE OTHERS

Method	Task1	Task2	Task3	Average
w/o uncertainty quantification	57.1	62.7	69.4	63.1
w/o causal perception	63.7	64.8	69.3	65.9
w/o influence filter	62.8	63.0	73.5	66.4
w/o diversity selection	65.2	67.4	74.6	69.1
w/o hybrid optimization	59.4	62.3	70.3	64.0
SCDA	**68.5**	**71.6**	**79.2**	**73.1**

and hence easy to detect, while the "S" domain is relatively difficult to detect due to the large domain bias between the seen and unseen classes.

E. Ablation Studies

To further understand the effect of each component in our model, we conduct a set of ablation experiments on Rotated MNIST, CIFAR-10 & STL-10, and Colored PACS datasets. Due to space limitation, Table VI shows the results of ablation analysis for three representative tasks (Task1: 30° & 60° →0° & 90°, Task2: Im-S→Im-C, and Task3: P & C & S→A). In our experiments, the impact of different components can be estimated by controlling major constraint terms, including the effect of without the uncertainty partitioning ("w/o uncertainty quantification"), the effect of without the causality inference ("w/o causal perception"), without the influence minimization ("w/o influence filter"), without the diversity maximization ("w/o diversity selection"), and without the minimax optimization ("w/o hybrid optimization").

From Table VI, we can make the following observations. First, when removing the uncertainty partitioning ("w/o uncertainty quantification"), the classification accuracies dramatically degrade due to the lack of goal orientation, which demonstrates the effectiveness of data stratification based on density ratio estimation for secure and robust knowledge transfer. Second, our scheme achieves 4.8%, 6.8% and 9.9% gains over "w/o causal perception" on the 30° & 60° →0° & 90°, Im-S→Im-C and P & C & S→A tasks, respectively. This clearly validates that both the basic features and state features should be causally inferred to enhance the performance and alleviate the interference-free and interference-active shifts. Third, our scheme outperforms "w/o influence filter" and "w/o diversity selection", showing that it is beneficial to narrow distribution discrepancy across domains in the candidate pool. Finally, "w/o hybrid optimization" works significantly worse than our scheme, which clearly implies the superiority of mining the complementary information through the influence minimization and the diversity maximization.

F. Causality Analysis

To quantitatively analyze the effectiveness of causal reasoning, the effects of interference-free and interference-active shifts are quantified by normalization of the predicted results. We first report the classification results of "w/o causal perception" and our SCDA for each class on the Im-S→Im-C task in Table VII, and then illustrate the effects of the two shifts in Fig. 7.

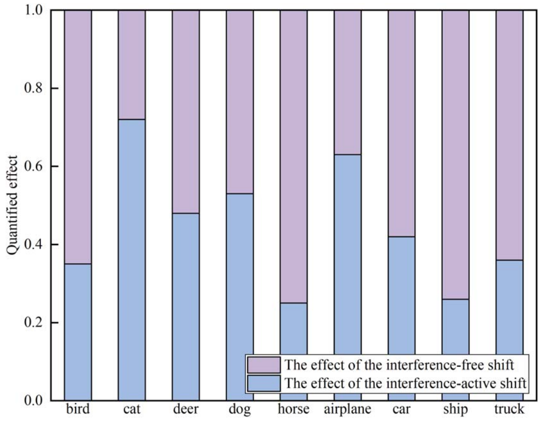


Fig. 8. Quantified effects of the interference-free and interference-active shifts on the Im-S→Im-C task. The purple and blue bars represent the effects of the interference-free and interference-active shifts, respectively. The horizontal axis denotes different categories and the vertical axis is the quantified value of the effect.

From the results in Table VII, it is interesting to find that the proposed SCDA outperforms "w/o causal perception" not only in animal classes (*i.e.*, 5.4% gains on "bird" and 8.9% gains on "cat") but also on vehicle classes (*i.e.*, 7.5% gains on "airplane" and 6.4% gains on "car"). The promising results indicate the effectiveness of stratified causality perception on exploring the transferable causal knowledge for secure and robust knowledge transfer.

