

Topology Change Aware Distributed State Estimation Based on Unsupervised Bipartite Graph-enabled Causality-inspired Sparse Learning

Abstract—Topology changes in a distribution network are common due to planned reconfigurations and unintentional switching events during practical operations. Once topology changes occur, it is difficult for existing optimization-based and learning-based distribution system state estimation (DSSE) methods to perform state estimation due to the lack of accurate structural information of the new topology and the labeled data (recorded state variables) used for training. To this end, this paper proposes an **unsupervised-on-target** learning-based state estimation method for the distribution network after topology changes without relying on the topology information and labeled data. In particular, a bipartite graph learning (BGL) method with rank constraints is first designed to learn the representation of each topology with a restricted set of measurements. Then, the Euclidean distance is employed to select the best-matched source domain historical topology according to the representation learned by the BGL. To extract invariant causal structures across the two topologies, a **causality-inspired sparse structure learning** for domain adaptation network is further designed. **It relaxes the correlations between the selected historical and new topologies into an associative structure, represented by attention scores derived from the proposed inter- and intra-variable attention networks.** This allows the leverage of the causality to enhance the state estimation performance of the distribution network after topology changes without relying on accurate topology information and recorded labels used for training. The comparison results on two standard IEEE test systems validate the efficacy of the proposed method.

Index Terms—unsupervised state estimation, distribution system state estimation, domain adaption, bipartite graph, attention mechanism.

I. INTRODUCTION

The increasing penetration of renewable power generation poses huge challenges for the safe operation of distribution systems. Distribution System State Estimation (DSSE) aims to provide estimates of system state according to the real-time measurements and pseudo-measurements [1], [2]. It plays a critical role in enhancing the situational awareness of the distribution network, which is beneficial for the decision-making of system operators. However, the challenges in DSSE arise from factors such as inaccurate physical models, lack of real-time measurements, and frequent topology changes [3], [4].

Current approaches for DSSE tasks can generally be categorized into two main types. The first type of methods is optimization-based strategies. The Weighted Least Squares (WLS) method is a well-established optimization technique commonly used in DSSE and has been extensively researched in the literature [5]. [6] presents a novel convexification framework to reformulate the WLS framework into a semi-definite program through convex relaxation techniques. A graph computing-enabled algorithm based on the WLS method is proposed in [7]. Considering the asynchrony of the measurements, [8] proposes a proximal-point type model that is conducted in an online fashion. To alleviate the impact of inaccurate topology, [9] proposes a robust state estimation that can deal with the topology error. [10], [11] design a compressive sensing framework to jointly tackle the state estimation and the inaccurate topology. In addition, robust sequential estimation approaches such as the Ensemble

Kalman Filter have been adapted to distribution systems to improve noise resilience and uncertainty quantification [12]. Although the optimization-based methods can achieve satisfying performance, they rely on either precise line parameters or accurate structural information of the distribution network, which are often challenging to obtain in practical scenarios. In addition, the optimization-based DSSE methods suffer from a high calculation burden when dealing with large-scale distribution systems [13]. Comprehensive reviews have also summarized these optimization-based and sequential estimation techniques, discussing their strengths and limitations in the context of smart distribution networks [14]. The above reasons bring challenges to the implementation of optimization-based methods in practice.

The second category consists of learning-based methods. Compared with optimization-based ones, such methods can uncover underlying patterns and correlations between observed measurements and state variables by analyzing historical time-series data [15]. Therefore, accurate physical information is not required. In addition, since the sophisticated optimization task is reduced to matrix-vector multiplications once the training process is completed, the computational efficiency can be significantly improved. The above advantages make the learning-based methods a promising solution for the online DSSE task [16], [17]. However, distribution systems with embedded microgrids pose additional challenges for learning-based DSSE because of bidirectional power flows and greater variability, motivating specialized estimation strategies [18]. Moreover, recent reviews have highlighted the growing role of deep learning, transfer learning, and ensemble methods in real-time DSSE applications, while also pointing out the need for large labeled datasets and challenges in adapting to topology changes [19]. Therefore, the effectiveness of the learning-based methods is achieved on the basis of two assumptions, namely sufficient training samples and fixed system topology. In practice, a distribution network may undergo regular topology alterations owing to planned reconfiguration and unintended switching actions [20]. When such events occur, the existing learning-based methods necessitate re-training to adapt to the topology change [21]. During the retraining process, the labels generated according to the new topology, which refer to the state variables in the context of DSSE, are typically required for the optimization of the model parameters. But the state variables under the new topology cannot be obtained in practice. Although several transfer learning-driven DSSE approaches have been designed to tackle the topology change issue, they still demand a certain number of labels to align with the new topology [22]. This brings huge challenges for the implementation of the existing learning-based DSSE methods.

A comprehensive comparison of existing DSSE methods for addressing variations in network topology is summarized in Table I. It can be seen from the table that, the optimization-driven approaches [5]–[11] depend on the correct topology information and line parameters of the distribution system,

TABLE I: Requirements for recent study and proposed method when the topology changes.

Ref.	Accurate line parameters	Observation for data labels	Topology information
[5]–[8]	Required	Not Required	Required
[9]–[11]	Required	Not Required	Partial
[16]	Not Required	Required	Not Required
[15], [17]	Not Required	Required	Not Required
[23], [24]	Not Required	Required	Required
[22], [25]	Not Required	Required	Partial
Pro	Not Required	Target labels Not Required	Not Required

while the learning-based approaches [15]–[17], [22]–[25] require the labels for the training of the model. However, both premises are difficult to meet in practice when a topology change occurs.

To bridge the above gaps, this paper proposes an unsupervised learning-based state estimation approach for the distribution network after topology changes. **It requires labeled data from historical topologies (source domains) but no labels are available in the new topology (target domain).** The main contributions are:

- We propose a fully unsupervised state estimation framework that achieves accurate estimation results for distribution networks experiencing topology changes without requiring physical models or labeled state variables. **This is achieved by combining a novel bipartite graph learning (BGL) module with a causality-inspired sparse structure learning domain adaptation (CDA) network.** Unlike traditional optimization-driven DSSE approaches [5]–[11] that require precise topology information, and learning-based methods [15]–[17], [22]–[25] that rely on labeled data for retraining after topology changes, our approach eliminates both dependencies.
- We introduce a bipartite graph learning method with rank constraints to automatically identify the most similar historical topology to the unseen new topology using only limited measurement data. This graph-structure-aware selection addresses a gap left by conventional feature extraction approaches [26], which typically focus on feature-wise relationships while ignoring inter-topology structural similarities.
- Our method enables effective transfer learning without labels for the target topology by leveraging **causality-inspired sparse learning domain adaptation.** By modeling sparse associative structures across different topologies, the CDA module captures domain-invariant patterns that improve estimation accuracy, even under significant topology variations.
- The proposed framework demonstrates strong scalability and robustness, validated on both standard IEEE 33-node and 119-node systems. Comparative experiments and ablation studies confirm its advantages over existing optimization- and learning-based DSSE techniques in handling topology changes without access to accurate topology or labeled data.

The rest of this paper is organized as follows. Section II introduces the DSSE task. In Section III, our proposed BGL-CDA method is presented in detail. Section IV exhibits the

performance of our method by the comparative experiments and the ablation study. Finally, conclusions are depicted in Section V.

II. PROBLEM DEFINITION

Consider a distribution network consisting of the topology structure, line parameters, state variables \mathbf{x} , and measurements \mathbf{z} collected from various devices. Since the measurements are limited by the the lack of measurement equipment in practice, pseudo-measurements are considered as additional choices. In general, based on the state variables \mathbf{x} and the measurements \mathbf{z} , the model can be formulated as follows:

$$\mathbf{z} = h(\mathbf{x}) + \mathbf{v} \quad (1)$$

where $h(\mathbf{x})$ denotes the measurement function and \mathbf{v} represents the measurement error. The traditional optimization-based DSSE method WLS tends to minimize the following specified objective function by incorporating weighted adjustments:

$$J' = [\mathbf{z} - h(\mathbf{x})]^T \mathbf{W} [\mathbf{z} - h(\mathbf{x})] \quad (2)$$

where \mathbf{W} refers to the coefficient matrix of various measurements. However, such methods require accurate physical knowledge of the model to perform the iteration process and a slight topology change may lead to a deterioration of the results. Learning-based methods based on deep neural networks are proposed to establish a mapping function between measurements \mathbf{z} and state variables \mathbf{x} without reliance on the topology information. The model can be simply defined as follows:

$$\mathbf{x}_t = f_t(\mathbf{z}_t | \theta_t) \quad (3)$$

where f represents the multi-layer neural networks and θ denotes the learned parameters. We would like to emphasize that the mapping function f is topology specific and t represents the corresponding topology scenario, which means that its parameter set θ needs be retrained when a topology change occurs. However, the state variables \mathbf{x}_t associated with the new topology are typically unknown. The lack of labeled data under the new topology brings huge challenges for traditional supervised learning-based DSSE methods. To address this issue, an unsupervised bipartite graph-enabled **causality-inspired sparse structure learning** method is proposed for the DSSE after topology changes, the details of which are shown in section III.

III. METHODOLOGY

The proposed method comprises two primary components. The first component involves a bipartite graph learning approach designed to identify the optimal topology from extensive historical topologies. This method aims to select the most suitable topology that matches the requirements of the downstream task. The second component, **causality-inspired sparse learning domain adaptation**, concentrates on extracting the sparse associative structure inherent in the selected topology. This process facilitates the migration of knowledge from the source topology to the target topology. It is noteworthy that both components operate as unsupervised processes, a facet that has not been extensively explored in recent literature.

