A Transfer Learning Framework for Power System Event Identification

Haoran Li, Student Member, IEEE, Zhihao Ma, Student Member, IEEE, and Yang Weng, Senior Member, IEEE

Abstract—The increasing uncertain components of power systems foster the wide applications of Machine Learning (ML) techniques. While traditional ML models demand a large set of data, data-scarce dilemmas exist for new meters, devices, and new grids. Further, for rich historical measurements, valuable data may still be limited, especially for targets like identifying system events that rarely occur in the power system. To enhance the event type differentiation and localization for a data-limited grid, we propose a Transfer Learning (TL) framework to transfer knowledge from a data-rich grid (source grid) to the target grid, using measurements from Phasor Measurement Units (PMUs). The transferring process is challenging because of (1) high-volume data with redundant information, (2) different measurement dimensionalities, (3) dissimilar data distributions, and (4) disjoint event-location-label spaces for two grids. To handle the challenges of (1) to (3), we propose a joint optimization to reduce dimensionality and maximize common knowledge in a shared low-dimensional feature space, where the commonality lies in the same dimensions and close data distributions. Such an optimization-based procedure is verified via rigid mathematical theorems given the same label space, i.e., event-type-label space. However, for event localization, challenge (4) obstructs the optimization. Therefore, we design a label space alignment method to relabel the event location by the event zone location and build an event zone estimation problem. Then, the framework is generalized to both tasks. Finally, comprehensive experiments demonstrate the advantages of the proposed methods over state-of-the-art transfer learning models.

Index Terms—Event identification, power systems, limited data, transfer learning, dimensionality reduction, distribution adaptation, label transfer.

I. INTRODUCTION

POWER systems are dramatically integrating highly uncertain components for cleaner, more efficient, and lower-cost energy generations and consumption. For example, renewable energies like Photovoltaic (PV) and wind power depend largely on the changing weather. The charging of Electric Vehicles (EVs) also brings randomness to the electric grid. This unprecedented system evolution, despite promising values, deteriorates traditional system monitoring and prevents the achievement of situational awareness.

To enhance the system monitoring under changing operating points, Machine Learning (ML) methods are more and more popular due to their high capacity of extracting informative features from dynamic data streams. This advantage is especially expanded with the increasing deployment of Phasor Measurement Units (PMUs) to provide high-resolution phasor data. PMU-based ML models achieve great success in state estimation [1], event identification [2]–[4], resilience improvement [5], and cyber-attack detection [6], etc. These applications, however, strictly assume that there are enough data for training.

Such an assumption is easily violated in the following scenarios. 1) For new grids, data scarcity prevents ML model training. 2) For old grids with new meters or components (e.g., lines, generators, and loads), the data dimensionality and distributions change after the new construction. Thus, old ML models will fail, while new models are hard to train with limited data. 3) Even for stable grids with long-term operations and few events, the number of event labels is limited. Thus, directly training of a Machine Learning (ML) model is hard due to the limited labels. For example, the fault data is intrinsically limited for a highly reliable grid. In general, we demand new principles to fundamentally change the dilemma.

Therefore, we propose to utilize Transfer Learning (TL) for knowledge migration from a data-rich grid (source grid) to a data-limited grid (target grid). TL employs well-established feature extraction techniques to obtain common features for enhancing the decision-making in the target domain [7]. This approach promotes many TL applications to fields like object localization [8], image classification [9], video pattern recognition [10], sentiment analysis [11], and vehicle routing detection [12], etc. For power systems, TL also has few applications on dynamic security assessment [13], load file generation [14] and event forecasting [15]. However, these methods are restricted to a small scope and have limited theoretical foundations. In this paper, we aim to propose a general TL framework for power system transfer learning with both strong theoretical foundations and high generalizability. Notably, although we focus on event type differentiation and localization, these two tasks are representative for situations of data space homogeneity/heterogeneity and label space identicalness/disjointing.

More specifically, we assume two grids can have either the same or different number of PMUs for homogeneous and heterogeneous measurement spaces [16], respectively. As for the label space, two grids share the same event-type set but completely

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/
different event-location sets. For example, the location label of a line trip for one grid doesn’t have any physical meanings for the other grid. Taking all the above characteristics into consideration, we try to answer the following questions. (i) Can we propose a unified framework to incorporate datasets with the same or different dimensionalities and minimize the knowledge transfer error? (ii) Can we bridge two disjoint label spaces for physically meaningful label transfer?

For question (i), we focus on heterogeneous TL and treat the homogeneous scenario as one special case. To align the dimensionality, existing work aims to find proper projections. [17] find the mapping from the source to the target by minimizing the Euclidean distance. This projection is inefficient and inaccurate for power system datasets since a) the Euclidean distance is an inaccurate measure to evaluate the data distribution difference [16] which implies that there is a distance between the distribution of the source and the target spaces. For example, for two power systems, the PMU measurements usually have certain distribution distances due to different operating conditions.

For adaptation distribution difference, Domain Adaptation (DA) in TL minimizes distribution discrepancy for the source and the target domains to obtain the common knowledge, which is theoretically sound and computationally inexpensive [7]. For example, [18] improves the problem a) by introducing Entropic Gromov-Wasserstein Discrepancy for a better evaluation but still meets the problem of b). [19] utilizes the so-called Maximum Mean Discrepancy (MMD) [20] as the measure for problem a). For problem b), Principal Component Analysis (PCA) is employed to pre-process the target domain’s data, thus making the conversion from the source to the target to be an efficient dimensionality reduction process. However, PCA fixes the mapping for the target grid to a subspace, thus reducing the capacity of finding common knowledge.

To increase the capacity, we propose to find two projections for both the source and the target data to a common low-dimensional space. Within this space, we utilize MMD to measure the distribution divergence. Further, to support the proposed method, we exhibit strong mathematical foundations for the optimization formulation and the introduction of MMD measure. After projections, any ML models can be further trained using the transformed data to perform tasks in the target grid as long as the label spaces are the same. For disjoint label spaces in question (ii), a label-transfer process is required. We find that usually, the disjoint labels indicate a local change to the system, e.g., the label for a local event. These changes have higher impacts on PMUs that are closer to the change location. Thus, we can select a fixed number of representative PMUs for two grids to relabel the local change. Consequently, disjoint label spaces are mapped to one common label space for the label transfer, thus complementing the total TL framework.