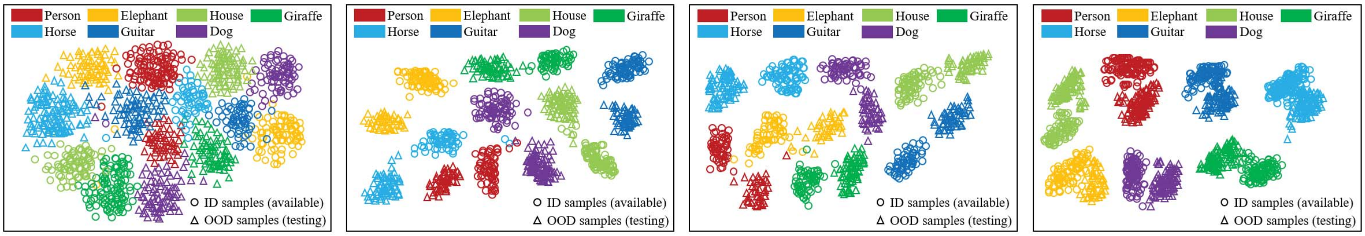
In Fig. 8, the higher quantified effect of the interference-free shift (purple bars) means that the interference-free shift is more significant to secure and robust generalization, and the larger quantified effect of the interference-active shift (blue bars) indicates that the interference-active shift is more crucial. From the results, we find that the effects inferred by causal perception accurately reflect the importance of different shifts. For example, "horse" has a higher quantified effect of the interference-free shift than "cat" since "horse" contains more basic representations than "cat" in this imbalanced task, and hence the interference-free shift in "horse" deserves much more attention than that in "cat". This shows that stratified causal perception can infer the contribution or influence of the basic and state features to mitigate the interference-free and interference-active shifts for the improvement of adaptation and generalization through adaptive distribution adjustment.

G. Feature Visualization

To further illustrate the effectiveness of the proposed SCDA scheme, we visualize the data distribution of original features ("Original"), the learned features without uncertainty partitioning ("w/o uncertainty quantification"), the learned features without causality inference ("w/o causal perception") and the learned features with our scheme ("SCDA") in the same feature space on the A→P task, as presented in Fig. 9(a)–(d), respectively. For clarity, we visualize two domains (different shapes) and seven classes (different colors) by the t-distributed stochastic neighbor embedding (t-SNE).

TABLE VII
CLASSIFICATION ACCURACY (%) OF "W/O CAUSAL PERCEPTION" AND OUR SCDA FOR EACH CLASS ON THE Im-S→Im-C TASK UNDER INTELLIGENT COLLABORATIVE SCENARIOS

Method	Bird	Cat	Deer	Dog	Horse	Airplane	Car	Ship	Truck	Average
w/o causal perception	61.6	59.5	61.6	62.7	49.6	73.2	77.1	67.2	70.9	64.8
SCDA	67.0	68.4	63.9	75.6	56.3	80.7	83.5	70.6	78.4	71.6



(a) Original (b) w/o uncertainty quantification (c) w/o causal perception (d) SCDA

Fig. 9. Feature visualization on the A→P task. The circles and triangles denote the ID features and OOD features, respectively.

We draw several interesting observations and insights arising from Fig. 9. First, there is a large distribution shift between the "A" and "P" domains as shown in Fig. 9(a), and even some available and testing features of the same class belong to different clusters. Second, as presented in Fig. 9(b) and 9(c), our scheme generalizes the learned in-distribution (ID) features to the out-of-distribution (OOD) features better than "w/o uncertainty quantification" and "w/o causal perception" owing to the realization of uncertainty stratification and causality inference. Finally, in Fig. 9(d), the learned ID features in the diverse classes are aligned and centralized with the OOD features, clearly showing that our scheme can successfully bridge the gap between available and testing samples to guarantee the security and robustness of knowledge transfer in intelligent collaborative services.