A. Bipartite Graph Learning Method

In this part, our proposed bipartite graph learning method is divided into three modules: 1) the bipartite graph learning module, which establishes a bipartite graph to model the connections between different historical topologies. By applying connectivity constraints on the built bipartite graph, we can learn the embeddings of each topology. 2) the optimization strategy module provides a novel optimization strategy for the proposed bipartite graph learning module. 3) After obtaining the embeddings of each topology, the predicting module selects the best-matched topology for the downstream causal domain adaptation method according to the Euclidean distance.

1) Bipartite Graph Learning Module

We first aggregate all the node features in the topology into a vector for each historical topology. Then we can achieve $X \in \mathcal{R}^{d \times n}$, which includes n distribution network topologies with d features. Traditional feature-extraction methods like the autoencoder (AE) and principal component analysis (PCA) mainly focus on the relationship between different features and don't consider the structural information between different data. While subspace clustering operates under the premise that each topology can be described as a weighted sum of other topologies within the same subspace. This matrix of combination coefficients serves as the similarity graph, thereby encapsulating the global structure of all historical topologies. Typically, the model can be expressed as [27]:

$$\min_S \|X - XS^T\|_F^2 + \alpha f(S) \quad s.t. \quad S \geq 0, S\mathbf{1} = \mathbf{1}, \quad (4)$$

in which the non-negative similarity matrix $S \in \mathcal{R}^{n \times n}$ depicts the relationship between different topologies and $\alpha > 0$ acts as a role for balancing the parameter. The reconstruction error $\|X - XS^T\|_F^2$ and the regularizer function $f(\cdot)$ add up to an objective function. $S\mathbf{1} = \mathbf{1}$ implies that the total of each row in S equals one.

However, Eq. (4) only simply learns the relationship between the different topologies without considering the local cluster structure. In other words, the topology changes of the distribution network are generally depicted by several changes of similar topologies. Therefore, aiming at mining the local structure of the topology, a connectivity constraint is applied on the matrix S to make the matrix have predefined k connected components. Then, the model can be listed as follows:

$$\min_S \|X - XS^T\|_F^2 + \alpha \|S\|_F^2 \quad (5)$$

$$s.t. \quad 0 \leq S, S\mathbf{1} = \mathbf{1}, S \in \Omega.$$

Then, in order to investigate the local cluster structure of the different topologies underlying S , a bipartite graph is employed. Specifically, a bipartite graph Z is combined with S as $Z = \begin{bmatrix} S & \\ S^T & \end{bmatrix} \in \mathcal{R}^{2n \times 2n}$. Correspondingly, the normalized Laplacian matrix L can be calculated as $L = I - D^{-\frac{1}{2}} Z D^{-\frac{1}{2}}$ and each row element of diagonal matrix D is defined as $d_i = \sum_{j=1}^{2n} z_{ij}$. [28] demonstrates that the count of connected components in the bipartite graph corresponding to S matches the multiplicity k of the zero eigenvalue in the Laplacian matrix L .

According to the above analysis, the left n distribution network nodes and the right n nodes can be organized into k clusters by enforcing exactly k connected components in the bipartite graph structure. This requirement is mathematically

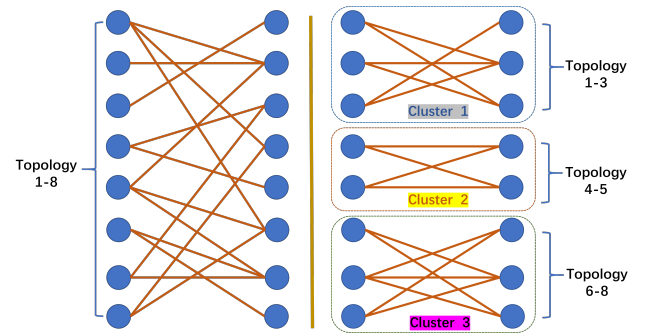


Fig. 1: The bipartite graph that depicts the relationship between different topologies with rank constraints. Before applying the constraints, the distribution networks on the left side of the bipartite graph are randomly linked to those on the right. Upon imposing the restriction, the built graph exhibits a specific count of connected subgraphs.

expressed by $\text{rank}(L) = 2n - k$, leveraging the spectral graph theory result that the number of connected components in a graph equals the multiplicity of the zero eigenvalue of its Laplacian matrix. Physically, enforcing k connected components ensures that measurements are partitioned into k disjoint, non-overlapping topology clusters, each representing a distinct and interpretable configuration mode. This aligns with the prior knowledge that distribution networks typically operate in k fundamental topology patterns, driven by different switch states or planned operational scenarios. Fig. 1 illustrates the relationships between these topologies via the bipartite graph. Based on this idea, Eq. (5) can be reformulated as:

$$\min_{S,F} \|X - XS^T\|_F^2 + \alpha \|S\|_F^2 \quad (6)$$

$$s.t. \quad 0 \leq S, S\mathbf{1} = \mathbf{1}, \text{rank}(L) = 2n - k.$$

Since the rank constraint is inherently combinatorial and non-convex, making direct optimization intractable, we adopt a relaxation approach inspired by [29]. Specifically, we replace the hard rank constraint with a spectral regularization term that encourages L to have k near-zero eigenvalues. This nuclear-norm-inspired relaxation preserves the goal of finding k well-separated connected components while ensuring the optimization problem remains convex and tractable. Consequently, our proposed Bipartite Graph Learning (BGL) module based on subspace clustering is formulated as follows:

$$\min_{S,F} \|X - XS^T\|_F^2 + \alpha \|S\|_F^2 + \beta \text{Tr}(F^T L F) \quad (7)$$

$$s.t. \quad 0 \leq S, S\mathbf{1} = \mathbf{1}, F^T F = I,$$

where $F \in \mathcal{R}^{2n \times k}$. Compared with common feature-extraction methods, this module is capable of capturing the overall structure of different topologies and exploring the local graph structure of historical topologies. The embeddings achieved based on the structural information are more suitable for selecting a similar topology. It is noteworthy to mention that Eq. (7) is an optimization-based problem, which could be addressed using an alternating approach to achieve the desired solution.

2) Optimization Strategy Module

To solve Eq. (7), an alternating optimization strategy is adopted. It fixes F and solves S and then fixes S and solves F iteratively until the iteration converges. The details are listed as follows:

Fix F and Solve S : consider that $L = I - D^{-\frac{1}{2}} Z D^{-\frac{1}{2}}$, where both Z and D are influenced by variable S . To proceed, we apply the following equation:

$$\text{Tr}(F^T L F) = \frac{1}{2} \sum_{i=1}^{2n} \sum_{j=1}^{2n} z_{ij} \left\| \frac{F_{i,:}}{\sqrt{d_i}} - \frac{F_{j,:}}{\sqrt{d_j}} \right\|_2^2 \quad (8)$$

Considering the structure of Z , we can further convert above expression into the following form:

$$\text{Tr}(F^T L F) = \sum_{i=1}^n \sum_{j=1}^n s_{ij} \left\| \frac{F_{i,:}}{\sqrt{d_i}} - \frac{F_{n+j,:}}{\sqrt{d_{n+j}}} \right\|_2^2 \quad (9)$$

Defining $\left\| \frac{F_{i,:}}{\sqrt{d_i}} - \frac{F_{n+j,:}}{\sqrt{d_{n+j}}} \right\|_2^2$ as T_{ij} , our problem can be solved row by row as:

$$\begin{aligned} \min_{S_{i,:}} \quad & S_{i,:} X^T X S_{i,:}^T - 2 X_{:,i} X S_{i,:}^T + \alpha S_{i,:} S_{i,:}^T \\ & + \beta T_{i,:} Z_{i,:}^T \quad \text{s.t. } 0 \leq S_{ij} \leq 1, \sum_j S_{ij} = 1. \end{aligned} \quad (10)$$

The issue can be efficiently tackled using convex quadratic optimization techniques.

Fix S and Solve F : when S is held constant, the initial and subsequent terms in Eq. (7) become invariant. The problem can thus be reformulated as:

$$\max_{F \in \mathbb{R}^{2n \times k}, F^T F = I} \text{Tr}\left(F^T D^{-\frac{1}{2}} Z D^{-\frac{1}{2}} F\right) \quad (11)$$

According to matrix theory, the optimal F for this problem is the top k singular vectors of L .

3) Predicting Module

After learning the embedding matrix $F \in \mathbb{R}^{2n \times k}$, the top n rows $U = \{u_1, \dots, u_n\} \in \mathbb{R}^{n \times k}$ are selected as the embeddings for all distribution networks and each row represents the embeddings of one historical topology. Then, the most widely used Euclidean distance is adopted to evaluate the similarity between different distribution networks. By calculating Euclidean distance, for every new topology, we can calculate all similarities between the target topology and all historical topologies. Finally, we choose the historical topology with the minimum Euclidean distance as the source domain dataset for subsequent tasks. The complete algorithm for BGL is summarized in Algorithm 1.

B. Causality-inspired Sparse Structure Learning Method

In section III-A, we obtain the high-quality source topology dataset for subsequent training tasks. Now we can aggregate all measurements of all nodes to form a vector. Since the topology dataset is time series data, we can define the topology as follows:

$$\mathbf{x} = \{\mathbf{x}_{t-T+1}, \dots, \mathbf{x}_{t-1}, \mathbf{x}_t\} \quad (12)$$

where $\mathbf{x}_t \in \mathbb{R}^M$ represents the measurements of the topology and M is the count of the measurements for one certain topology, $\mathbf{y}_t \in \mathbb{R}^V$ denotes the results of the predicting state variables of all nodes in the topology. Then the distribution of the source topology obtained in III-A can be defined as $P_S(\mathbf{x}, \mathbf{y})$, where each source domain sample x_S has a unique corresponding label y_S and the labels are known for us, while the distribution of the target topology is formulated as $P_T(\mathbf{x}, \mathbf{y})$. Different from the source topology, the labels y_T are unknown to us and our purpose is to predict the state

Algorithm 1 Bipartite Graph Learning

Input: n distribution network topologies $T = \{t_1, \dots, t_n\}$, parameters α , predefined subgraphs k , parameters β .

Output: Topology t_i .

- 1: Establish historical topologies matrix $X = \{x_1, \dots, x_n\} \in \mathbb{R}^{d \times n}$ by aggregating the node features of each topology.
- 2: Assign random values to the matrix F .
- 3: **while** the convergence criterion is not satisfied **do**
- 4: Adjust S in Eq. (10) through convex quadratic programming.
- 5: Adjust F in Eq. (11) by computing the top k singular vectors of L .
- 6: **end while**
- 7: Select the top n rows of F to obtain $U = \{u_1, \dots, u_n\} \in \mathbb{R}^{n \times k}$ as the embeddings for the new topology matrix.
- 8: Compute the Euclidean distance between the target topology representation and each historical topology representation in U , and select the topology index t_i with the minimum distance.

variables y_T by utilizing the source domain $P_S(\mathbf{x}, \mathbf{y})$ and data samples x_T .

Considering that both the source topology and the target topology are typically generated by the same underlying causal mechanisms, we aim to learn a structural representation that maintains predictive stability under varying topology scenarios $e \in \mathcal{E}$. Ideally, we seek for each node i a set of causal parents $\text{PA}(i)$ such that the conditional distribution

$$P_e(x_i | x_{\text{PA}(i)}) \quad (13)$$

remains invariant across all environments $e \in \mathcal{E}$. Under the framework of structural equation models (SEM), the causal relationships among variables can be modeled as:

$$\mathbf{x} = B^* \mathbf{x} + \boldsymbol{\varepsilon}^{(e)}, \quad (14)$$

where B^* denotes the true adjacency matrix of the causal graph, and $\boldsymbol{\varepsilon}^{(e)}$ represents the environment-specific noise. Causal invariance requires that the conditional distribution of x_i given $\text{PA}(i)$ remains unchanged for all e . However, in practical distribution systems, identifying B^* precisely faces two major challenges. When unobserved confounders or measurement noise exist, the causal structure is generally unidentifiable due to Markov equivalence among different DAGs. Even under ideal assumptions (complete observability, no confounding, linear Gaussianity), identifying the optimal DAG still requires enumeration over all possible directed acyclic graphs, with the search space given by:

$$O\left(d! \cdot 2^{\frac{d(d-1)}{2}}\right), \quad (15)$$

which is NP-hard. To address this, we relax our objective from extracting the full invariant causal structure to learning a sparse associative structure. This structure is modeled via attention mechanisms: for each target node i , we introduce a set of weights which quantify the strength of dependency from node j to node i . These weights are obtained using the sparsemax function, which enforces sparsity by forcing most weights to be exactly zero. Essentially, the learned sparse associative structure provides a computationally feasible and trainable approximation to the underlying causal structure. Rather than aiming for exact causality, we target predictive

stability—selecting a set of neighbor nodes that remain effective for predicting the target node across different topological configurations. This strategy improves model scalability and aligns with engineering intuition in power systems, where critical causal chains tend to remain invariant during topology transitions.

Therefore, we propose a method to acquire fine-grain segments of the time-series topology, thereby alleviating the challenge posed by offsets. Then, taking into account time lags from various domains, two attention mechanisms are implemented for exploring the underlying correlation between measurements and the relationship within the measurements. **Our causality-inspired domain adaptation (CDA) model exploits the intrinsic properties of the distribution network and aligns shared relational structures across different topologies, thereby deriving domain-independent representations for unsupervised regression.**

1) Fine-grain Segment Preprocess

Considering the topology dataset is a long time series data, it is impossible to input the whole dataset. Besides, the whole dataset as input is unable to perceive the trend of measurement changes and the mutual influence between measurements. To this end, given a predefined parameter T , for each variable x^i , we divided it into several time-series segments with different lengths and the maximum length is T . Then it yields:

$$\tilde{x}^i = \{x_{t:t}^i, x_{t-1:t}^i, \dots, x_{t-\tau+1:t}^i, \dots, x_{t-T+1:t}^i\}. \quad (16)$$

In this study, we set $T = 24$ since the measurement data are sampled on an hourly basis and the distribution network states generally follow a strong daily periodicity of 24 hours. This setting allows each segment to fully capture one natural day of variations in voltage, current, and load. If T were smaller than 24, each segment would fail to represent a complete daily cycle; if T were larger than 24, it would mix inter-day heterogeneity and increase computational burden. Therefore, $T = 24$ provides a reasonable balance between capturing meaningful daily patterns and maintaining tractability. Then to investigate the impact of all measurements, each variable $x_{t-\tau+1:t}^i$ comes with a single LSTM. It can be formulated as:

$$h_\tau^i = LSTM(x_{t-\tau+1:t}^i; \theta^i) \quad (17)$$

where θ^i denotes the set of learnable parameters (weights and biases) of the LSTM corresponding to variable i . Then the new embedding of the time-series segments can be described as:

$$h^i = \{h_1^i, h_2^i, \dots, h_\tau^i, \dots, h_N^i\}. \quad (18)$$

Now we achieve the adaptive segment representations from the original topology.

2) Sparse Associative Structure Establish

After obtaining the adaptive segment representation h^i , we can find that Eq. (18) does not consider the weights for different length segments. It simply combines all the segments with the same importance. Therefore, to fix on some important segments, the self-attention mechanism is proposed for feature exacting [30]. The new segments Z is described as:

$$Z^i = \sum_{\tau=1}^N \alpha_\tau^i \cdot (h_\tau^i \mathbf{W}^V); u_\tau^i = \frac{1}{N} \sum_{k=1}^N \frac{(h_k^i \mathbf{W}^Q) (h_k^i \mathbf{W}^K)^\top}{\sqrt{d_h}} \quad (19)$$

$$\alpha^i = \{\alpha_1^i, \dots, \alpha_\tau^i, \dots, \alpha_N^i\} = \text{spm}(u_1^i, \dots, u_\tau^i, \dots, u_N^i)$$

where \mathbf{W}^Q , \mathbf{W}^K , \mathbf{W}^V are parameters for training and $\sqrt{d_h}$ is the scaling factor. Aiming at selecting the most representative segment representations, a sparsemax function is proposed in [31] for computing the weights as follow:

$$\text{spm}(\mathbf{u}) = \arg \min_{\mathbf{q} \in \Delta^{K-1}} \|\mathbf{q} - \mathbf{u}\|^2 \quad (20)$$

It yields the Euclidean projection onto the probability space Δ^{K-1} .

Then we can establish the sparse associative structure by exploring the underlying relationship between different representations Z^i . Another form of attention mechanism proposed in [32] is adopted to evaluate the importance between two representations:

$$e^{ij} = \frac{Z^i \cdot Z^j}{\|Z^i\| \cdot \|Z^j\|} \quad (21)$$

Unfortunately, Eq. (21) does not consider the effects of the time lags, which could lead to a misleading associative structure. Therefore, we redefined e^{ij} as follows:

$$e_\tau^{ij} = \frac{Z^i \cdot h_\tau^j}{\|Z^i\| \cdot \|h_\tau^j\|}; e^{ij} = \{e_1^{ij}, e_2^{ij}, \dots, e_\tau^{ij}, \dots, e_N^{ij}\}. \quad (22)$$

Then we do the same operation in Eq. (20) to obtain the weights. To be precise, the associative structure between two measurements is formulated as follows:

$$\beta^i = \{\beta^{i1}, \beta^{i2}, \dots, \beta^{ij}, \dots, \beta^{iM}\} \\ = \text{spm}(\{e^{i1}, e^{i2}, \dots, e^{ij}, \dots, e^{iM}\}) (j \neq i). \quad (23)$$

where β^{ij} represents the associative value between variable i and variable j .

3) Predicting Module

According to the learned representations β in Eq. (23), we can establish the associative structure representations U and aggregate the representations with fine-grain segment representation Z to form the final representations J for subsequent regression tasks. It yields:

$$U^{ij} = \sum_{\tau=1}^N \beta_\tau^{ij} \cdot h_\tau^j; U^i = \sum_{m=1, m \neq i}^M U^{im}; J^i = [Z^i; U^i] \quad (24)$$

Then we can regard the whole process as a deep neural network G_J with two different attention mechanisms [33]. It can be depicted as follows:

$$\mathbf{J} = G_J(f(\mathbf{x}; \Theta); \mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V), \quad (25)$$

where $\mathbf{J} = [J^1; J^2; \dots; J^M]$ represents the exacting representation and Θ denotes the parameters of each LSTM. Then we simply input the representation into the regression neural network to yield the results of the state variable of all nodes in the topology. The whole network can be defined as follows:

$$y_{\text{res}} = G_y(G_J(f(\mathbf{x}; \Theta); \mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V), \phi). \quad (26)$$

4) Loss Function

In this part, we elaborate on the loss function of the proposed method. Our loss function consists of three parts. α proposed in Subsection III-B2 represents the probability of predefined length of certain segments. We assume that the similar topology should share the same pattern of the duration of the segment. By applying this constraint, the subsequent

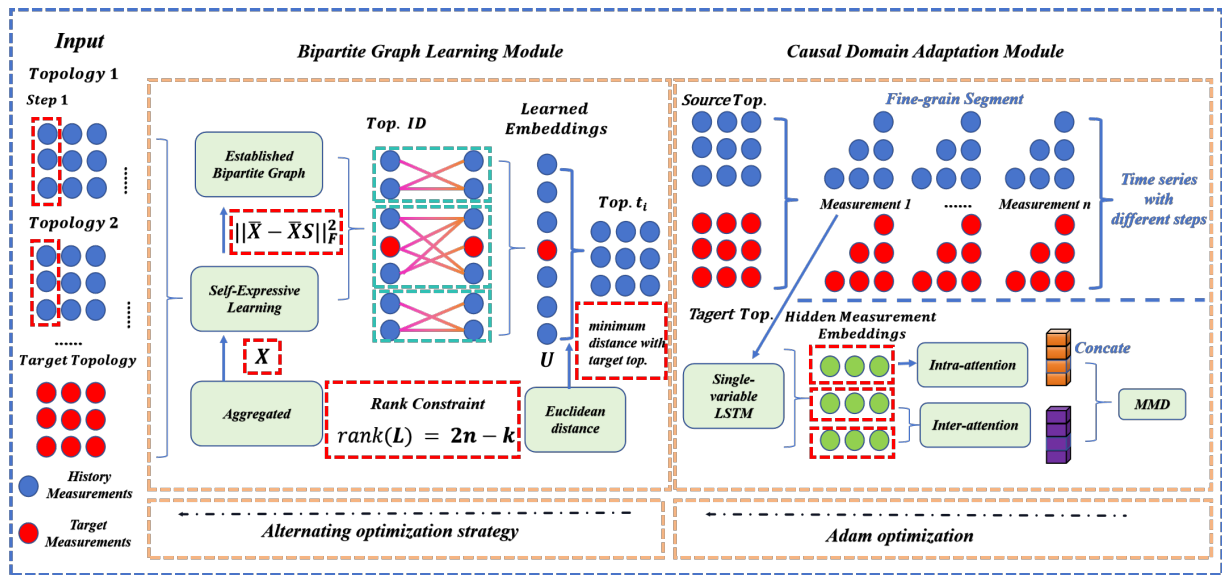


Fig. 2: The framework of the proposed method.

learned associative structure could be captured more efficiently. Besides, the learned associative structure should also follow this assumption to ensure the effectiveness of mining the domain-invariant associative structure. Therefore inspired by the most widely used metric maximum mean discrepancy (MMD), we have the following two loss function terms [34]:

$$\mathcal{L}_\alpha = \sum_{m=1}^M \left\| \frac{1}{|\mathcal{X}_S|} \sum_{x_S \in \mathcal{X}_S} \alpha_S^m - \frac{1}{|\mathcal{X}_T|} \sum_{x_T \in \mathcal{X}_T} \alpha_T^m \right\|, \quad (27)$$

$$\mathcal{L}_\beta = \sum_{m=1}^M \left\| \frac{1}{|\mathcal{X}_S|} \sum_{x_S \in \mathcal{X}_S} \beta_S^m - \frac{1}{|\mathcal{X}_T|} \sum_{x_T \in \mathcal{X}_T} \beta_T^m \right\|.$$

where α_S^m , α_T^m , β_S^m , β_T^m , denote the weights of segments and the associative structure of the m -th variable from the source topology and the target topology obtain in Eq. (19) and Eq. (23).

For the regression network $G_y(\cdot; \phi)$, the MAE is adopted as the objective function. Finally, the comprehensive loss function of the proposed structure alignment model for time series domain adaptation is expressed as:

$$\mathcal{L}(\Theta, \mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V, \phi) = \mathcal{L}_y + \gamma(\mathcal{L}_\alpha + \mathcal{L}_\beta) \quad (28)$$

where γ is hyper-parameter. The whole regression neural network, utilizing the trained optimal parameters, is tailored for the target domain. Following the objective function outlined above, the proposed model is further trained on both the source and target topologies using the subsequent steps:

$$\left(\Theta, \mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V, \phi \right) = \arg \min_{\Theta, \mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V, \phi} \mathcal{L}(\Theta, \mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V, \phi). \quad (29)$$

The comprehensive algorithm is shown in Algorithm 2. Finally, the whole process of our method is depicted in Fig 2.

IV. CASE STUDY

This section presents comparative experiments on the IEEE 33-node and 119-node distribution networks to assess the effectiveness of the proposed approach.

Algorithm 2 Causality-inspired Sparse Structure Learning Domain Adaptation

Input: Source topology distribution $P_S(x, y)$ and target topology distribution $P_T(x, y)$, hyper-parameter γ

Output: State variables y_{res} .

- 1: Initialize all parameter of the network randomly.
- 2: Obtain the adaptive segment representation h^i in Eq. (18)
- 3: **while** convergence condition does not meet **do**
- 4: Randomly select a batch of data B from $P_S(x, y)$ and $P_T(x, y)$.
- 5: Compute the outcomes for the data batch B using the model.
- 6: Calculate the loss function by Eq. (28).
- 7: Update the parameter by Adam optimization.
- 8: **end while**
- 9: Output the parameters of the model and predict the state variables y_{res} in the target topology.

A. Experimental Setup

The original topology of 33-node system is illustrated in Fig. 3. We install photovoltaic (PV) units at bus 5, 12 and 30 with a capacity of 400 kW. The real-time measurements consist of the active injection power and reactive injection power of lines 1-5 and 6-11. 1% uniform noise is added to the real-time measurements. The active and reactive power of all nodes in the DN constitute the input pseudo-measurements. A uniform noise of 50% is incorporated into the pseudo-measurements. The paper obtained 8760 sets of simulation data for the whole year from each topology structure through a simulation model, which includes 24 sets of data per day for the whole year (365 days). **In practice, the state labels are typically obtained through conventional state estimation approaches (such as WLS-based DSSE) or extracted from offline validated operational records, which are routinely accessible in real-world distribution systems.**

Besides, based on the original topology, another 12 topologies are built to test the performance by alternating the status of the switches. The specific settings for each topology are provided in Table. II. When one topology is set as the new topology, the other 12 topologies are regarded as historical

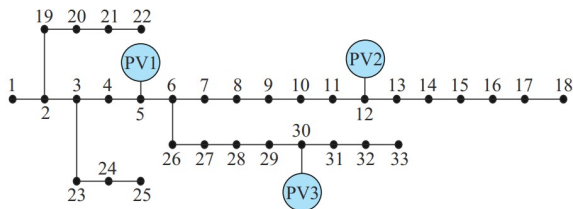


Fig. 3: The original topology of the IEEE 33-node system.

topologies. The training data of our model consists of two parts. The first part is the training samples of the historical topology, which include 1000 instances of measurements and corresponding labeled data (state variables). The second part is the 50 instances of real-time and pseudo-measurements collected from the new topology. The training process of the proposed method can be divided into two stages. In the initial stage, 50 instances of measurements collected from the new topology and each historical topology are first processed by the BGL method. The unsupervised learning process allows the selection of the best-matched topology from historical topologies. For the BGL stage, the regularization parameters α and β in Eq. (7) were fixed at 1 and 0.1, respectively, for all experiments. These choices were based on unsupervised analyses of reconstruction error, sparsity, and block-structure quality across scenarios, ensuring stable, consistent, and interpretable results without the need for topology-specific tuning. Then, the CDA method is employed to enhance the knowledge transfer using the 50 measurements of the new topology and 1000 instances of training samples of the selected topology. We would like to emphasize that our experimental setting is in accord with the practical conditions since only 50 instances of the new topology are used. Although the labeled data of the historical topology are utilized in the second stage, we believe this can be achieved for the historical topology by existing DSSE methods.

To prove the effectiveness of our proposed model, four learning-based methods, including BPN, CNN, GP, and GAT are implemented. BPN contains four FCN layers and adopts the Tanh as the activation function. CNN consists of three one-dimensional layers and adopts Leakyrelu as the activation function. GAT adopts a two-layer structure with an attention mechanism. The learning rate of these methods is set to $1e-4$. All the models use Adam optimization during the training process. In the case of Gaussian process regression (GP), the radial basis function (RBF) is employed as the kernel function to model the underlying data relationships. Since the CNN, BPN, and GP methods don't rely on the structure of the topology, they are trained using the samples of all historical topologies. For GAT, 12 models are trained under all historical topologies and averaged values of all the models are utilized as the estimated results for the new topology. Two WLS-based methods, namely WLS-W and WLS-optimal methods, are considered as benchmarks. For the WLS-W method, the accurate structural information of the new topology obtained after a topology change is not known and that of the original topology is used. For the WLS-optimal method, it is assumed that the structural information is accurately known. The WLS is implemented in MATLAB, whereas the remaining learning-based techniques are executed using Python. Specifically, all learning-based methods are implemented with PyTorch 2.3.1 and executed on a server equipped with an Intel i9-14900K CPU, a GeForce RTX 4080S GPU, and 96 GB of RAM,

TABLE II: The alternations of the topology based on the original topology.

Topology ID	Alternations of the switch status
2	14-15 disconnected, 9-15 connected
3	20-21 disconnected, 8-21 connected
4	28-29 disconnected, 25-29 connected
5	32-33 disconnected, 18-33 connected
6	24-25 disconnected, 29-25 connected
7	17-18 disconnected, 33-18 connected
8	21-22 disconnected, 12-22 connected
9	14-15, 20-21 disconnected, 9-15, 8-21 connected
10	14-15, 28-29 disconnected, 9-15, 25-29 connected
11	14-15, 24-25 disconnected, 9-15, 29-25 connected
12	14-15, 17-18 disconnected, 9-15, 33-18 connected
13	14-15, 21-22 disconnected, 9-15, 12-22 connected

allowing for a comprehensive and consistent comparison across platforms.

B. Evaluation Metrics

The mean absolute error (MAE) is adopted to evaluate the performance of our proposed method. Specifically, given the actual state variable and the predicted results, it is calculated according to:

$$MAE = \sum_{i=1}^n |x_i - \hat{x}_i| / n \quad (30)$$

Here, x_i and \hat{x}_i denote true values and predicted results of state variables for the i -th nodes, respectively. n represents the overall count of nodes within the evaluation system.

C. Evaluation of the Proposed Method

The results regarding voltage magnitude and angle are presented in Table III. Each row of topology represents the changed topology, and the other 12 topologies are used as historical topologies. As seen in the table, the WLS-optimal method can achieve satisfying performance when accurate topology information is used. However, the DN may experience frequent topology alterations. When a topology change takes place, it's challenging to obtain accurate structural information. This complicates the WLS method's ability to adapt to new topology since it relies on accurate structural information. As a result, the WLS-W method, which uses wrong topology information for DSSE tasks, suffers from obvious performance degradation compared with that obtained by the WLS-optimal method. The huge impact of accurate topology on the performance of WLS-based DSSE approaches are observed. Although the topology information is not required by the BPN, CNN, and GP methods, they rely on labeled data for the adjustment of their parameters to adapt to the unseen DN. In the context of state estimation, the labeled data refer to the state variables recorded under the new topology, which are impossible to obtain. As a sequence, large estimation errors are observed for the BPN, CNN, and GP methods. It demonstrate the deficiency of conventional learning-based DSSE approaches in tackling topology changes. Different from traditional learning-based methods, the integration of BGL and CDA method enables the proposed approach to achieve effective adaption to the unseen DN without accurate topology information and labeled data. Its performances are close to that obtained by the WLS-optimal method in most cases. Note that the WLS method depends on the precise information of the new topology and line parameters that

TABLE III: The MAEs of magnitude and angle tasks under different cases for various DSSE methods on IEEE 33-node test system. The units for MAEs of voltage magnitude and voltage angle are $1e-4$ p.u. and $1e-3$ p.u. respectively.

Test Topology ID	BPN		CNN		WLS-W		GP		GAT		Proposed		WLS-optimal	
	Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang
2	26.14	34.4	24.84	35.4	7.98	26.9	21.74	38.5	48.96	61.9	7.91	23.5	5.42	21.1
3	44.18	51.7	37.45	59.4	26.16	48.8	37.66	55.8	88.12	101.3	8.49	19.8	5.73	25.3
4	37.69	71.3	31.48	66.8	36.56	78.2	34.15	67.4	59.15	99.4	6.31	15.9	6.80	19.2
5	33.21	46.4	35.56	42.4	8.79	30.5	26.89	52.9	45.78	78.1	9.92	35.5	4.23	18.1
6	43.77	63.8	39.15	59.5	19.88	45.7	38.97	49.4	61.55	110.4	11.29	46.7	7.13	22.4
7	34.46	46.9	37.96	38.4	10.78	33.6	37.65	38.8	58.41	74.2	8.13	31.2	5.43	16.1
8	37.86	49.6	36.23	41.7	16.78	35.8	39.86	47.5	53.69	65.7	8.09	19.3	3.95	18.8
9	40.73	56.2	45.75	52.8	28.46	58.3	34.57	42.6	71.56	73.2	9.41	47.4	6.41	24.1
10	39.48	84.2	42.98	79.6	37.84	84.5	33.49	77.9	76.25	89.9	11.23	13.9	7.92	19.6
11	39.53	58.2	43.47	61.8	21.99	56.3	41.78	61.2	48.16	73.6	8.83	38.6	4.45	23.5
12	38.64	53.1	36.89	49.9	13.88	42.6	34.17	51.9	48.65	69.4	13.12	37.8	9.43	19.9
13	32.72	43.6	31.71	47.1	18.45	43.7	35.48	38.4	51.93	61.2	7.29	27.1	5.93	17.4

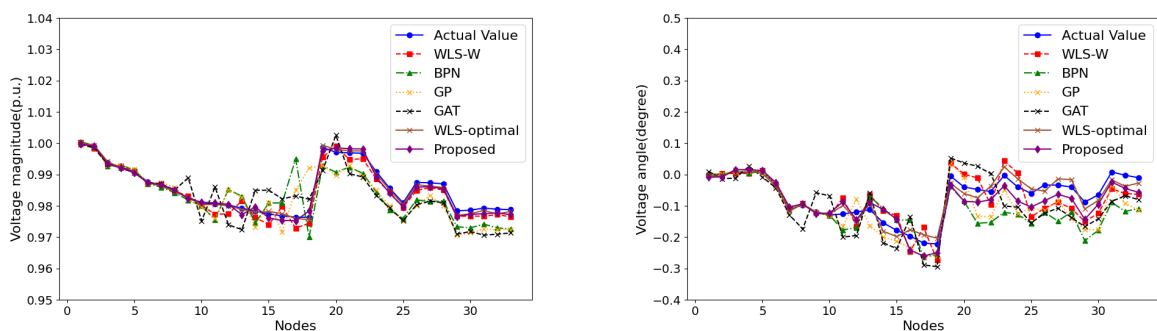


Fig. 4: The voltage estimated by different DSSE methods under topology 7: voltage magnitude (left); voltage angle (right).

are challenging to acquire in practice, while our method can avoid the dependence on that by learning from data. The results demonstrate the adaptability to topology changes. To further evaluate the performance achieved by different DSSE methods, the profiles of voltage for various approaches on topology 7 are plotted in Fig. 4. It can be seen that WLS-W, BPN, CNN, GP and GAT methods significantly deviate from actual values of the voltage magnitude and angle in most nodes. In contrast to the WLS-optimal method, our proposed approach achieves comparable outcomes without relying on precise structure information and line parameters.

D. Ablation Test

1) Evaluation of the Impact of the BGL Module

In this subsection, comparative tests are carried out to verify the effectiveness of the proposed BGL module. Several benchmarks are considered: 1) Non-CDA method, where the original topology is directly used as the source topology and the CDA method is employed for the training of the state estimator; 2) Normal-CDA, PCA-CDA, and AE-CDA methods, where the original features, the features extracted by the PCA and autoencoder (AE) are first utilized to select the most similar topology according to the Euclidean distance. The CDA method is then used for the training of the estimator based on the selected topology. The comparison results are listed in Table IV.

When the original topology is directly used as the source topology, the estimation errors are relatively large in most cases for the Non-CDA method. This demonstrates that the naive use of the original topology doesn't help in the state

TABLE IV: Ablation study of the BGL module.

Topology ID	Proposed		Normal-CDA		PCA-CDA		AE-CDA		Non-CDA	
	Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang
2	7.91	23.5	32.67	103.2	14.67	34.1	12.12	35.5	23.45	52.8
3	8.49	19.8	45.68	82.9	33.57	75.3	8.63	22.1	17.66	28.8
4	6.31	15.9	6.94	19.4	6.31	15.9	6.31	15.9	26.37	61.9
5	9.92	35.5	38.76	53.5	9.92	35.5	21.76	53.8	9.92	35.5
6	11.29	46.7	36.88	79.4	33.37	67.8	41.89	108.2	38.78	92.4
7	8.13	41.2	27.69	111.7	12.78	36.2	24.66	55.1	21.44	74.3
8	8.09	19.3	8.09	19.3	8.09	19.3	8.09	19.3	13.76	22.0
9	9.41	47.4	37.44	79.1	7.76	22.7	39.88	71.4	9.41	47.4
10	11.23	13.9	35.94	61.4	11.23	13.9	12.68	16.9	12.68	16.9
11	8.83	38.6	55.17	93.5	13.36	44.9	13.36	44.9	15.88	59.8
12	13.12	37.8	29.48	64.4	34.72	81.5	13.12	37.8	18.33	55.4
13	7.29	27.1	27.86	41.4	32.65	56.2	7.29	27.1	25.54	69.8

estimation task under the new topology. When similar topologies are selected according to the features extracted by the PCA and AE methods, the PCA-CDA and AE-CDA methods can achieve better performance than the Non-CDA method. This demonstrates that the feature extraction process can aid in the selection of a similar topology. However, since they ignore the local graph-structure information among the measurement data, the performance enhancements are limited. By contrast, our proposed method explores the underlying graph structure hidden in the data by applying prior rank constraints. This allows the proposed method to achieve the best performances on 11 out of 12 conditions. The results illustrate the effectiveness of the proposed bipartite graph

learning method.

2) Evaluation of the Impact of the CDA Module

To examine the superiority of our proposed CDA module, comparative tests are performed among various benchmarking methods: 1) the BGL-BPN method, where the CDA module of our approach is replaced by the BPN method. The BPN-based method is trained to utilize the data recorded under the similar topology selected by BGL; 2) the BGL-WLS method, where the state variables are calculated by the WLS method according to the structural information of the similar topology selected by the BGL module; 3) the BGL-DANN method, where the CDA module is replaced by a Domain-Adversarial Neural Network (DANN) [35]. In this variant, the DANN model is trained to extract domain-invariant features between the selected similar historical topology (source domain with labels) and the new topology (target domain without labels) using adversarial learning to reduce domain discrepancy; 4) This variant replaces the CDA module with the Source Hypothesis Transfer (SHOT) method [36]. SHOT aligns source and target distributions in the hypothesis space by freezing the source classifier and adapting the target features to preserve source decision boundaries, thereby avoiding target label reliance. The results obtained by different methods are listed in Table V.

TABLE V: Ablation study of the CDA module.

Topology ID	Proposed		BGL-DANN		BGL-SHOT		BGL-BPN		BGL-WLS	
	Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang	Mag	Ang
2	7.91	23.5	8.45	25.5	9.30	24.7	11.49	38.2	13.18	25.1
3	8.49	19.8	9.67	25.0	8.91	21.4	14.86	36.1	13.71	55.3
4	6.31	15.9	7.50	19.2	7.74	20.9	7.64	23.3	11.31	24.6
5	9.92	35.5	11.7	32.5	11.02	37.8	11.49	43.8	13.42	37.7
6	11.29	46.7	12.44	49.6	18.66	54.37	21.68	64.7	17.21	58.4
7	8.13	41.2	11.2	45.8	9.42	44.5	15.88	53.5	12.78	47.3
8	8.09	19.3	7.94	21.1	8.35	23.0	8.46	24.8	9.59	14.3
9	9.41	47.4	11.52	50.5	13.45	55.5	14.74	58.2	10.88	72.7
10	11.23	13.9	13.32	18.5	12.27	17.4	16.49	28.4	19.36	23.4
11	8.83	38.6	9.94	41.3	11.72	45.5	11.37	46.5	10.97	48.0
12	13.12	37.8	18.25	39.2	16.82	41.0	21.44	43.0	18.46	47.4
13	7.29	27.1	8.14	28.8	8.71	28.4	9.14	31.7	10.34	29.8

When the BPN is trained using the data recorded under a similar topology, the performances on the test topology outperform that obtained using all historical topologies in Table III. However, the BGL-BPN method only utilizes the data of the historical topology. The potential values of the measurements recorded under the new topology are not exploited. Different from the BGL-BPN method, the proposed method leverages the advantages of transfer learning and explores the causal domain-invariant features between the source topology and target topology. This allows it to fully exploit the values of the measurement data of the new topology. As a result, it achieves better performance than that of the BGL-BPN method. The proposed approach also surpasses the performance of the BGL-WLS method, which depends on precise line parameters that are challenging to acquire in real-world scenarios.

Furthermore, we also compare our approach with the BGL-DANN and BGL-SHOT baselines. These methods replace the CDA module with standard unsupervised domain adaptation techniques: DANN uses adversarial learning to enforce domain-invariant representations between the selected similar historical topology (source domain with labels) and the new

topology (target domain without labels) [35], while SHOT transfers source hypotheses to the target domain without accessing target labels by optimizing feature clustering and hypothesis consistency [36].

However, these standard domain adaptation methods have a key limitation: they rely solely on generic feature alignment or prediction consistency mechanisms without explicitly modeling the underlying causal relationships and physical associations between different power system topologies. In distribution network topology switching scenarios, changes in measurement data involve not just simple distribution shifts but also complex dependency structure changes induced by the network reconfiguration. DANN and SHOT may align or transfer features indiscriminately across all dimensions, lacking the selective and physically interpretable mechanism needed to ensure that the learned predictive relationships remain stable and useful under different topologies.

By contrast, the proposed method can achieve the best performance without reliance on the system parameters and the labeled data under the new topology. The results illustrate the effectiveness of the proposed CDA module.

E. Hyperparameter Sensitivity Analysis

In our bipartite graph learning module, the rank parameter k controls the rank constraint on the Laplacian matrix, thereby determining the number and granularity of matched clusters between historical and target topologies. Theoretically, since our bipartite graph contains 12 historical topology nodes and 12 target topology nodes, the maximum number of connected components is 12, which would imply each historical and target topology forms an isolated one-to-one match. However, such extreme fragmentation would lose the potential for shared structure and cross-domain transfer. Conversely, setting $k = 1$ would force all topologies into a single cluster, oversimplifying the system's diversity. To balance these extremes, we selected a moderate range of $k = 3$ to 9 in our experiments.

Figures 5 illustrate the prediction error (MAE, scaled by 10^{-4}) for two representative target topologies as k varies. For Topology 2, the error shows a U-shaped trend: it decreases until $k = 5$ and then increases again at larger k . For Topology 11, the error decreases sharply from $k = 3$ to 5 and remains flat and minimal between $k = 5$ and 8, indicating robustness across a moderate k range. These results confirm that while k is important, the model is not overly sensitive within [4, 8], ensuring stable alignment without precise tuning.

We further evaluated the sensitivity of the regularization weights α and β by varying them in $\{0.01, 0.1, 1, 10, 100\}$ on IEEE-33 (Topology 7). As shown in Fig. 6, the model remains very stable when $\alpha, \beta \in [0.1, 10]$ ($\text{MAE} \approx 8.1 \times 10^{-4}$), while extreme values cause performance degradation (up to 2.7×10^{-3}). This indicates that the BGL module is not overly sensitive to α and β , and default values such as $(\alpha, \beta) = (1, 1)$ are sufficient.

Finally, we analyzed the influence of γ , the weight of the MMD alignment term in the CDA loss. Table VI reports the results on IEEE-33 (Topologies 2 and 5). When $\gamma = 0$, no transfer is performed, leading to large errors (15.3 and 17.8). Increasing γ significantly improves performance, with the best trade-off achieved at $\gamma = 1-2$ (7.91 for Topology 2, 9.82 for Topology 5). However, when γ is too large (e.g., 100), the errors increase sharply (31.6 and 48.2) and training

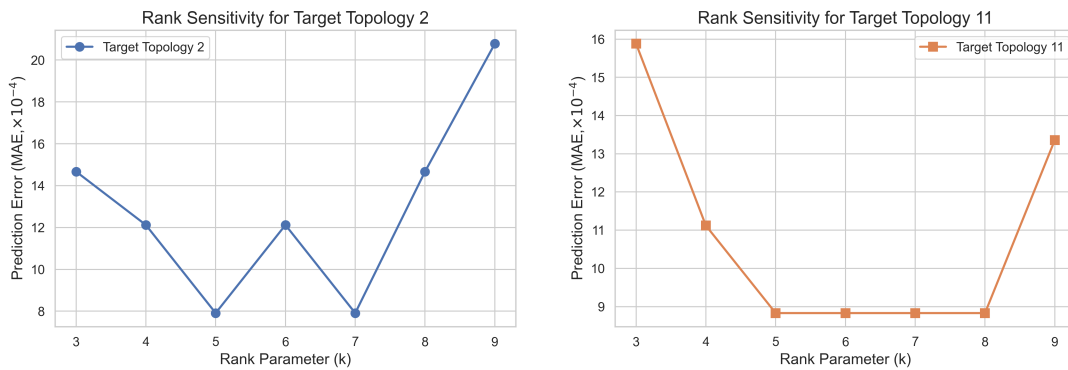


Fig. 5: Prediction error (MAE, $\times 10^{-4}$) across different rank parameter values: Topology 2 (left); Topology 11 (right).

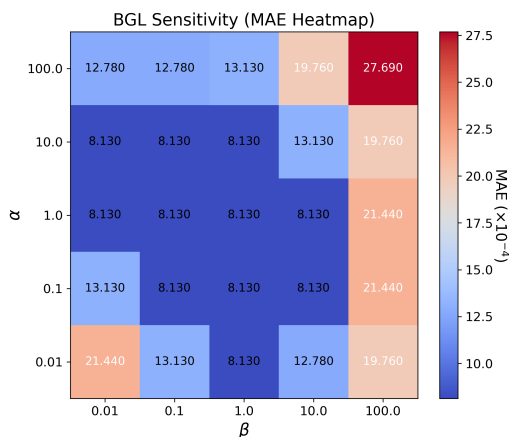


Fig. 6: Sensitivity analysis of α, β in BGL on IEEE-33 (Topology 7). MAE ($\times 10^{-4}$) for each configuration.

TABLE VI: Impact of different γ on target-domain voltage magnitude MAE under IEEE-33 Topologies 2 and 5. $\gamma = 0$ indicates no transfer. MAE ($\times 10^{-4}$).

γ	Topology 2	Topology 5
0	15.3	17.8
0.1	11.2	13.5
1	7.91	9.92
2	8.03	9.82
10	12.7	14.9
50	14.8	16.9
100	31.6	48.2

convergence slows down, as the loss becomes dominated by MMD and the regression objective is weakened.