The complete framework is illustrated in Fig. 1. We aim to transfer knowledge from a long-operated source grid to a new target grid, which boosts the performance of an ML model to identify events in the target grid. To capture event patterns, we utilize PMU data with high resolutions and labels to train a supervised ML model like a Deep Neural Network (DNN). However, as shown in the middle of Fig. 1, the measurement spaces of the source and the target system (1) have different distributions due to different operating conditions, and (2) have different dimensions that cause heterogeneity. Further, though two systems can have shared event type labels, they own (3) disjoint event-location label spaces. Thus, traditional ML models fail to utilize all the information together, and we propose a heterogeneous Transfer Learning (TL) framework to tackle these problems. Specifically, as shown in the middle of Fig. 1, the data transfer process handles (1) and (2) via finding the common feature space with the same dimensions and similar feature distributions from the source and the target systems. Thus, these features can be input to the ML model (e.g., the DNN model at the right of Fig. 1) for training. Secondly, the label transfer process deals with (3) via categorizing event locations into zones that have the same number and relative locations (e.g., Northwest and Southeast, etc.) in the source and the target grid. Then, zone indices are treated as labels. Finally, the right part of Fig. 1 illustrates that the event type or event zone labels are the output of the ML models for training and predictions. In general, we have the following contributions:

- We formalize the problem of power system TL with a thorough consideration of the data-space and label-space differences. With domain knowledge, we show that though data/label differences exist, there is shared knowledge, and positive transfer can always happen.
- We propose a general framework to minimize the differences with strong mathematical guarantees for highly efficient knowledge transfer.
- We will conduct extensive experiments to demonstrate the high performance of our framework with datasets from different systems.
Remark: to the best of the author’s knowledge, our paper is the first to consider both the data space heterogeneity and label space non-overlaps. Primarily, we combine the domain knowledge of event responses and rigorous mathematical proofs to build the model. With extensive experiments, we show the effectiveness of our proposed models.

The rest of the paper is organized as follows: Section II formalizes the problem. Section III demonstrates the data transfer process. Section IV illustrates the label transfer approach. Section V evaluates the performance of the proposed framework, Section VI adds discussions for the future work and Section VII concludes the paper.

II. PROBLEM DEFINITION

In this paper, the goal is to train a classifier that can analyze a period of PMU streams and output the underlying event type and location. Note that some smart meter data with high resolutions and qualities [21]–[23] can also be utilized as long as the measurements can capture event dynamics. However, in this paper, we focus on PMU data. To prepare the input data, we employ a moving window to segment PMU streams. As shown in Fig. 2, the window has fixed width and moving gap for the source and the target measurements. Then, each time a small period of PMU streams and output the underlying event type is similar [30]. Then, the aggregation of them in a system can bring transfer positivity for most of the power systems. (1) for different PMUs numbers in two grids. However, our models can also handle the case when \( d_S = d_T \). Then, let \( y \) and \( \tilde{y} \) be the label variable of the input vectors for the source and the target grid. The label can be the normal status, event types, and event location label variable of the input vectors for the source and the target grid, respectively. Note that

\[
\{I_1, I_2, \ldots, I_h\} \cap \{J_1, J_2, \ldots, J_l\} = \emptyset \text{ for two disconnected power networks.}
\]

After the moving window segmentation, we denote \( \{(x_i, y_i)_{i=1}^{n_S}\} \) and \( \{(\tilde{x}_i, \tilde{y}_i)_{i=1}^{n_T}\} \) as a set of labeled training instances from the source grid and the target grid, respectively, where \( n_S \) and \( n_T \) are the number of samples for the source and the target grids, and we assume \( n_S \gg n_T \). With above notations, we have the following problem formulations

- **Problem:** PMU-based Transfer learning for event type differentiation and event zone estimation.
- **Given:** a set of PMU-based samples with labels \( \{(x_i, y_i)_{i=1}^{n_S}\} \) and \( \{(\tilde{x}_i, \tilde{y}_i)_{i=1}^{n_T}\} \) from the source and the target grids, respectively.
- **Output:** a classifier \( f \) to distinguish event types or identify event locations in the target grid.

The defined problem with high \( d_S \) and \( d_T \) forces us to find a cost-efficient data transfer process. We show in the next section that this procedure is achieved via finding a common low-dimensional feature space.

III. DATA TRANSFER: PROJECTION TO LOW-DIMENSIONAL FEATURE SPACE

A. Validation of the Transfer Positivity

Before elaborating on data transfer details, a more fundamental issue should be first figured out: will the data transfer be positive to enhance rather than deteriorate the ML model training for the target grid? Namely, we want to avoid the so-called negative transfer [24]. To illustrate the positivity of our transfer learning, we start from physics. Generally, as shown in Fig. 3(a), we utilize simulated data from Illinois 200-bus system [25] and South Carolina 500-bus system [26], and realistic data from Western Electricity Coordinating Council (WECC) system [27] and our utility partner in Arizona to demonstrate the Generator Trip (GT) and Line Fault (LF) events. Note that we have another IEEE 14-bus simulation system [28] and its event performances are similar. Due to space limit, we don’t show this system’s event curves. Further, with geographical information of the 200-bus and 500-bus systems, we utilize Fig. 3(b) to demonstrate LF event impacts on different PMUs from event centers to edges. Thus, we show the similarity of the event geographical propagation.

Specifically, we have the following statements to support transfer positivity for most of the power systems. (1) for different systems, the dynamical behaviors of each electric device are similar [29]. Then, the aggregation of them in a system can bring similar dynamics in response to a certain event. (2) the protection and response mechanism to a specific event is similar [30]. For example, for GT events, the inertial and governor responses help to stabilize the frequency, leading to similar behaviors of frequency for the first 3 curves in Fig. 3(a). Based on (1) and (2), we observe similar behaviors of the GT and LF in simulated data. For realistic data, we find that though they have disturbances, they have similar patterns of the curve change tendency after a specific event.
For event zone estimation, we claim that the common knowledge lies in the event impact propagation from the event locations to other areas. Specifically, we present the voltage magnitude drop after an LF event for 200- and 500-bus systems, as shown in Fig. 3(b). The impact of the event decreases as the location is further and further away from the event center. Such a propagation, based on the system damping characteristic, is general for different systems. Finally, the red dotted block in Fig. 3(a) and the sharp decreasing in the left part of Fig. 3(b) illustrate that the shared knowledge lies in the low-dimensional space compared to the ultra high-dimensional input data. Thus, we propose the following assumption that lays the foundations for further transfer learning.