VII. CONCLUSION

In this article, we present a secure and robust knowledge transfer framework for intelligent collaborative services through stratified-causality distribution adjustment. The following conclusions can be drawn from this research work: 1) an innovative uncertainty quantification via density ratio estimation can mine the distribution rules of data points during the preprocessing, which provides a basis for determining the distribution areas that need to be focused on; 2) the causal perception in a stratified manner is proposed to infer the effects of interference-free and interference-active shifts, which sufficiently ensures that the synthetic causality-guided features can adaptively bridge gaps in data distribution to promote the security and robustness of knowledge transfer under intelligent collaborative scenarios; 3) this article brings a new perspective that the cycle-consistent minimax optimization can be exploited to further alleviate the false alignment across different domains via minimizing the influence and maximizing the diversity; and 4) extensive experimental results show that our proposed scheme is more appropriate for practical and complicated knowledge transfer tasks through goal-directed knowledge delivery, and can simultaneously mitigate the threat of negative transfer under the condition of severe distribution shifts. Moreover, data privacy leakage and model copyright infringement can be avoided by adaptive distribution adjustment. In the near future, we are planning to investigate the uncertainty quantification criterion that can accurately assist to explore inherent causal relationships for the implementation of flexible and efficient knowledge transfer in device-edge-cloud collaborative services.

ACKNOWLEDGMENT

Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not reflect the views of National Research Foundation, Singapore and Cyber Security Agency of Singapore.

REFERENCES

[1] T. Hao, K. Hwang, J. Zhan, Y. Li, and Y. Cao, "Scenario-based AI benchmark evaluation of distributed cloud/edge computing systems," *IEEE Trans. Comput.*, vol. 72, no. 3, pp. 719–731, Mar. 2023.

[2] D. He, N. Kumar, M. K. Khan, L. Wang, and J. Shen, "Efficient privacy-aware authentication scheme for mobile cloud computing services," *IEEE Syst. J.*, vol. 12, no. 2, pp. 1621–1631, Jun. 2018.

[3] Y. Bai, L. Chen, S. Ren, and J. Xu, "Automated customization of on-device inference for quality-of-experience enhancement," *IEEE Trans. Comput.*, vol. 72, no. 5, pp. 1329–1342, May 2023.

[4] D. Li, Y. Yang, Y. Song, and T. M. Hospedales, "Learning to generalize: Meta-learning for domain generalization," in *Proc. 32th AAAI Conf. Artif. Intell.*, 2018, pp. 3490–3497.

[5] K. You, M. Long, Z. Cao, J. Wang, and M. I. Jordan, "Universal domain adaptation," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 2720–2729.

[6] J. Jia, L. Zhai, W. Ren, L. Wang, Y. Ren, and L. Zhang, "Transferable heterogeneous feature subspace learning for JPEG mismatched steganalysis," *Pattern Recognit.*, vol. 100, May 2020, Art. no. 107105.

[7] J. Jia, M. Luo, S. Ma, and L. Wang, "Partial knowledge transfer in visual recognition systems via joint loss-aware consistency learning," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 7463–7474, 2022.

[8] M. M. Kamani, S. Farhang, M. Mahdavi, and J. Z. Wang, "Targeted data-driven regularization for out-of-distribution generalization," in *Proc. 26th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, 2020, pp. 882–891.

[9] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," 2019, *arXiv:1907.02893*.

[10] J. Jia, M. Luo, J. Liu, W. Ren, and L. Wang, "Multiperspective progressive structure adaptation for JPEG steganography detection across

domains," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3660–3674, Aug. 2022.

[11] M. Koyama and S. Yamaguchi, "Out-of-distribution generalization with maximal invariant predictor," 2020, *arXiv:2008.01883*.

[12] X. Hu et al., "A systematic view of model leakage risks in deep neural network systems," *IEEE Trans. Comput.*, vol. 71, no. 12, pp. 3254–3267, Dec. 2022.

[13] J. Jia, Y. Wu, A. Li, S. Ma, and Y. Liu, "Subnetwork-lossless robust watermarking for hostile theft attacks in deep transfer learning models," *IEEE Trans. Dependable Secure Comput.*, to be published, doi: 10.1109/TDSC.2022.3194704.

[14] P. Li, X. Rao, J. Blase, Y. Zhang, X. Chu, and C. Zhang, "CleanML: A study for evaluating the impact of data cleaning on ML classification tasks," in *Proc. 37th IEEE Int. Conf. Data Eng.*, 2021, pp. 13–24.