In summary, the sensitivity experiments confirm that the proposed method is robust to hyperparameter settings. The rank parameter k and BGL regularization parameters α, β have only mild effects within reasonable ranges, while γ achieves stable balance between alignment and regression around 0.1–10, with excessively large values leading to degradation. This demonstrates the practical usability of our framework without requiring precise hyperparameter tuning.

F. Time Complexity and Computation Time Experiments

This section provides a detailed analysis of the computational cost of the proposed method, focusing on its suitability for real-time online state estimation with second-level response requirements in distribution networks. The method is designed with a two-stage architecture consisting of offline training and online inference. Accordingly, we separately analyze the computational demands of the offline domain

adaptation training phase and the online single-pass inference phase.

1) Theoretical Time Complexity Analysis of BGL and CDA

The proposed method consists of two key computational components: the BGL module and the CDA module. We analyze their theoretical time complexity to demonstrate their scalability and practical feasibility.

The BGL module aims to perform subspace clustering by minimizing reconstruction error with rank constraints or their relaxed forms. Its optimization typically involves two main steps. First, in the update of the coefficient matrix S , the problem reduces to solving a regularized constrained quadratic program. Given an input observation matrix of size $n \times d$ and a clustering dimension k , the matrix multiplication and projection updates have a complexity of approximately $O(ndk)$. This step can also be efficiently parallelized or block-optimized in practice. Second, in the update of the graph Laplacian L or its low-rank embedding F , eigenvalue decomposition is required to obtain the leading k dimensions. While full eigendecomposition of an $n \times n$ matrix has complexity $O(n^3)$, practical implementations typically extract only the top r eigenvectors, reducing the complexity to $O(rn^2)$ where $r \ll n$. As a result, the overall per-iteration complexity of the BGL module can be approximated as $O(ndk + rn^2)$. Given that in distribution network applications n, d , and k are of moderate, controllable size, and leveraging efficient numerical linear algebra libraries, the BGL module can converge in a reasonable time during offline training.

The CDA module includes two key stages: time-series segment encoding and attention-based sparse associative structure learning. In the LSTM-based encoding stage, for a time-series segment length T and hidden dimension d , the per-variable complexity is $O(Td^2)$. For M observed variables, the total encoding complexity becomes $O(MTd^2)$. Since T is typically a short window (tens of steps) and d is a design-controlled medium dimension, this stage is linearly scalable and efficient in practice. In the attention mechanism, the module computes intra-variable attention among N time segments with complexity $O(Nd^2)$ per variable, and inter-variable attention across all M variables with complexity $O(Md^2)$. Therefore, the overall forward-pass complexity of the CDA module can be expressed as $O(MTd^2 + MNd^2 + Md^2)$. Given that T, N , and M are selected as moderate hyperparameters in distribution network applications, the CDA module ensures efficient online inference performance suitable for real-time state estimation requirements.

This theoretical analysis confirms that both the BGL and CDA modules are designed with well-controlled, scalable

computational complexity, supporting their integration into practical distribution network state estimation systems.

2) Experimental Analysis of Training Time and Scalability

In this experiment, we analyze the time complexity characteristics of the BGL module through simulation. Given that real-world distribution networks typically have a limited number of historical topology samples, we employed a GAN-based generative adversarial approach to expand and simulate larger-scale historical datasets, with sample sizes set to 12, 120, 1200, and 12000. The experimental results demonstrate that the total training time of the BGL module shows clear nonlinear growth as the number of historical topologies increases, consistent with the theoretical complexity driven by matrix operations and eigenvalue decomposition that scale rapidly with data size. At the largest scale of 12000 samples, the training time exceeded 5000 seconds, illustrating the computational cost of performing global subspace learning under extreme scenarios. However, it is important to note that in typical distribution network planning and operation, the number of observable historical topologies rarely reaches such extreme scales and generally remains in the range of tens to a few hundreds. Therefore, the proposed method maintains reasonable computational cost and acceptable offline training times in practical applications.

On the other hand, the CDA module's unsupervised domain adaptation stage is designed to adapt to a single new target topology at a time, without processing all historical topologies simultaneously. As a result, its offline adaptation training time remains effectively constant in our experiments, measured at approximately 89.12 seconds. This training duration is fully acceptable in real-world grid operations, where new topology integration is not a high-frequency event and adaptation can be prepared in advance. After completing offline BGL and CDA training, the final deployed model requires only a single forward pass to estimate all node voltage states. Based on the size of the model and the experimental setup, the typical single-pass inference time is around 40-60 milliseconds in GPU environments and can be maintained within 100-200 milliseconds on standard CPU hardware, well below the typical second-level refresh cycles required for estimating the state of the distribution network, thus fully satisfying engineering real-time deployment requirements.

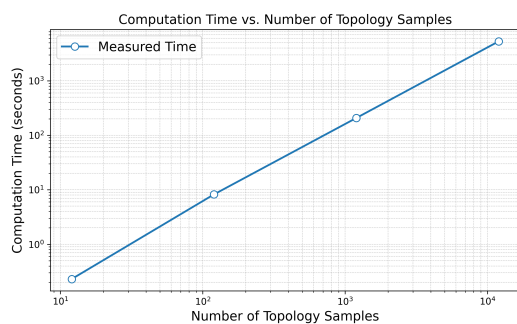


Fig. 7: Measured computation time versus number of historical topology samples.

G. Experiments on large-scale system

To prove the scalability of our model, we further conduct experiments on the IEEE 119-node system. In detail, based on the 33-node system, three photovoltaic (PV) units were added, all with the same capacity. The settings for

the pseudo-measurements and real-time measurements remain unchanged. Six topologies are utilized in our test, for each topology, the other five topologies are selected as historical data. The topology changes are as follows: Topology 2 disconnects 10-11 and connects 10-19; Topology 3 disconnects 27-28 and connects 27-51; Topology 4 disconnects 100-101 and connects 77-101; Topology 5 disconnects 10-11, 27-28, and connects 10-19, 27-51; Topology 6 disconnects 27-28, 100-101, and connects 27-51, 77-101. Every historical topology comprises 1000 samples of training data. As the scale of the data increases, the availability of real-time measurement data increases correspondingly. The learning-based methods can exhibit strong advantages in such tasks. Therefore we compared our proposed method with BPN and GP. All the historical topologies are applied for BP and GP to yield the results. The results are listed in Table. VII.

TABLE VII: MAEs of voltage magnitude and angle are obtained by different methods under different topologies.

Topology ID	Proposed		GP		BPN	
	Mag	Ang	Mag	Ang	Mag	Ang
2	3.51	9.7	4.16	11.1	5.04	12.0
3	3.65	6.8	3.92	6.5	5.44	8.1
4	4.33	7.0	5.69	9.6	5.46	8.4
5	7.42	11.9	13.42	17.4	11.16	14.5
6	6.85	10.7	8.37	12.2	13.44	14.9

From Table. VII, our proposed approach surpasses BP and GP in nearly every scenario. Besides, we can find that BP and GP can achieve similar performance with our method in topology 2-4. This is because a slight change in a large system may not lead to a significant performance degradation. The results verify the scalability of our proposed model. In other words, our proposed model can work well in a large-scale distribution system.

H. Practical Deployment and Limitations

In real-world power system operation, our method is designed to handle previously unseen and unknown topology configurations while performing accurate state estimation without requiring target-domain labels. When a new target topology is encountered, our approach does not simply rely on a static pre-trained model. Instead, it incorporates the new topology's measurement data together with all historical topology data into a joint unsupervised domain adaptation training process.

The advantage of this design is that it requires no labeled data from the target domain while leveraging historical experience from similar scenarios to support efficient knowledge transfer. Even if the new topology has never been seen before, as long as a reasonably similar historical topology exists in the library, the model can effectively utilize that source-domain information to improve prediction accuracy.

However, we acknowledge that this is fundamentally a similarity-based transfer strategy, which has limitations when the new topology is significantly different from all historical configurations. In such cases, the adaptation performance may be affected. **To further study this risk of negative transfer, we enlarged the source pool to 30 historical topologies and tested on five target topologies in IEEE-33. Results show that Top-1 selection sometimes suffers from mis-matched sources, while Top-3 significantly reduces the average MAE (from 11.3 to 9.2). In contrast, Top-5 degrades performance (average MAE increases to 12.2), since introducing too many less-related sources brings in noise and misleading patterns. In**

TABLE VIII: Target-domain voltage magnitude MAE ($\times 10^{-4}$) under different Top- k source selection strategies (IEEE-33, 30 historical sources).

Method	Topology 2	Topology 5	Topology 7	Topology 11	Topology 13	Avg.
Top-1 (nearest)	10.2	13.5	9.8	11.2	12.0	11.3
Top-3 (weighted)	8.6	10.4	8.3	9.0	9.8	9.2
Top-5 (weighted)	11.0	14.2	10.6	12.1	12.9	12.2

these experiments, the multiple sources in Top-3 and Top-5 were combined using a simple inverse-distance weighting scheme, i.e., $w_i = \frac{(d_i + \varepsilon)^{-1}}{\sum_{j=1}^k (d_j + \varepsilon)^{-1}}$ with $\varepsilon = 10^{-6}$. This confirms that using a small ensemble of sources (e.g., Top-3) improves robustness, whereas blindly adding more sources may aggravate negative transfer. The detailed results are reported in Table VIII. To mitigate this limitation, our method can be extended beyond selecting a single most similar source to instead leverage multiple top- k similar historical topologies through weighted fusion, enhancing robustness and generalization when similarity is limited.

Moreover, real-world power system operation provides a favorable setting for our method. As utilities continue to accumulate new topology scenarios and measurement data over time, the historical topology library can be continuously expanded and updated. This dynamic growth reduces the risk of encountering completely unmatched scenarios in the future and enables the model to achieve better adaptability and scalability in long-term practical deployment.

V. CONCLUSION

This paper proposes an unsupervised state estimation method for distribution networks under topology changes. We first design a bipartite graph learning (BGL) module with rank constraints to learn representations of both new and historical topologies, enabling the selection of the most similar historical topology based on Euclidean distance. Then, a causality-inspired sparse structure learning (CDA) network is used to extract invariant causal structures across topologies and predict state estimation results on the new topology without requiring its state variable labels. The experimental results on the IEEE 33-bus and 119-bus systems demonstrate that our method achieves consistently accurate state estimation under various topology change scenarios. Notably, it outperforms existing optimization-based and learning-based DSSE methods that rely on precise topology information or require retraining with new labeled data. The approach reduces dependence on new labeled data, improves adaptability to real-world reconfiguration events, and provides a promising direction for robust DSSE in practical distribution networks.

REFERENCES

- [1] W. Song, J. He, J. Lin, H. Ye, X. Ling, and C. Lu, "Bias analysis of pmu-based state estimation and its linear bayesian improvement," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 2, pp. 1607–1617, 2024.
- [2] J. Hu, W. Hu, D. Cao, S. Li, J. Chen, Y. Huang, Z. Chen, and F. Blaabjerg, "Robust multiarea distribution system state estimation based on structure-informed graphic network and multitask gaussian process," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 8, pp. 10 599–10 612, 2024.
- [3] U. Kuhar, M. Pantoš, G. Kosec, and A. Švigelj, "The impact of model and measurement uncertainties on a state estimation in three-phase distribution networks," *IEEE Transactions on Smart Grid*, vol. 10, no. 3, pp. 3301–3310, 2019.
- [4] Y. Zhu, X. Xu, and Z. Yan, "Accelerated matrix completion-based state estimation for unobservable distribution networks," *IEEE Transactions on Industrial Informatics*, vol. 20, no. 12, pp. 13 798–13 810, 2024.
- [5] F. C. Schweppe, "Power system static-state estimation, part iii: Implementation," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-89, no. 1, pp. 130–135, 1970.
- [6] Y. Zhang, R. Madani, and J. Lavaei, "Conic relaxations for power system state estimation with line measurements," *IEEE Transactions on Control of Network Systems*, vol. 5, no. 3, pp. 1193–1205, 2018.
- [7] C. Yuan, Y. Zhou, G. Liu, R. Dai, Y. Lu, and Z. Wang, "Graph computing-based wls fast decoupled state estimation," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2440–2451, 2020.
- [8] G. Cavararo, J. Comden, E. Dall'Anese, and A. Bernstein, "Real-time distribution system state estimation with asynchronous measurements," *IEEE Transactions on Smart Grid*, vol. 13, no. 5, pp. 3813–3822, 2022.
- [9] Z. Soltani, S. Ma, M. Khorsand, and V. Vittal, "Simultaneous robust state estimation, topology error processing, and outage detection for unbalanced distribution systems," *IEEE Transactions on Power Systems*, vol. 38, no. 3, pp. 2018–2034, 2023.
- [10] H. S. Karimi and B. Natarajan, "Joint topology identification and state estimation in unobservable distribution grids," *IEEE Transactions on Smart Grid*, vol. 12, no. 6, pp. 5299–5309, 2021.
- [11] Z. Xu, W. Jiang, J. Chen, R. Fu, and B. Weng, "Joint topology and state estimation using ttu monitoring data considering three-phase imbalance," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 10, pp. 9955–9967, 2023.
- [12] M. Huang and Z. Wei and J. Zhao and R. A. Jabr and M. Pau and G. Sun, "Robust Ensemble Kalman Filter for Medium-Voltage Distribution System State Estimation," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 7, pp. 4114–4124, 2019.
- [13] C. Xu and A. Abur, "A fast and robust linear state estimator for very large scale interconnected power grids," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4975–4982, 2018.
- [14] Xu, Junjun and Jin, Yulong and Zheng, Tao and Meng, Gaojun, "On state estimation modeling of smart distribution networks: a technical review," *Energies*, vol. 16, no. 4, p. 1891, 2023.
- [15] K. R. Mestav, J. Luengo-Rozas, and L. Tong, "Bayesian state estimation for unobservable distribution systems via deep learning," *IEEE Transactions on Power Systems*, vol. 34, no. 6, pp. 4910–4920, 2019.
- [16] E. Manitsas, R. Singh, B. C. Pal, and G. Strbac, "Distribution system state estimation using an artificial neural network approach for pseudo measurement modeling," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 1888–1896, 2012.
- [17] P. N. P. Barbeiro, J. Krstulovic, H. Teixeira, J. Pereira, and J. P. Iria, "State estimation in distribution smart grids using autoencoders," in *2014 IEEE 8th International Power Engineering and Optimization Conference (PEOCO)*, 2014.
- [18] N. M. Manousakis and G. N. Korres, "Application of State Estimation in Distribution Systems with Embedded Microgrids," *Energies*, vol. 14, no. 23, p. 7933, 2021.
- [19] Azimian, Behrouz and Biswas, Reetam Sen and Moshtagh, Shiva and Pal, Anamitra and Tong, Lang and Dasarathy, Gautam, "State and topology estimation for unobservable distribution systems using deep neural networks," *IEEE transactions on instrumentation and measurement*, vol. 71, pp. 1–14, 2022.
- [20] H. Li, Y. Weng, V. Vittal, and E. Blasch, "Distribution grid topology and parameter estimation using deep-shallow neural network with physical consistency," *IEEE Transactions on Smart Grid*, vol. 15, no. 1, pp. 655–666, 2024.
- [21] H. Li, Z. Ma, and Y. Weng, "A transfer learning framework for power system event identification," *IEEE Transactions on Power Systems*, vol. 37, no. 6, pp. 4424–4435, 2022.
- [22] D. Cao, J. Zhao, W. Hu, Q. Liao, Q. Huang, and Z. Chen, "Topology change aware data-driven probabilistic distribution state estimation based on gaussian process," *IEEE Transactions on Smart Grid*, vol. 14, no. 2, pp. 1317–1320, 2023.
- [23] X. Pang and J. Gill, "Spike and slab prior distributions for simultaneous bayesian hypothesis testing, model selection, and prediction, of nonlinear outcomes," *political analysis*.
- [24] L. Pagnier and M. Chertkov, "Physics-informed graphical neural network for parameter & state estimations in power systems," *arXiv preprint arXiv:2102.06349*, 2021.
- [25] D. Cao, J. Zhao, W. Hu, N. Yu, J. Hu, and Z. Chen, "Physics-informed graphical learning and bayesian averaging for robust distribution state estimation," *IEEE Transactions on Power Systems*, 2023.

- 1
2 [26] M. Tschannen, O. Bachem, and M. Lucic, "Recent advances
3 in autoencoder-based representation learning," *arXiv preprint*
4 *arXiv:1812.05069*, 2018.
- 5 [27] Z. Kang, C. Peng, and Q. Cheng, "Twin learning for similarity and
6 clustering: A unified kernel approach," in *Thirty-First AAAI Conference*
7 *on Artificial Intelligence*, 2017.
- 8 [28] F. R. Chung and F. C. Graham, *Spectral graph theory*. American
9 Mathematical Soc., 1997, no. 92.
- 10 [29] Z. Kang, C. Peng, Q. Cheng, X. Liu, X. Peng, Z. Xu, and L. Tian,
11 "Structured graph learning for clustering and semi-supervised classification," *Pattern Recognition*, vol. 110, p. 107627, 2021.
- 12 [30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N.
13 Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
- 14 [31] A. F. T. Martins and R. F. Astudillo, "From softmax to sparsemax:
15 A sparse model of attention and multi-label classification," *JMLR.org*,
16 2016.
- 17 [32] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by
18 jointly learning to align and translate," in *International Conference on*
19 *Learning Representations*, 2014.
- 20 [33] R. Cai, J. Chen, Z. Li, W. Chen, K. Zhang, J. Ye, Z. Li, X. Yang, and
21 Z. Zhang, "Time series domain adaptation via sparse associative structure
22 alignment," in *Proceedings of the AAAI Conference on Artificial*
23 *Intelligence*, vol. 35, no. 8, 2021, pp. 6859–6867.
- 24 [34] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep
25 domain confusion: Maximizing for domain invariance," *arXiv preprint*
26 *arXiv:1412.3474*, 2014.
- 27 [35] Y. Ganin and E. Ustinova and H. Ajakan and P. Germain and H.
28 Larochelle and F. Laviolette and M. Marchand and V. Lempitsky,
29 "Domain-Adversarial Training of Neural Networks," *Journal of Ma-*
30 *chine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016.
- 31 [36] Liang, Jian and Hu, Dapeng and Feng, Jiashi, "Do we really need
32 to access the source data? source hypothesis transfer for unsupervised
33 domain adaptation," in *International conference on machine learning*.
34 PMLR, 2020, pp. 6028–6039.
- 35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60