**Assumption 1:** There exists a low-dimensional feature space that is extracted from the source and the target data spaces and represents the common knowledge to the system event.

**Remark:** There may be some special systems that lose the event similarity compared to most of the power systems, leading to negative transfer [24]. However, in this paper, we focus on the potentials of transfer learning for most of the standard systems and the method design. The investigation and quantification of negative transfer on power systems can be treated as a future topic.

### B. Data Transfer Objective Formulation

Based on Assumption 1, we propose to project \( \mathbf{x} \) and \( \tilde{\mathbf{x}} \) to the \( d \)-dimensional latent feature space with linear transformations \( \mathbf{T} \in \mathbb{R}^{d \times d_S} \) and \( \mathbf{T} \in \mathbb{R}^{d \times d_T} \), where \( d \ll d_S \) and \( d \ll d_T \). Namely, we have

\[
F(\mathbf{x}) = \mathbf{T}_S \mathbf{x}, \quad \tilde{F}(\tilde{\mathbf{x}}) = \mathbf{T}_T \tilde{\mathbf{x}}.
\]  

(1)

We can add kernel mapping to \( \mathbf{x} \) and \( \tilde{\mathbf{x}} \) for non-linearity. However, we find the linear mapping is generally good enough in testing, and adding kernels don’t bring significant improvements. In the following derivations, we focus on the linear mapping.

The optimal \( \mathbf{T} \) and \( \tilde{\mathbf{T}} \) ultimately help to train a good classifier for event identification in the target grid. Namely, we want the conditional probabilities are approximately equal:

\[
P(y|F(x)) \approx P(\tilde{y}|\tilde{F}(\tilde{x})),
\]  

(2)

where \( P(\cdot) \) represents the probability or probability density for the random variables. (2) indicates that the decision boundaries of the source and the target for each label are close. Thus, we can utilize all the feature data to train the classifier.

**Remark:** (2) as the goal can only be proposed and analyzed when the label spaces for the source and the target grid are the same, e.g., label space for event types. For disjoint label spaces for event locations, we propose to first align the label space in the next section. Then, all the derivations based on (2) can be continued. For the convenience of later derivations, we still assume the final common label space includes \( K \) unique labels.

To further understand the objective in (2), we rewrite the equation according to Bayes’ rule:

\[
P(F(x)|y) * P(y) \over \sum_i P(F(x)|y = i) * P(y = i)
\]  

\[
\approx P(\tilde{F}(\tilde{x})|\tilde{y}) * P(\tilde{y}) \over \sum_i P(\tilde{F}(\tilde{x})|\tilde{y} = i) * P(\tilde{y} = i).
\]  

(3)

Equation (3) shows that for the source and the target grids’ data, if \( P(F(x)|y) \approx P(\tilde{F}(\tilde{x})|\tilde{y}) \) and \( P(y) \approx P(\tilde{y}) \), (2) can be achieved. The former equation inspires a minimization of the conditional probability of features given labels, but the latter condition is generally not true. For instance, if the source grid is much older than the target grid, the probability of an event is generally higher. Luckily, based on the probability theory, we have the following equations.

\[
P(F(x)) = \sum_i P(F(x)|y = i) * P(y = i),
\]  

\[
P(\tilde{F}(\tilde{x})) = \sum_i P(\tilde{F}(\tilde{x})|\tilde{y} = i) * P(\tilde{y} = i).
\]  

(4)

Fig. 3 Illustrations of the existence of shared knowledge for event type differentiation and localization.
Therefore, guaranteeing $P(F(x)) \approx P(\hat{F}(\bar{x}))$ and $P(F(x)|y) \approx P(\hat{F}(\bar{x})|\bar{y})$ together can lead to the final goal of (2). Based on this observation, we can jointly minimize the distance between $P(F(x))$ and $P(\hat{F}(\bar{x}))$ and the distance between $P(F(x)|y)$ and $P(\hat{F}(\bar{x})|\bar{y})$. Namely, we have the following optimization problem:

$$
\min_{T, \bar{T}} \sum_{y, \bar{y}} D(Tx, \bar{T}\bar{x}|y = \bar{y}) + \eta D(Tx, \bar{T}\bar{x}) + \lambda \|T\|^2_F + \lambda \|\bar{T}\|^2_F,
$$

where $D$ represents the distance of two distributions and $\eta$ is a hyper-parameter to weight two objectives and avoid numerical issues. $\lambda$ is the penalty term and $\|\cdot\|_F$ represents the Frobenius norm. The last two terms help to prevent overfitting. We term our optimization in (5) as Heterogeneous Joint conditional and Marginal Distribution Adaptation (HJDA).

Remark: HJDA is different from Joint Distribution Adaptation (JDA) in [20], which utilizes the same projection for two domains. HJDA considers different projection matrices $T$ and $\bar{T}$ since (1) nodes in two grids are misaligned and can’t be treated as one input variable and (2) cross-grid transfer learning is often heterogeneous, i.e., $d_S \neq d_T$.

There are many measures for probability distance $D$ in (5). The most popular one is a nonparametric distance estimate: Maximum Mean Discrepancy (MMD) for its simplicity in calculations and solidness with mathematical foundations [19], [20], [31]. In general, MMD evaluates the sample distribution difference over any unit-ball function class in the characteristic Reproducing Kernel Hilbert Spaces (RKHS). Thus, the MMD metric has two distinguished properties: (1) the unit-ball function class is rich enough to make MMD powerful to represent the distribution difference. (2) RKHS brings fast convergence of the estimate MMD to the true value as the sample size increases. One can find more strict mathematical proofs of MMD’s properties in [31]. Based on the definition of MMD, we specify the optimization problems as follows.

$$
\min_{T, \bar{T}} \sum_{y, \bar{y}} D(Tx, \bar{T}\bar{x}|y = \bar{y}) + \eta D(Tx, \bar{T}\bar{x}) + \lambda \|T\|^2 + \lambda \|\bar{T}\|^2,
$$

where $n_S^{(k)}$ and $n_T^{(k)}$ represent number of samples for label $k$, where $1 \leq k \leq K$.

C. Data Transfer Optimization Algorithm

We can utilize gradient descent method to solve the optimization. The gradient with respect to $T$ is as follows.

$$
\nabla_T = \sum_k \left( \frac{4}{n_S^{(k)}} \sum_{i,j=1}^{n_T^{(k)}} T x_i (x_j) + \frac{2}{n_S^{(k)} n_T^{(k)}} \sum_{i=1}^{n_T^{(k)}} \sum_{j=1}^{n_T^{(k)}} \bar{T} \bar{x}_j (x_i) \right) + \frac{4\eta}{n_S^{(k)}} \sum_{i,j=1}^{n_T^{(k)}} T x_i (x_j) + 2\lambda T,
$$

$$
= \sum_k \left( T X_k M_k X_k^\top - \bar{T} \bar{X}_k N_k \bar{X}_k^\top \right) + \eta \left( T X M_0 X^\top - \bar{T} \bar{X} N_0 \bar{X}^\top \right) + 2\lambda T,
$$

where $X$ is the stack of all source samples, i.e., $X = [x_1, \ldots, x_{n_S}]$, and $X_k$ is the stack of source samples with label $k$. The same stacking happens on $\bar{X}$ and $\bar{X}_k$ using target samples. Each element for $M_k$, $N_k$, $M_0$, and $N_0$ is $4/(n_S^{(k)})^2$, $2/(n_S^{(k)} n_T^{(k)})$, $4/n_S^2$, and $2/(n_S n_T)$, respectively. Similarly, we can calculate $\nabla_{\bar{T}}$ by symmetry. Finally, we solve the optimization by iteratively updating $T$ and $\bar{T}$ using the calculated gradients.

IV. LABEL TRANSFER: LABEL ALIGNMENT VIA PMU INDICES

While HJDA in Section III guarantees a well-established data transfer process with solid mathematical explanations, this procedure can only handle the identical label spaces for the source and the target grid. For the event zone estimation problem, however, the label spaces are disjoint. Thus, we propose a label alignment approach to achieve the cross-system label transfer.

As shown in Fig. 3(b), system event has a decreasing impact from the event center to further areas. Thus, the event location can be reflected by the PMU locations. Therefore, we can select some responsible PMUs, each of which aims at monitoring potential events in one local area. For example, in Fig. 4, each system can be divided into 6 zones, i.e., the Northwest, North, Northeast, Southwest, South, and Southeast areas for zone $1 \sim 6$.

We divide the zone based on the similarity measure of event impacts to responsible PMUs. Namely, we select the responsible PMUs as the event zone centers. Then, the event impact similarity helps to identify event zone boundaries, leading to the event zone division. Specifically, we have the following steps.

(i) we choose the event zone center via selecting responsible PMUs. In general, the selection is based on the prior knowledge of the system, including the system topology, the factors to
determine the event frequency like loads, generation reserves, environments, and device ages, and the availability of PMUs, etc. In our paper, we focus more on demonstrating the potentials of event zone divisions for Transfer Learning. Thus, we assume the responsible PMUs are evenly distributed among the system. Then, for both the source and the target systems, we pick up the same number of responsible PMUs, thus bringing the same number of event zones. Mathematically, the indices of the responsible PMUs (i.e., 1, 2, ..., M for M responsible PMUs) can form the common set of event zone labels for two grids.

(ii) we identify the event zone boundaries via measuring the similarity of event impacts on the responsible PMUs. Mathematically, the problem can be solved via finding the PMU that suffers the largest distribution change before and after a certain event. Thus, we utilize Maximum Mean Discrepancy (MMD) for evaluations. Specifically, for the \(j^{th}\) PMU in the source or target grid, let \(w_j^i\) be the sample at the \(i^{th}\) time slot. Then, we can calculate the MMD distance of samples before and after an event:

\[
D_j = \left\| \frac{1}{n_0} \sum_{i=1}^{n_0} w_j^i - \frac{1}{n_1} \sum_{k=n_0}^{n_0+n_1} w_j^k \right\|^2,
\]

where \(n_0\) and \(n_1\) are the total number of time slots before and after the event, respectively. Therefore, the MMD value \(D_j\) represents the corresponding event’s impacts on PMU \(j\). Using (8) to calculate one event’s impacts on all the responsible PMUs, we can rank and find the mostly-affected PMU as the true responsible PMU for the specific event.

The complete process essentially relabels one event with its responsible PMU index (i.e., the zone index), and the new label with the categorized geographical information can be directly applied to another grid. Finally, though we only utilize a subset of PMUs for relabeling, all PMU measurements can be inputted into the data transfer framework. This is because in the data transfer process, we have two projections to maximize the common features. This data-driven process will automatically decide which PMU is important by weighting the PMU values using the transformation matrices.

Remark: in our label transfer process, we have a fixed number of event zones (i.e., \(M\) event zones for \(M\) responsible PMUs) for the source and the target systems. However, we can do the fine-grained selection by the prior knowledge of the system, including the system topology, the factors to determine the event frequency like loads, generation reserves, environments, and device ages, and the availability of PMUs, etc. Then, the zone numbers can be different for the source and the target grids. In future work, we will extend the topic for Transfer Learning under overlapped but not the same zone label spaces.

V. EXPERIMENT

A. Dataset Description

To diversify the testing scenarios, we utilize IEEE 14-bus system [28], Illinois 200-bus system [25], and South Carolina 500-bus system [26] to simulate events. The simulation platform is based on a business-level software Positive Sequence Load Flow (PSLF) [32] from General Electric (GE) company. For each system event simulation, we change the loading conditions and event locations to demonstrate the robustness of our models. Table I displays all event types in our simulations. Further, as shown in Fig. 5, we visualize the event using Voltage Magnitude (VM) data obtained from PSLF. The data are randomly selected from 10 PMUs in the 200-bus system. As for the plots, similar event patterns can be found in [33], which demonstrates the correctness of the simulation. Subsequently, we generate electric measurement streams with 30 samples per second to mimic the realistic PMU streams, including voltage magnitude, voltage angle, current magnitude, current angle, and frequency. To simulate the data-rich source grid and the data-limit target grid, we consider \(n_S = 8\) and \(n_T = 50\) to test different scenarios, where we add noise and vary PMU penetration levels. Secondly, we conduct sensitivity analysis with respect to the data points. We denote \(N = n_S + n_T\) and vary \(N \in \{400, 450, 500, 550, 600, 650, 700\}\) and keep the ratio \(n_S \over n_T = 8\).

We add white Gaussian noise with signal-to-noise-ratio (SNR) to be 125 to mimic realistic PMU noise [34]. Then, we
Fig. 5. The visualization of five types of events in experiments.

Fig. 6. The visualization of PMU locations for the 200-bus system.

denote $\gamma \in \{0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$ as the PMU penetration level to evaluate the model performance with different PMU numbers. For example, for the 200-bus system, $\gamma = 0.05$ implies that we randomly select 10 PMUs to prepare the training data. For this case, Fig. 6 visualizes the PMU locations in the 200-bus system, where the red and green circles represent the nodes with generators and loads, respectively. The circle size represents the relative generation/load values. Finally, we add labels from PMU 1 to PMU 10 for the PMU locations. For the event zone estimation problem, the selected PMUs should cover the 6 monitoring regions in Fig. 4 so that we can form the common zone-index space for label transfer. Finally, we denote $A \rightarrow B$ as transfer learning from source grid $A$ to target grid $B$.

B. Model Description and Implementation Details

To comprehensively demonstrate the high-performance of our models, we have implemented the following models for comparison.

- Principal Component Analysis (PCA) + Resnet [35]. This is a benchmark method to train a classifier without transfer learning. PCA aims to reduce the dimensionality of the PMU data, and Deep Residual Network (Resnet) is the classifier for event type differentiation and event zone estimation. The reason for using Resnet is that Resnet has excellent performances in various ML tasks [36]–[38].
- Heterogeneous Joint Distribution Adaptation (HJDA) + Resnet. We can use our proposed HJDA to convert source and target data to a common feature space. Then, Resnet can be employed to train classifiers based on the common features.
- Transfer Component Analysis (TCA) [39] + Resnet. TCA also employs MMD measure to find projections to a latent space from the source and the target measurement spaces. However, the objective only considers minimizing the marginal data distribution. Then, Resnet can be employed to train the features in the latent space.
- Cross-Domain Landmark Selection (CDLS) [19] + Resnet. The supervised version of CDLS finds the optimal matrix to project the source data to the PCA-based subspace of the target data, where MMD is also employed as the objective for both the marginal and the conditional distributions. With transferred datasets, Resnet can be employed to train the classifier.

All the above methods can be utilized to event zone estimation problems with the help of our proposed label transfer process.

We utilize 3-fold cross-validation to fine-tune the hyper-parameters and obtain the testing accuracy for the final classifiers. Notably, the hyper-parameters of Resnet should stay the same: i.e., the same hidden layers, number of neurons, activation function, learning rate, and so on. Thus, we create fair comparisons among different transfer learning distribution adaptation methods.

C. Transfer Learning-Based Event Type Differentiation: Highly-Accurate Performance of the Proposed HJDA

Fig. 7(a) to Fig. 7(f) illustrate the results of power system event identification using different methods. We conduct comprehensive experiments for every two of the synthetic systems and with changing $\gamma$. For each scenario, we find that our proposed HJDA (red box) performs the best over other methods. Specifically, the average testing accuracy among all scenarios is 81.7%, 76.7%, 73.6%, and 64.6% for HJDA, CDLS, TCA, and PCA, respectively. This proves the general high performance of our method. Further, we have the following analyses to verify the principles of our transfer learning framework.

Positive transferability. Compared to the PCA method (grey box) without transfer learning, other 3 transfer learning methods have a significant improvement for the testing accuracy, which demonstrates the positive transferability of the power system PMU data for event type differentiation. As we discussed in Section III-A, the positive effect of transfer learning exists for power systems, which is highly supported by our results.

HJDA vs. TCA. In each trial, HJDA has significant accuracy increasing compared to TCA (shallow green box). The average line in the box plot for HJDA is always higher than TCA, and the box plots for these two methods do not have overlaps. The reason is that TCA only focuses on minimizing the distance of marginal data distributions in (4). However, this does not mean the general goal of conditional distribution in (2) can be
minimized. Especially, if the label distributions are different for two systems, equations (3) and (4) indicate that conditional distributions in (2) cannot match. In our simulation, we randomly implement the event, so the label distribution for the source and the target is not the same. Thus, TCA does not perform well. On the other hand, HJDA has a thorough consideration of the joint marginal and conditional distributions, thus obtaining a much better performance.

**HJDA vs. CDLS**. Compared to the CDLS method (blue box), HJDA has a big improvement when we test transfer learning between 14-bus and 500-bus systems, and a minor improvement when we test transfer learning between 14-bus and 200-bus system or 200-bus and 500-bus system. We have the following explanations for our observations. (i) CDLS method employs PCA to reduce the dimensionality of the target data and finds the optimal projection from the source space to the PCA-based target space with the same objectives of our methods. Thus, our method wins with the ability to search for a better target projection rather than a fixed PCA transformation. Especially, the possibility of finding a better projection is promoted under the event label supervision. (ii) the data distribution difference is largest between 14-bus and 500-bus system, and it is vital to utilize a better projection for the target grid. Thus, HJDA performs much better for transfer learning in 14-bus and 500-bus systems. (iii) since CDLS employ the same objectives for optimization with strong mathematical supports, CDLS can report good results for other transfer learning scenarios when the source and the target grids’ data have relatively smaller distribution difference.

**Performance with respect to system size**. Intuitively, a larger system seems to have richer information and can bring a better result. However, we observe an interesting and counter-intuitive phenomenon: using a smaller-size system as the source grid can
achieve a better performance than treating it as the target grid. This is especially observed for transfer learning between 14-bus and 200-bus/500-bus systems.

The reason is that when we minimize the conditional distribution given labels, i.e., the first sum term in (5), using an integer to represent a label with high complexity may cause troubles. More specifically, for one specific event type, different event locations and magnitudes essentially demand a more careful event categorization in minimizing the conditional distribution differences. For example, the active power of generators for the 14-bus system has less variability than that for the 500-bus system, making the data variance of the 500-bus system larger but the same label "generator trip". Therefore, HJDA suffers an over-minimization that compresses too much information in the 500-bus system to match the distribution in the 14-bus system. We treat the 500-bus system as the source grid, more data are used and over-compressed, thus deteriorating the performance. To mitigate this issue, methods like label complexity analysis, regularization, or non-linear domain adaptation can be utilized, and we will explore these methods in the next paper.

Table II: The Testing Accuracy (% (Mean ± Standard Deviation) of Different Methods + Resnet for Event Zone Estimation

|                | HJDA  | CDLS  | TCA   | PCA   |
|----------------|-------|-------|-------|-------|
| 14 → 200       | 80.2 ± 4.4 | 74.4 ± 6.2 | 68.7 ± 3.5 | 60.6 ± 4.3 |
| 200 → 14       | 73.6 ± 3.6 | 66.5 ± 5.8 | 64.4 ± 4.7 | 56.3 ± 7.8 |
| 14 → 500       | 76.5 ± 5.4 | 73.5 ± 3.3 | 69.7 ± 4.3 | 58.3 ± 8.4 |
| 500 → 140500   | 70.9 ± 2.4 | 68.4 ± 3.3 | 60.6 ± 6.9 | 56.3 ± 5.6 |
| 200 → 500      | 75.0 ± 6.7 | 68.3 ± 4.9 | 70.0 ± 6.7 | 58.3 ± 6.4 |
| 500 → 200      | 72.6 ± 4.1 | 69.7 ± 2.5 | 64.6 ± 5.0 | 60.6 ± 5.9 |

D. Transfer Learning-Based Event Zone Estimation: Relabeling Technique Enable Positive Transfer and High Accuracy of the Proposed HJDA

For event zone estimation, we divide each system into 6 zones to align the label, as shown in Fig. 4. For each zone, we assume there is a PMU (i.e., the responsible PMU) for monitoring. However, as we discussed, the grid division can be flexible according to the availability of PMUs. Then, we conduct line fault for different lines that are uniformly distributed in the monitoring regions. Subsequently, for each line-fault event, we calculate its responsible PMU based on the ranking of the calculated MMD value in (8). If the $k^{th}$ ($1 \leq k \leq 6$) responsible PMU has the largest MMD value, we relabel the certain line fault as $k$.

After our relabeling techniques, we conduct different methods for data transfer and implement one fixed Resnet to be the event zone estimation classifier. The final result is shown in Table II. We observe similar behaviors discussed in Section V-C. Specifically, we have (i) HJDA performs better than PCA, TCA, and CDLS. The averaged testing accuracy among all scenarios is 74.7%, 70.1%, 68.7%, and 60.6% for HJDA, CDLS, TCA, and PCA, respectively. (ii) Information migration from a small system to a large system still brings better results. This indicates that the over-compression phenomenon still exists, explained in Section V-C. Basically, for a larger system, the monitoring sub-region is wider with more variability for its inner events. For example, some line faults have a bigger impact on the systems while others don’t. Thus, compressing all of these event data to match the conditional distribution hurt the final classifier performance. We demand further theoretical frameworks to analyze and prevent the over-compression problem.

E. Sensitivity Analysis With Respect to Data Numbers: Increasing Data Points Boost the Performance of HJDA

In Section V-C, we utilize $N = 450$ to demonstrate the results and show large improvements of our HJDA over other methods. To further illustrate the improvements of all methods giving more data, we conduct a sensitivity analysis with respect to data numbers for event type differentiation. Specifically, we vary $N \in \{400, 450, 500, 550, 600, 650, 700\}$. Then, we report the mean of the testing accuracy, and the results are shown from Fig. 8(a) to Fig. 8(f). We find that if $N \geq 550$, our method can achieve the average accuracy larger than or equal to 90.5%. When $N = 700$, the average accuracy can be 95.7%. This shows that we can provide a reliable result for the utility partners. Further, we observe that for each scenario, our method enjoys the largest improvement when the number of data points increases. This illustrates the advantages of our method to efficiently absorb new information from new data samples. Such a significant result is due to our method’s complete consideration of conditional distribution and marginal distribution differences (compared to TCA + Resnet method) and flexible projection process (compared to CDLS + Resnet method), as illustrated in Section V-C.

F. Computational Time for Event Simulation and Model Training and Testing

The computational time is one of the most important evaluations for real-time applications. Thus, we provide the time for developing both the power network for the dynamic simulation and the deep residual network (Resnet) for classification. The former represents the simulation time in PSLF and the latter includes the training/testing time of the classification model.

For the PSLF simulation of the power networks, we observe that the simulation time depends on the system size. Specifically, for 14-bus, 200-bus, and 500-bus systems, the average simulation times are 2.5 s, 3.3 s and 4.2 s, respectively. This implies that PSLF can quickly generate high-quality event data for the research usage.

For Resnet, our experiments include different transfer learning methods, including HJDA + Resnet, CDLS + Resnet, TCA + Resnet, and PCA + Resnet. Specifically, HJDA, CDLS, TCA, and PCA all include the dimensionality reduction process and we restrict the projected data to have the same dimensionality and the Resnet to have the same architecture. Thus, we expect close training and testing time for HJDA + Resnet, CDLS + Resnet, TCA + Resnet, and PCA + Resnet. As for the benchmark PCA + Resnet without transfer learning, we only utilize the target data to train and test, leading to much smaller training and testing time.

Table III illustrates the result when we utilize the example of 200 → 500 with $N = 450$ for transfer learning-based methods.
Fig. 8. Sensitivity analysis for event type differentiation.

(a) Testing accuracy (%) for scenario 14 → 200.

(b) Testing accuracy (%) for scenario 200 → 14.

(c) Testing accuracy (%) for scenario 14 → 500.

(d) Testing accuracy (%) for scenario 500 → 14.

(e) Testing accuracy (%) for scenario 200 → 500.

(f) Testing accuracy (%) for scenario 500 → 200.
and \( N = n_T = 50 \) for the benchmark method. Finally, we consider 3-fold cross validation. Thus, for transfer learning-based methods, we obtain 300 data points for training and 150 points for testing. For the benchmark method, we obtain 33 data points for training and 17 data points for testing.

We find that (1) it takes around 13 s for training with 300 data points. This training time is affordable as the training process can be treated off-line to prepare the classifier, which doesn’t affect the real-time classification. (2) it takes around 0.23 s for testing with 150 data points, which indicates that we can achieve the real-time event identification.

VI. DISCUSSION

Our paper shows the great novelty of investigating power system knowledge transfer. To the best of the authors’ knowledge, limited work [13]–[15] has a concern on this topic, and their methods either can’t tackle power grid measurement heterogeneity or don’t provide solid mathematical foundations. Thus, our research contributions include both model design with solid theories and extensive numerical verification. In Section III, we show how the optimization model is developed based on Bayes’ theorem.

Then, in Section V, we conduct 6 transfer learning scenarios among 3 power systems with the changing PMU numbers, leading to 36 testing cases. Results in each testing case show that our proposed HJDA performs better than other methods. Averagely, our method has 5.0%, 8.1%, and 17.1% accuracy improvements over TCA, CDLS, and PCA methods. Secondly, we conduct sensitivity analysis with respect to the number of data. We find that when the number of data is equal to or larger than 550, the testing accuracy of our method can reach 90.5%. Further, when the number of data is larger than 700, the average accuracy can reach 95.7%, which shows the reliability for realistic applications.

Finally, we propose 3 challenges that need to be addressed in future papers. (i) optimal event zone categorization. In our paper, we develop (8) as a measure to assign each event to an event zone. However, we can’t guarantee that the process is optimal. In the future, we will develop an optimization model to find the optimal zones with strict mathematical guarantees. (ii) different number of optimal zones for two grids. The optimal zones can have different numbers for two systems, which indicates that the event-zone-index space is overlapped but not the same. Under this condition, we will propose a novel training method to efficiently tackle the label space difference and guarantee the ML model performance in the target system. (iii) fine-grained label division. As shown in the last Paragraph of Section V-C, we find that for two physical systems, the label within one event type can have different variability. For example, 500-bus system has more levels of generator output than that of the 14-bus system. Thus, treating both systems’ generator trips as one integer causes errors for transfer learning. This implies a more fine-grained label division to guarantee the improved performance of transfer learning, which we will investigate in future work.

VII. CONCLUSION

The evolution of modern power systems encourages a Machine Learning (ML) solution for fast and accurate event identification. However, ML-based methods require enough data for training, which is often non-applicable for new systems. Thus, in this paper, we aim at transferring knowledge from a data-rich source grid to a data-limited target grid to boost the ML performance in the target grid.

To achieve this goal, we systematically design a PMU-based Transfer Learning (TL) framework with strict mathematical guarantees and experimental verification. Specifically, we first verify the existence of the shared knowledge, i.e., the common behaviors to an event across systems, thus proving the transferability of our problem. Secondly, to extract and transfer common knowledge from data, we propose a Heterogeneous Joint Domain Adaptation (HJDA) with strong support from the probability theory. HJDA projects source/target data to a low-dimensional subspace with close feature distributions to represent the shared knowledge. Using the common features, any ML model can be trained to differentiate event types.

For event zone estimation, however, the event-location label spaces are disjoint for two grids, which prevents the optimization of HJDA and the following ML training. To handle this issue, we propose an event-label transfer method to align and relabel events by the geographical information of the event’s most-affected PMU. Thus, the two disjoint label spaces are mapped to a common responsible-PMU-location space. With the label transfer technique, HJDA and the ML models can be used for event zone estimation. Finally, comprehensive experiments demonstrate the advantages of our model over some state-of-the-art transfer learning models.

REFERENCES

[1] Y. Weng, R. Negi, and M. D. Ilić, “Probabilistic joint state estimation for operational planning,” IEEE Trans. Smart Grid, vol. 10, no. 1, pp. 601–612, Jan. 2019.
[2] H. Li, Y. Weng, E. Farantatos, and M. Patel, “A hybrid machine learning framework for enhancing PMU-based event identification with limited labels,” in Proc. IEEE Int. Conf. Smart Grid Synchronized Meas. Analytics, 2019, pp. 1–8.
[3] H. Li, Y. Weng, E. Farantatos, and M. Patel, “An unsupervised learning framework for event detection, type identification and localization using PMUs without any historical labels,” in Proc. IEEE Power Energy Soc. Gen. Meeting, 2019, pp. 1–5.
[4] Y. Yuan, Y. Guo, K. Dehghanpour, Z. Wang, and Y. Wang, “Learning-based real-time event identification using rich real PMU data,” IEEE Trans. Power Syst., vol. 36, no. 6, pp. 5044–5055, Nov. 2021.
[5] O. A. Alini, K. Onahada, and A. M. Abu-Mahfouz, “A review of machine learning approaches to power system security and stability,” IEEE Access, vol. 8, pp. 113 512–113 531, 2020.
[6] M. Mohammadpourfarid, Y. Weng, M. Pechenizkiy, M. Tadjianian, and B. Mohammadi-Ivatloo, “Ensuring cybersecurity of smart grid against data integrity attacks under concept drift,” Int. J. Elect. Power Energy Syst., vol. 119, 2020, Art. no. 105947.

TABLE III

| TABLE III | THE TRAINING AND TESTING TIME (S) FOR RESNET WITH DIFFERENT TRANSFER LEARNING METHODS FOR THE SCENARIO 200 \( \rightarrow \) 500 |
|---|---|---|---|---|
|   | HJDA | CDLS | TCA | PCA |
| TRAINING TIME | 12.51 | 13.65 | 13.01 | 1.64 |
| TESTING TIME  | 0.14 | 0.15 | 0.15 | 0.03 |
[7] S. J. Pan and Q. Yang, “A survey on transfer learning,” *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[8] M. Guillaumin and V. Ferrari, “Large-scale knowledge transfer for object localization in imagenet,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 3202–3209.

[9] Y. Zha et al., “Heterogeneous transfer learning for image classification,” in *Proc. AAAI Conf. Artif. Intell.*, 2011, pp. 1304–1309.

[10] L. Duan, I. W. Tsang, D. Xu, and T.-S. Chua, “Domain adaptation from multiple sources via auxiliary classifiers,” in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 289–296.

[11] J. Blitzer, R. McDonald, and F. Pereira, “Domain adaptation with structural correspondence learning,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2006, pp. 120–128.

[12] L. Feng, Y.-S. Ong, I.W.-H. Tsang, and A.-H. Tan, “An evolutionary search paradigm that learns with past experiences,” in *Proc. IEEE Congr. Evol. Computation*, 2012, pp. 1–8.

[13] C. Ren and Y. Xu, “Transfer learning-based power system online dynamic security assessment: Using one model to assess many unlearned faults,” *IEEE Trans. Power Syst.*, vol. 35, no. 1, pp. 821–824, Jan. 2020.

[14] H. Li, Y. Weng, and H. Tong, “Heterogeneous transfer learning on power systems: A merged multi-modal Gaussian graphical model,” in *Proc. IEEE Int. Conf. Data Mining*, 2020, pp. 1088–1093.

[15] X. He, Q. Liu, J. Tang, Q. Ma, G. Wang, and M. Liu, “Power system frequency situation prediction method based on transfer learning,” in *Proc. IEEE Asia-Pacific Power Energy Eng. Conf.*, 2020, pp. 1–5.

[16] O. Day and T. M. Khoshgoftaar, “A survey on heterogeneous transfer learning,” *J. Big Data*, vol. 4, no. 1, pp. 1–42, 2017.

[17] J. T. Zhou, I. W. Tsang, S. J. Pan, and M. Tan, “Heterogeneous domain adaptation for multiple classes,” in *Proc. Artif. Intell. Statist.*, 2014, pp. 1095–1103.

[18] Y. Yan, W. Li, H. Wu, H. Min, M. Tan, and Q. Wu, “Semi-supervised optimal transport for heterogeneous domain adaptation,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, vol. 7, pp. 2969–2975.

[19] Y.-H. H. Tsai, Y.-R. Yeh, and Y.-C. F. Wang, “Learning cross-domain landmarks for heterogeneous domain adaptation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 5081–5090.

[20] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, “Transfer feature learning with joint distribution adaptation,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2200–2207.

[21] C. Stegner, J. Bogenrieder, P. Luchscheider, and C. J. Brabec, “First year of smart metering with a high time resolution–realistic self-sufficiency rates for households with solar batteries,” *Energy Procedia*, vol. 99, pp. 360–369, 2016.

[22] G. Liu, J. Gu, J. Zhao, F. Wen, and G. Liang, “Super resolution perception for smart meter data,” *Inf. Sci.*, vol. 526, pp. 263–273, 2020.

[23] C. Zha and A. Reinhardt, “Reliable streaming and synchronization of smart meter data over intermittent data connections,” in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids*, 2019, pp. 1–6.

[24] M. T. Rosenstein, Z. Marx, L. P. Kaelbling, and T. G. Dietterich, “To transfer or not to transfer,” in *Proc. Neural Inf. Process. Syst. Workshop Transfer Learn.*, vol. 898, 2005, pp. 1–4.

[25] Engineering Texas A&M University, “Illinois 200-Bus system: ACTIVSg200,” 2016. [Online]. Available: https://electricgrids.engr.tamu.edu/electric-grid-test-cases/activsg200/

[26] Engineering Texas A&M University, “SouthCarolina 500-Bus system: ACTIVSg500,” 2016. [Online]. Available: https://electricgrids.engr.tamu.edu/electric-grid-test-cases/activsg500/

[27] M. U. Usman and M. O. Faruque, “Applications of synchrophasor technologies in power systems,” *J. Modern Power Syst. Clean Energy*, vol. 7, no. 2, pp. 211–226, 2019.

[28] Illinois Center for a Smarter Electric Grid, “IEEE 14-Bus system,” 2013. [Online]. Available: https://icseg.iti.illinois.edu/ieee-14-bus-system/

[29] North American Electric Reliability Corporation, “Real-time application of synchrophasors for improving reliability,” 2017. [Online]. Available: https://www.naspi.org/sites/default/files/reference_documents/capr_final_20101017.pdf?fileID=519

[30] J. Qin, Z. Li, and Y. Zha, “Protection mechanism in reliability evaluation approach to multistate system with common cause failure,” *Complexity*, vol. 2021, pp. 1–12, 2021, doi: 10.1155/2021/6691362.

[31] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, “A kernel two-sample test,” *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 723–773, 2012.

[32] General Electric Energy Consulting, “General electric concorda PSLF,” 2018. [Online]. Available: https://www.geenergyconsulting.com/practice-area/software-products/pslf

[33] R. Ma, S. Basumallick, and S. Eftekharnejad, “A PMU-based data-driven approach for classifying power system events considering cyberattacks,” *IEEE Syst. J.*, vol. 14, no. 3, pp. 3558–3569, Sep. 2020.

[34] Y. Yuan, S. Low, O. Ardakanian, and C. Tomlin, “On the inverse power flow problem,” 2020, arXiv:1610.06631.

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

[36] M. Guo and Y. Du, “Classification of thyroid ultrasound standard plane images using ResNet-18 networks,” in *Proc. IEEE Int. Conf. Anti-Counterfeiting, Secur., Identification*, 2019, pp. 324–328.

[37] L. Ma, R. Shuai, X. Ran, W. Liu, and C. Ye, “Combining DC-GAN with ResNet for blood cell image classification,” *Med. Biol. Eng. Comput.*, vol. 58, no. 6, pp. 1251–1264, 2020.

[38] H. Choi, S. Ryu, and H. Kim, “Short-term load forecasting based on ResNet and LSTM,” in *Proc. IEEE Int. Conf. Commun., Control, Comput. Technol. Smart Grids*, 2018, pp. 1–6.

[39] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, “Domain adaptation via transfer component analysis,” *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.