[15] D. Krueger et al., "Out-of-distribution generalization via risk extrapolation (REX)," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, vol. 139, pp. 5815–5826.

[16] J. Jia, M. Luo, S. Ma, L. Wang, and Y. Liu, "Consensus-clustering-based automatic distribution matching for cross-domain image steganalysis," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 6, pp. 5665–5679, Jun. 2023.

[17] S. Huang, C. Lin, S. Chen, Y. Wu, P. Hsu, and S. Lai, "AugGAN: Cross domain adaptation with GAN-based data augmentation," in *Proc. Eur. Conf. Comput. Vision*, 2018, vol. 11213, pp. 731–744.

[18] W. Hsu, Y. Zhang, and J. R. Glass, "Unsupervised domain adaptation for robust speech recognition via variational autoencoder-based data augmentation," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, 2017, pp. 16–23.

[19] N. Ng, K. Cho, and M. Ghassemi, "SSMBA: Self-supervised manifold based data augmentation for improving out-of-domain robustness," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2020, pp. 1268–1283.

[20] T. VanderWeele, *Explanation in Causal Inference: Methods for Mediation and Interaction*. London, U.K.: Oxford Univ. Press, 2015.

[21] S. Disabato, M. Roveri, and C. Alippi, "Distributed deep convolutional neural networks for the Internet-of-Things," *IEEE Trans. Comput.*, vol. 70, no. 8, pp. 1239–1252, Aug. 2021.

[22] Z. Yue, T. Wang, Q. Sun, X. Hua, and H. Zhang, "Counterfactual zero-shot and open-set visual recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 15404–15414.

[23] W. Wang, F. Feng, X. He, H. Zhang, and T. Chua, "Clicks can be cheating: Counterfactual recommendation for mitigating clickbait issue," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1288–1297.

[24] T. Wang, J. Huang, H. Zhang, and Q. Sun, "Visual commonsense R-CNN," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 10757–10767.

[25] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE Conf. Comput. Vision and Pattern Recognit.*, 2020, pp. 3713–3722.

[26] H. Huang and X. Liu, "Local migration model of images based on deep learning against adversarial attacks," *IEEE Trans. Comput.*, to be published, doi: 10.1109/TC.2021.3075715.

[27] H. Bahng, S. Chun, S. Yun, J. Choo, and S. J. Oh, "Learning de-biased representations with biased representations," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 528–539.

[28] C. Chen, H. Wu, J. Su, L. Lyu, X. Zheng, and L. Wang, "Differential private knowledge transfer for privacy-preserving cross-domain recommendation," in *Proc. ACM Web Conf.*, 2022, pp. 1455–1465.

[29] X. Liao, W. Liu, X. Zheng, B. Yao, and C. Chen, "PPGenCDR: A stable and robust framework for privacy-preserving cross-domain recommendation," in Proc. *37th AAAI Conf. Artif. Intell.*, 2023, pp. 4453–4461.

[30] D. P. Kingma and P. Dhariwal, "Glow: Generative flow with invertible 1x1 convolutions," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10236–10245.

[31] J. Pearl, M. Glymour, and N. P. Jewell, *Causal Inference in Statistics: A Primer*. Hoboken, NJ, USA: Wiley, 2016.

[32] Y. Rao, G. Chen, J. Lu, and J. Zhou, "Counterfactual attention learning for fine-grained visual categorization and re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2021, pp. 1025–1034.

[33] J. Jacobsen, A. W. M. Smeulders, and E. Oyallon, "i-RevNet: Deep invertible networks," in *Proc. 6th Int. Conf. Learn. Representations*, 2018, pp. 1–11.

[34] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 1885–1894.

[35] V. Piratla, P. Netrapalli, and S. Sarawagi, "Efficient domain generalization via common-specific low-rank decomposition," in *Proc. 37th Int. Conf. Mach. Learn.*, 2020, vol. 119, pp. 7728–7738.

[36] D. Mahajan, S. Tople, and A. Sharma, "Domain generalization using causal matching," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, vol. 139, pp. 7313–7324.

[37] D. Li, Y. Yang, Y. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 5543–5551.

[38] F. Ahmed, Y. Bengio, H. van Seijen, and A. C. Courville, "Systematic generalisation with group invariant predictions," in *Proc. 9th Int. Conf. Learn. Representations*, 2021, pp. 1–19.

[39] J. Li, M. Jing, K. Lu, Z. Ding, L. Zhu, and Z. Huang, "Leveraging the invariant side of generative zero-shot learning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 7402–7411.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

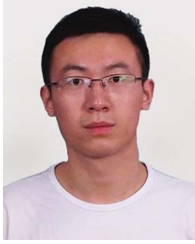
[41] X. Sun, Z. Yang, C. Zhang, K. V. Ling, and G. Peng, "Conditional Gaussian distribution learning for open set recognition," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 13477–13486.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–15.

[43] L. Cui, Z. Chen, S. Yang, R. Chen, and Z. Ming, "A secure and decentralized DLaaS platform for edge resource scheduling against adversarial attacks," *IEEE Trans. Comput.*, to be published, doi: 10.1109/TC.2021.3074806.

[44] A. T. Nguyen, P. Torr, and S. N. Lim, "FEDSR: A simple and effective domain generalization method for federated learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 38831–38843.

[45] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern*, 2009, pp. 248–255.



Ju Jia received the Ph.D. degree in cyberspace security from Wuhan University, Wuhan, China, in 2021. He was a Research Fellow with the School of Computer Science and Engineering at Nanyang Technological University, Singapore. He is currently a Research Associate Professor with the School of Cyber Science and Engineering, Southeast University, Nanjing, China. His research interests include data security, intelligent collaborative security, multimedia content security, and artificial intelligence security.

Siqi Ma (Member, IEEE) received the B.S. degree in computer science from Xidian University, Xi'an, China, in 2013, and the Ph.D. degree in information system from Singapore Management University, in 2018. She was a Research Fellow with Distinguished System Security group at CSIRO, and then was a Lecturer with the University of Queensland. She is currently a Senior Lecturer with the University of New South Wales, Canberra, ACT, Australia. Her research interests include data security, IoT security, and software security.

Lina Wang (Member, IEEE) received the B.S. degree from Hefei University of Technology, Hefei, China, in 1986, and the M.S. and Ph.D. degrees from Northeastern University, Shenyang, China, in 1989 and 2001, respectively. She is currently a Professor with the Cyber Science and Engineering School at Wuhan University. Her research interests include multimedia content security, cloud security, data security, and machine learning methods in network security detection.

Yang Liu (Senior Member, IEEE) received the B.Comp. degree (Hons.) from the National University of Singapore (NUS), in 2005, and the Ph.D. degree from NUS and MIT, in 2010. He started his Postdoctoral work with NUS and MIT. In 2012, he joined Nanyang Technological University (NTU). He is currently a Full Professor and the Director of the Cybersecurity Laboratory, NTU. He specializes in software verification, security, and software engineering. His research has bridged the gap between the theory and practical usage of formal methods and program analysis to evaluate the design and implementation of software for high assurance and security. By now, he has more than 400 publications in top tier conferences and journals. He received a number of prestigious awards, including MSRA Fellowship, TRF Fellowship, Tan Chin Tuan Fellowship, Nanyang Research Award 2019, ACM Distinguished Speaker, NRF Investigatorship, and 15 best paper awards.

Robert H. Deng (Fellow, IEEE) is currently an AXA Chair Professor in cybersecurity, the Director of the Secure Mobile Centre, and the Deputy Dean of the Faculty & Research, School of Information Systems, Singapore Management University (SMU). His research interests include data security and privacy, network security, and system security. He is a fellow of Academy of Engineering Singapore. He received the Outstanding University Researcher Award from the National University of Singapore, Lee Kuan Yew Fellowship for Research Excellence from SMU, and AsiaCPacific Information Security Leadership Achievements Community Service Star from International Information Systems Security Certification Consortium. He serves/served on many editorial boards and conference committees, including the editorial boards of ACM TRANSACTIONS ON PRIVACY AND SECURITY, the IEEE SECURITY AND PRIVACY, the IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY.