processes

MDPI

Article

Robust Distribution System State Estimation with Physics-Constrained Heterogeneous Graph Embedding and Cross-Modal Attention

Siyan Liu [1,*], Zhuang Tang [1], Bo Chai [1] and Ziyu Zeng [2]

[1] China Electric Power Research Institute Co., Ltd., Beijing 102209, China; tz_work@126.com (Z.T.); chaibo@epri.sgcc.com.cn (B.C.)
[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China; zengziyu23@mails.tsinghua.edu.cn
* Correspondence: liusiyan1992@126.com

Abstract

Real-time distribution system state estimation is hampered by limited observability, frequent topology changes, and measurement errors. Neural networks can capture the nonlinear characteristics of power-grid operation through a data-driven approach that possesses important theoretical value and is promising for engineering applications. In that context, we develop a deep learning framework that leverages General Attributed Multiplex Heterogeneous Network Embedding to explicitly encode the multiplex, heterogeneous structure of distribution networks and to support inductive learning that adapts to dynamic topology. A cross-modal attention mechanism further models fine-grained interactions between input measurements and node/edge attributes, enabling the capture of nonlinear correlations essential for accurate state estimation. To ensure physical feasibility, soft power-flow residuals are incorporated into training as a physics-constrained regularization, guiding predictions toward consistency with grid operation. Extensive studies on IEEE/CIGRE 14-, 70-, and 179-bus systems show that the proposed method surpasses conventional weighted least squares and representative neural baselines in accuracy, convergence speed, and computational efficiency while exhibiting strong robustness to measurement noise and topological uncertainty.

Keywords: state estimation; graph neural network; deep learning; power flow

check for updates

Academic Editors: Yue Qiu, Suyang Zhou, Shichang Cui and Qirun Sun

Received: 2 September 2025
Revised: 21 September 2025
Accepted: 22 September 2025
Published: 25 September 2025

Citation: Liu, S.; Tang, Z.; Chai, B.; Zeng, Z. Robust Distribution System State Estimation with Physics-Constrained Heterogeneous Graph Embedding and Cross-Modal Attention. *Processes* **2025**, *13*, 3073. https://doi.org/10.3390/pr13103073

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Distribution system state estimation (DSSE) is a key component of modern power-system operation. It is critical to ensuring stable and efficient grid management [1]. The existing state estimation (SE) techniques are widely adopted in transmission systems, but they face notable challenges when they are applied to distribution networks [2]. First, the core challenge of the DSSE task lies in the dynamic changes in the network topology. In practical scenarios, the distribution network system will undergo dynamic changes in its topological structure due to load switching, line maintenance, and other events. Traditional methods were originally designed for simple transmission networks, which will render the predicted grid states invalid when the network structure changes [1]. Second, unlike transmission systems, which rely on a large number of redundant measurements, distribution systems often require higher accuracy and real-time performance. In distribution systems, the strong nonlinear relationships between voltage drop and power loss make it

difficult use the direct application of traditional algorithms for measurement. Moreover, distribution systems have insufficient data, so traditional methods will produce a large number of estimation errors under complex operating conditions [3]. Furthermore, the problem of persistent observability in distribution systems remains unsolvable. Due to cost constraints, distribution networks often deploy measurement equipment only at key nodes. Therefore, a large number of nodes need to obtain pseudomeasurements from historical data or predictive models, which leads to the occurrence of further uncertainties [4].

Consider the weighted least squares (WLS) method, the prevailing SE technique in transmission systems. Under ideal conditions—where measurement noise is purely white and Gaussian—WLS acts as an unbiased minimum-variance estimator, delivering optimal performance [5]. However, in distribution systems, measurement noise amplifies with network uncertainty, directly compromising the observability of active distribution networks [6,7]. Moreover, WLS suffers from two critical drawbacks for DSSE: its iterative computation is computationally intensive, and its performance degrades drastically under noisy measurements [8]. To mitigate these issues, researchers have proposed robust variants such as the maximum normalized residual test [9], weighted least absolute values [10], least median of squares [11], and least trimmed squares [12]. Despite these advancements, these models remain constrained by parametric dependencies and unresolved convergence/sensitivity tradeoffs, limiting their practical applicability to distribution systems with sparse measurements.

The Kalman filter (KF), leveraging temporal correlation through state-space modeling, offers an alternative by incorporating prior state estimates as auxiliary information to improve accuracy and convergence speed under low-observability scenarios [13,14]. However, its reliance on linear system assumptions clashes with the highly non-linear nature of real-world power systems, where power-flow equations exhibit strong non-linearities that degrade robustness and estimation accuracy in practical DSSE applications [15].

To address the problem of system observability, transmission grids rely on dense sensor deployments for real-time state measurements, a strategy that directly fulfills observability requirements [16]. However, this approach is infeasible for DSSE, where the limited number of installed measuring instruments leads to insufficient data redundancy. Specifically, the scarcity of measurements creates an underdetermined problem, making it challenging to iteratively derive accurate state estimates from sparse data—a fundamental observability limitation in DSSE [17].

To mitigate this issue, pseudo-measurements are employed. They are generated through observability-driven meter placement techniques: first, the unobservability index (UI) is calculated using information entropy to identify critical measurement locations; then, instruments are installed based on UI rankings. Historical data are then leveraged to derive predictive values for unmeasured states, a strategy that significantly reduces hardware costs compared to full sensor retrofitting. Despite these advantages, the generalizability of this method remains unvalidated for two critical scenarios: unbalanced multi-phase distribution systems and multi-time instance state estimation, where temporal dynamics and phase asymmetry introduce additional complexities.

Owing to the limitations of traditional methods, with the continuous advancement of artificial intelligence, a growing body of research has focused on applying deep learning to DSSE tasks. The limitations of traditional DSSE methods in terms of measurement sparsity, topological dynamism, nonlinear modeling, and multi-source data fusion stem essentially from their core model-driven logic, which relies on manual simplification and assumptions of the physical laws governing power grids. However, the complexity of modern distribution networks—characterized by high-penetration distributed generation, flexible loads, and multi-source data—has exceeded the limits of manual modeling capabilities.

Centered on a data-driven paradigm, deep learning can specifically address the intractable challenges faced by traditional methods through its capabilities such as automatic feature extraction, topological adaptability, robust nonlinear fitting, and multi-source data fusion. As such, it has become an indispensable technical approach used to enhance the accuracy, efficiency, and adaptability of DSSE. Its indispensability is evident not only in its use to solve existing problems but also in that it provides an extensible technical framework for the refined state estimation of future distribution networks.

Deep learning frameworks, which are capable of adaptively learning task-specific patterns from DSSE data, have demonstrated promising performance in recent studies. Current approaches include leveraging artificial neural networks (ANNs) to model specific power-system components alongside convolutional neural networks (CNNs) [18], multi-layer perceptrons (MLPs) [19], generative adversarial networks (GANs) [20], and Bayesian networks [15,21]. However, a critical bottleneck persists: the effectiveness of these neural networks hinges on the availability of large-scale, high-quality training datasets. For DSSE—where real-world measurements are often sparse, noisy, or incomplete—the challenge of ensuring data validity (e.g., addressing missing values, bad data, and non-stationary distributions) remains unresolved. While recent works attempt to mitigate this by incorporating inductive biases (e.g., topological priors) or physics-informed constraints [22,23], these hybrid methods still require substantial amounts of labeled data to achieve reliable generalization, limiting their applicability in scenarios with limited measurement resources.

Given the limitations of purely model-based or data-driven approaches, hybrid frameworks integrating physical insights with data-driven learning have emerged as a promising direction for DSSE [13]. Zhang et al. [24] proposed embedding physical regularization terms into deep neural networks to enforce power-system constraints, while Kumar et al. [25] developed an artificial neural network tailored for non-Gaussian noise corrupted by bad data—though their method requires prior knowledge of precise equipment states and measurement baselines. Rui et al. [26] introduced a Tapered DNN architecture that incorporates maximum entropy principles into DSSE, achieving accurate estimation via layer-wise unsupervised feature learning. Duan et al. [27] adopted a hybrid DL–ML strategy, fusing CNN and random forest models to extract dynamic features from time-series measurements. Gotti et al. [28] proposed a PCA–DBN framework: principal component analysis isolates noise-robust features, which are then fed into a deep belief network for topological structure identification; this approach demonstrates strong resilience to data loss and measurement noise. Ostrometzky et al. [29] developed a physics-informed dynamic DSSE framework that uses power-flow equations as regularization constraints, while Wang et al. [30] replaced the decoder of an autoencoder with a physical model to enable hybrid state estimation.

However, a common limitation of these methods is their neglect of explicit network-topology modeling, a critical shortcoming given the structural complexity of distribution grids. Graph Neural Networks (GNNs), by contrast, leverage network topology as inductive bias, inherently addressing the curse of dimensionality and demonstrating robust performance under topological perturbations [31,32]. Recent advances include EleGNN, as presented by Liu et al. [33], which improves traditional GNNs by incorporating physical connectivity and using node-edge feature propagation to model complex grid interactions. Madbhavi et al. [34] designed a GNN-based estimator that takes measurement matrices/tensors as input, introducing feature scaling and a pseudo-measurement generation module to improve generalization. Ngo et al. [35] further integrated knowledge of the physical field with the GNN architectures, allowing more effective processing of structural data to capture latent dependencies of the topological state.

In distribution-system operations, frequent topological changes [36] pose a critical challenge: retraining models to adapt to new configurations incurs substantial time and computational costs. Moreover, relying exclusively on graph structures for modeling often results in loss of critical node-specific attributes, as distribution systems are inherently multi-source and heterogeneous—each component carries rich attribute information beyond purely structural connections [1]. To address the limitations of traditional graph neural networks, which struggle to balance topological extraction and multi-source feature modeling in heterogeneous environments, this paper employs General Attributed Multiplex Heterogeneous Network Embedding (GATNE) [37]. By integrating diverse node attributes and structural multiplexity, GATNE effectively captures the nonlinear dependencies within distribution systems, overcoming the information loss that occurs in purely topological modeling. This approach enables the model to dynamically adapt to topological variations without full retraining while leveraging heterogeneous attributes to enhance the accuracy and robustness of DSSE. These features are critical to handling complex, real-world scenarios in distribution grids.

To tackle the critical dependency of DSSE on data quality, this paper integrates soft power-flow equations into the loss function, enforcing physical consistency by penalizing predictions that deviate from power-flow constraints. Unlike hard-constraint methods, this approach softly regularizes outputs to lie within the feasible operating region defined by power-flow dynamics, discarding implausible solutions that violate fundamental electrical laws. This ensures not only that the model's estimates are mathematically consistent but also that they maintain engineering viability, significantly enhancing robustness in handling noisy, incomplete, or uncertain real-world data.

Additionally, a cross-modal attention module is proposed to model the intricate interactions between input measurements and edge features, explicitly capturing the latent relationships between observed data and topological connectivity. The model adaptively weights the informative characteristics in heterogeneous modalities during the state estimation by merging the measurement inputs and the topological graph G within the GATNE framework. The output embeddings are fed into a power-flow constraint layer, which acts as a physics-informed filter to refine predictions against actual grid dynamics. Key contributions of this work include the following:

(1) A GATNE-based DSSE architecture is proposed. By modeling the multi-source heterogeneous structure of distribution systems and enabling inductive learning using measurement data, it can effectively capture the nonlinear relationships between nodes, thereby improving the model's accuracy and robustness.

(2) A cross-modal attention module is proposed to learn the correlations between model inputs and topological structure attributes, and it uses this correlation to better enable the model to mine hidden features, thereby improving the accuracy of the model in DSSE.

(3) The power-flow equations are introduced into the neural network architecture, combining the characteristics of data-driven models and restricting admissible solutions within a certain range to ensure that the model's output aligns with the objective laws of real-world physical scenarios, thereby enhancing the robustness and generalization capability of model predictions.

This paper introduces the proposed method in Sections 2 and 3, which respectively describe the application of the GATNE model to this task. Section 4 presents a case study that compares the performance of baseline algorithms and other data-driven models. Finally, Section 5 concludes the work.

2. Proposed Methods

2.1. Conventional Problem Formulation

The DSSE problem can be framed as the task of determining the state vector x from the measurement variable z. This process is fundamentally rooted in the interplay between network topology, bus parameters, and real-time measurements. Mathematically, this relationship can be expressed as follows:

$$z = h(x) + \epsilon \tag{1}$$

where z typically encompasses various measurements. The state variable x generally includes critical values. The term ϵ represents the noise vector. The function $h(\cdot)$ denotes the power-flow equations, which encapsulate the physical model parameters of the system. The specific formulations of $h(\cdot)$ are detailed in [38].

$$h(x) = \begin{cases} V_j = V_j \\ \phi_j = \phi_j \\ P_{j\to k} = V_j V_k Y_{jk} e^{i(\pi+\Delta\phi_{jk})} + V_j^2[\mathbb{R}(Y_{jk}) + \frac{\mathbb{R}(Y_{s_{jk}})}{2}] \\ P_{j\leftarrow k} = V_j V_k Y_{jk} e^{i(\pi-\Delta\phi_{jk})} + V_j^2[\mathbb{R}(Y_{jk}) + \frac{\mathbb{R}(Y_{s_{jk}})}{2}] \\ Q_{j\to k} = V_j V_k Y_{jk} e^{i(\Delta\phi_{jk}-\frac{\pi}{2})} - V_j^2[\mathbb{I}(Y_{jk}) + \frac{\mathbb{I}(Y_{s_{jk}})}{2}] \\ Q_{j\leftarrow k} = V_j V_k Y_{jk} e^{i(\frac{\pi}{2}-\Delta\phi_{jk})} - V_j^2[\mathbb{I}(Y_{jk}) + \frac{\mathbb{I}(Y_{s_{jk}})}{2}] \\ I_{j\to k} = \left| \frac{P_{j\to k} - jQ_{j\to k}}{\sqrt{3}V_i e^{-i\phi_j}} \right| = \frac{|P_{j\to k} - iQ_{j\to k}|}{\sqrt{3}V_j} \\ I_{j\leftarrow k} = \left| \frac{P_{j\leftarrow k} - iQ_{j\leftarrow k}}{\sqrt{3}V_k e^{-i\phi_j}} \right| = \frac{|P_{j\leftarrow k} - iQ_{j\leftarrow k}|}{\sqrt{3}V_k} \\ P_j = -\sum_{k\in N(j)} P_{j\leftarrow k} + P_{j\to k} \\ Q_j = -\sum_{k\in N(j)} Q_{j\leftarrow k} + Q_{j\to k} \end{cases} \tag{2}$$

where i represents the imaginary unit. The voltage difference between bus j and k is defined as $\Delta\phi_{jk} = \phi_j - \phi_k + \varphi_{jk}$ and denotes the shift angle of the transformer. φ_{jk} is the shift angle of the transformer. V_j and V_k are the voltage magnitudes at the starting and ending nodes of bus j to k. Y_{jk} and $Y_{s_{jk}}$ are the line admittance and shunt admittance of the line between bus j and k, respectively. In addition, the $j \to k$ is defined as the direction of power flow from bus j to k, and $j \leftarrow k$ is the opposite. P and Q represent the active and reactive powers of the bus, respectively. I represents the current.

The power-flow equations serve as a critical link between the state variables and various measurement equations within the network. Consequently, the state vector x can be estimated by determining the inverse relationship $h^{-1}(z)$ and correcting for the measurement errors ϵ. Traditional SE algorithms typically employ the iterative Newton–Raphson method to minimize the WLS objective function [38], as illustrated in Figure 1a. The formulation of the WLS is given as follows:

$$\hat{x} = \arg\min_x \left\{ \sum_{i=1}^{M} w_i (z_i - h_i(x))^2 \right\} = \arg\min_x \{[z - h(x)^T] W [z - h(x)]\} \tag{3}$$

In this equation, w_i represents the weight assigned to the i-th measurement, while W denotes the error covariance matrix. The purpose of the error covariance matrix is to ensure that any two measurement errors are statistically independent of one another. However, directly solving the state estimation problem can be quite challenging. Therefore, the Newton–Raphson iterative method is employed, with convergence achieved by setting the gradient of the objective function, denoted as J in Equation (3), to zero.

At the k-th time step, the Jacobian matrix of the state estimation value $x(k)$ is represented as $H(x(k))$. This matrix contains the partial derivatives of the measurement equations with respect to the state variables, providing essential information about how changes in the state variables affect the measurements. The gain matrix, denoted as G, is defined as follows:

$$G(x(k)) = H(x(k))^T W H(x(k)) \tag{4}$$

Therefore, the increment of the state estimation at the k-th time step can be calculated as follows:

$$\Delta x(k) = G(k)^{-1} H x(k)^T W(z - h(x(k))) \tag{5}$$

The state estimation value at the $k+1$-th time step can be obtained based on $x(k+1) = x(k) + \Delta x(k)$. Finally, through this iterative approach, convergence is achieved and the desired $\hat{x}$ can be obtained.

However, the Newton–Raphson iterative method is sensitive to the choice of initial conditions. In certain scenarios, the gain matrix derived from this method may produce multiple solutions, a phenomenon referred to as ill-conditioning. Furthermore, the method necessitates a substantial number of state measurements to ensure reliability in the results. In cases where redundancy is insufficient, the estimates may become inaccurate.

Additionally, for the Newton–Raphson method to be effective, the system must be observable. This requirement poses significant challenges in the context of Dynamic State Estimation (DSSE), where achieving full observability can be impractical. If the method is applied in situations of poor observability, the iterative outputs may diverge rather than converge, leading to unreliable state estimates [8].

These limitations highlight the need for alternative approaches or enhancements to the Newton–Raphson method, particularly in scenarios where measurement redundancy is limited or system observability is compromised. Addressing these challenges is crucial for improving the robustness and accuracy of state estimation in dynamic systems.

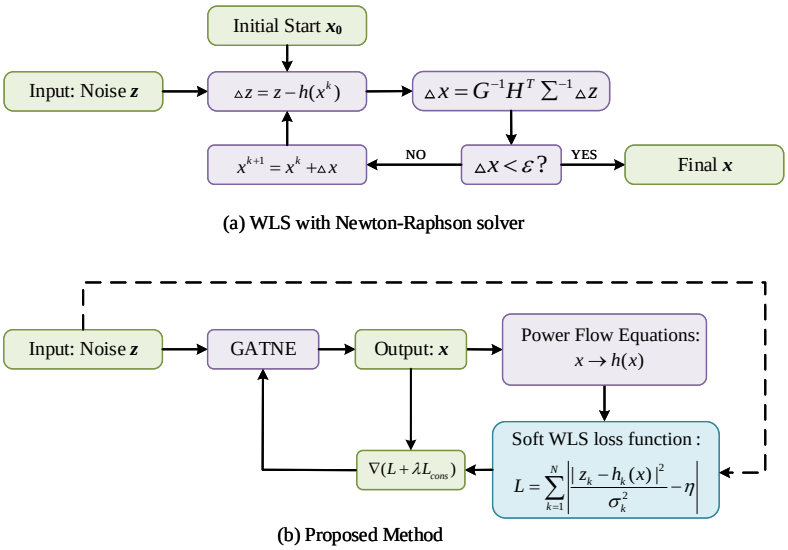


Figure 1. (**a**) SE with WLS with a Newton–Raphson solver. Starting from the initial state x_0, it updates x^k iteratively until the value of the residual Δx is less than a given threshold ϵ. (**b**) Combining deep learning with physical regularization terms, the task is formulated as a weakly supervised task, such that the output x of the GATNE model fits the output z. The physical constraint term L_{cons} is defined in Equation (7).

Although neural network methods are capable of effectively capturing complex mapping relationships, they often exhibit a strong dependence on high-quality data. To address

this limitation, this paper proposes a deep learning framework that integrates physical regularization terms. These physical regularization terms serve as a form of weak supervision, helping to mitigate the reliance on high-quality data.

By incorporating physical principles into the learning process, the proposed framework enhances the model's robustness and generalizability, allowing it to perform well even in scenarios where data quality may be suboptimal. This approach not only improves the accuracy of the neural network's predictions but also ensures that the learned representations remain consistent with the underlying physical laws governing the system. Thus, the integration of physical regularization terms represents a significant advancement in the development of data-driven models for dynamic state estimation, facilitating the production of more reliable and interpretable outcomes.

2.2. Physical Regularization Term

To embed physical knowledge within the DSSE framework, this paper integrated the power-flow equation (Equation (2)) directly into the training loop. Specifically, as illustrated in Figure 1b, the GATNE model ingests the feature vector z and produces a candidate state estimate x. This estimate is then propagated through the power-flow equation, and the resulting residuals are penalized via a weighted least-squares (WLS) loss. To avoid overconstraining the network by strict enforcement of the power-flow laws, this paper further introduced soft-bound constraints, ensuring that any solution lying within predefined intervals is accepted:

$$L(z, x) = \sum_{k \in M} \left| \frac{|z_k - h_k(x)|^2}{\sigma_k^2} - \mu \right| \tag{6}$$

where σ_k denotes the standard deviation, M the set of all measurements, and μ a prescribed accuracy threshold. The cardinality of the measurement set is given by $|M| = m$, and the soft regularization term μ governs the allowable deviation in Equation (6).

During training, this paper minimizes the discrepancy between the input feature vector z and the network's reconstructed output $h(x)$, treating each measurement's uncertainty as a weight to prevent the model from being dominated by any single datum. When the power-flow relationship of Equation (2) is embedded into the loss—and complemented with these soft bounds—the framework remains data-driven while still producing solutions that respect underlying physical laws, even when labels are incomplete or corrupted.

Furthermore, because the state variables in the power-flow equations are fully differentiable, their gradients can be obtained directly via the measurement Jacobian matrix $H(x) = \nabla h(x)$. Leveraging this structure enhances numerical stability, accelerates convergence, and ensures that the estimator meets stringent robustness and observability criteria.

2.3. The Constraints of a Physical Regularization Term

Using only WLS for learning may lead to the problem of falling into local minima. Inspired by [1], this paper adds penalty terms to the loss function to guide the model to learn physical rules better. Specifically, the constraint terms are voltage stability, phase-angle stability, and line-loading stability.

For voltage stability, to ensure that the voltage level of each node remains between 95% and 105%, this paper adds bilateral constraints $\max(0, V - 1.05) + \max(0, 0.95 - V)$ to the loss function to enforce this criterion. For phase-angle stability, the phase-angle difference in a stable system should be less than 0.25 rads. Therefore, another bilateral constraint $\max(0, \Delta\phi - 0.25) + \max(0, -0.25 - \Delta\phi)$ is added to limit the phase-angle difference. Finally, for line-loading stability, as the loading cannot exceed 100%, the third constraint is $\max(0, l - 1)$, where l represents the line loading.

Finally, the loss function employed in this paper is a combination of the above terms, as follows:

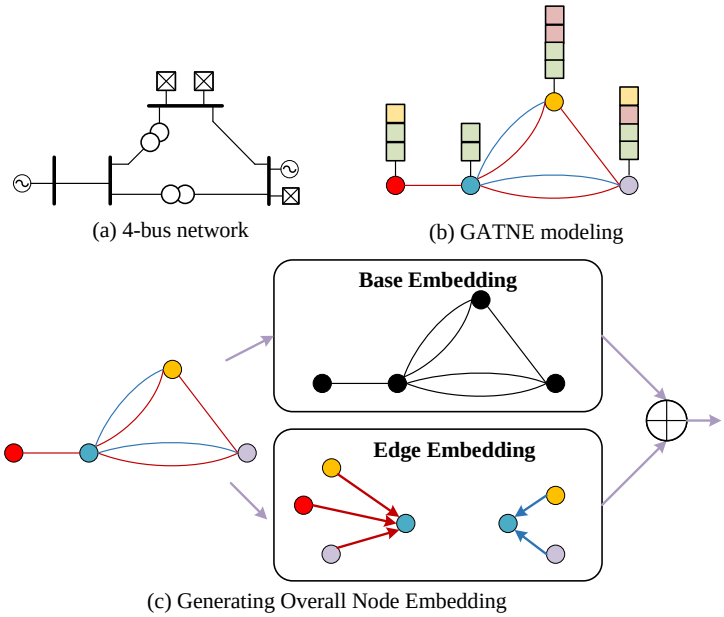
$$L(z,x) = \sum_{k\in M}\left|\frac{|z_k - h_k(x)|^2}{\sigma_k^2} - \mu\right| + \lambda_0[\lambda_1 \max(0, V - 1.05) + \lambda_2 \max(0, 0.95 - V) + \lambda_3 \max(0, \Delta\phi - 0.25) + \lambda_4 \max(0, -0.25 - \Delta\phi) + \lambda_5 \max(0, l - 1)] \quad (7)$$

Among these terms,$\lambda_i, i \in Z[0,5]$ are hyperparameters set to balance the effects of each mathematical term during the training process. These terms guide the model's output towards physically reasonable boundaries and prevent it from diverging towards local minima that far exceed the system's physical boundaries.

3. The GATNE Neural Networks for DSSE

3.1. General Attributed Multiplex Heterogeneous Network Embedding

Conventional graph neural networks typically model node relationships as homogeneous and untyped, treating each edge as a simple, one-way connection. In contrast, real-world power systems are naturally heterogeneous graphs in which multiple distinct edge types link different component classes (e.g., generators, buses, transformers, loads), as illustrated in Figure 2a. Naïvely applying a standard GNN to such a structure risks conflating these diverse relationships and discarding important semantic information. To address this risk, we adopt the GATNE framework (Figure 2b), which explicitly represents and attends over multiple edge types. By decomposing each node's embedding into a shared base vector plus edge-type–specific adjustments, GATNE enables both efficient reuse of node attributes and faithful modeling of the rich, multi-relational topology inherent in power-grid networks. Additionally, line parameters are incorporated into the node–edge adjacency matrix of the GATNE model, which helps the model identify the differences in electrical characteristics between different lines. Since distribution networks are inherently heterogeneous graphs, line parameters serve as key identifiers for the edge features in the graph. They enable the model to distinguish the electrical properties of various line segments and thus accurately model the topological connections and functional differences of the distribution network.



(a) 4-bus network (b) GATNE modeling

(c) Generating Overall Node Embedding

Figure 2. Modeling the network (**a**) with two generators, three loads, two lines, and two transformers as (**b**) a GATNE model and (**c**) a GATNE process.

GATNE models a power grid by jointly learning representations for both nodes and edge types within a unified framework. Each physical component—whether a bus, transformer or load—is cast as a node carrying attributes such as voltage magnitude, power injection, impedance, and rated capacity. Edges are partitioned into two semantically distinct classes: direct electrical-connection links (e.g., bus–line and line–device associations) and device-correlation links that encode parameter couplings introduced by transformers. During message passing, these edge types guide an adaptive attention mechanism that separately aggregates neighborhood information for each relationship, thereby preserving the rich, multi-relational structure of the grid and avoiding the information loss typical of homogeneous GNNs. Concretely, for each target node, the algorithm first gathers messages from its various edge-type-specific neighbors, then transforms each collection into a corresponding embedding vector, and finally fuses these vectors into a comprehensive node representation, as described in Algorithm 1.

Algorithm 1: The GATNE process.

Input: The edge embedding $e_{i,r}^{(k)}$, the edge type r and the base embedding b_i.
Output: The node vector $v_{i,r}$

Initialize: Randomly initialize edge embedding $e_{i,r}^{(0)}$
/* **Stage 1: Calculate the aggregation.**
repeat
  for $i = 1; i < m; i++$ **do**
    for $k = 1; k < K; r++$ **do**
      $e_{i,r}^{(k)} = aggregator(\{e_{i,r}^{(k-1)}, \forall v_i \in N_{i,r}\}) \leftarrow$ Equation (9)
until *reaching max iteration*;
// Combining the edge embedding
$E_i \leftarrow$ Equation (10)
/* **Stage 2: Calculate the node vector** $v_{i,r}$**.**
// Calculating attention score
$attn_{i,r} \leftarrow$ Equation (11)
// Calculating node vector
repeat
  for $i = 1; i < m; i++$ **do**
    $v_{i,r} = \alpha_r M_r^T E_i attn_{i,r} + b_i \leftarrow$ Equation (9)
until *reaching max iteration*;

First, node v_i aggregates the neighbors of the k order of the edge type r to obtain the embedding of the edge $e_{i,r}^{(k)}$.

$$e_{i,r}^{(k)} = aggregator(\{e_{j,r}^{(k-1)}, \forall v_j \in N_{i,r}\}) \tag{8}$$

where $N_{i,r}$ is the neighbor of node v_i whose edge type is r. Edges of type r connected to node v_i are randomly initialized as $e_{i,r}^{(0)}$. This paper adopts the mean aggregation method to aggregate nodes, as follows:

$$e_{i,r}^{(k)} = \sigma(\hat{W}^{(k)} mean(\{e_{j,r}^{(k-1)}, \forall v_j \in N_{i,r}\})) \tag{9}$$

Then, the edge embeddings corresponding to k-th order neighbors of different edge types are aggregated, resulting in the following:

$$E_i = concat([e_{i,1}, e_{i,2}, \ldots, e_{i,m}]) \tag{10}$$

Next, the self-attention is used to compute the coefficients $attn_{i,r} \in \mathbb{R}^m$ of linear combination of vector E_i, as follows:

$$attn_{i,r} = softmax(w_r^T \tanh(W_r E_i))^T \tag{11}$$

where w_r and W_r are learnable matrices for edge type r. Thus the overall embedding of node v_i for edge type r is as follows:

$$v_{i,r} = \alpha_r M_r^T E_i attn_{i,r} + b_i \tag{12}$$

where b_i is the base embedding for node v_i and α_r is a hyperparameter. The importance of edge embeddings is emphasized, and the value can be adjusted to control the training process. $M_r \in \mathbb{R}$ is defined as a learnable transformation matrix.

3.2. The Cross-Modal Attention Mechanism

The cross-modal attention mechanism learns to align and fuse two complementary streams of information: the noise vector z, which captures measurement perturbations and operational uncertainties, and the system attributes Att, such as nodal voltages and line power flows, which encode the network physical state. Rather than handling these modalities separately or combining them with fixed weights, the model projects both into a shared representation space and computes a compatibility score between each noise component and each physical attribute. High-scoring pairs reinforce one another, while inconsistent or corrupted inputs are downweighted.

In practice, this paper implements several parallel attention heads, each of which discovers a distinct pattern of interaction between noise and physical attributes. The outputs of all heads are concatenated and then passed through a linear transformation to produce a unified fused embedding that reflects the physical context and resists measurement errors. By integrating cross-modal attention directly into the GATNE backbone (Figure 3), the estimator reconciles noisy inputs and power-flow constraints within a single end-to-end training framework.

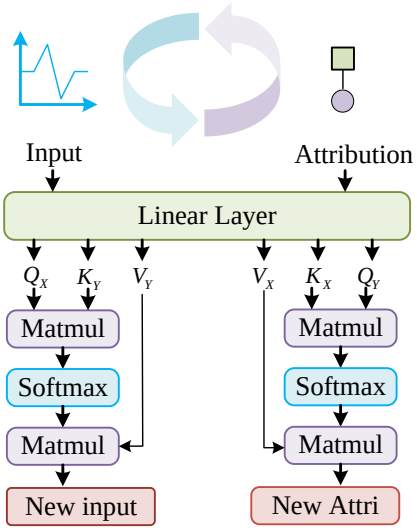


Figure 3. The cross-modal attention of z and attribution Att.

Experimentally, this design speeds up convergence because the network quickly learns which measurements deserve greater influence. It also improves the accuracy of the final estimation under conditions of high noise or partial observability. Moreover, the learned attention maps can be inspected after training, providing valuable insights into which sensors or measurements drive each part of the state estimate.

Suppose the noise inputs are $X = z = [z_1, z_2, \ldots, z_n]^T$ and the system attributes are $Y = Att = [Att_1, Att_2, \ldots, Att_m]^T$. After a linear transformation has been performed on them, we have the following:

$$\begin{cases} Q_X = W_X^Q X, K_X = W_X^K X, V_X = W_X^V X \\ Q_Y = W_Y^Q Y, K_Y = W_Y^K Y, V_Y = W_Y^V Y \end{cases} \tag{13}$$

where $W_X^Q, W_X^K, W_X^V, W_Y^Q, W_Y^Q, W_Y^Q$ are learnable parameters. Cross-modal attention is designed to capture interactions between measurement noise and system state attributes at both fine-grained and broad structural levels by using several parallel attention heads. Traditional approaches that focus on a single scale of interaction often struggle to reconcile local perturbations with the overall network topology. In our framework, learned projection matrices map raw noise and attribute vectors into a shared feature space, where their correlations can be computed more effectively. This transformation highlights the most informative signal components and suppresses spurious noise, leading to a more robust fusion of modalities. The multihead architecture further enables each attention path to specialize; for example, one head may focus on local voltage coupling while another captures long-distance power flow dependencies, but all operate within a single end-to-end trainable model. Empirically, this design accelerates convergence and yields higher estimation accuracy under noisy or partially observed conditions. Moreover, the resulting attention weights offer interpretable insights into which measurements and attributes drive each aspect of the state estimate.

Then, to compute the cross-modal attention scores, the feature vectors mapped by X and Y are utilized for calculation, as follows:

$$\begin{cases} score_X = softmax(\frac{Q_X K_Y^T}{\sqrt{d_{K_Y}}}) V_Y \\ score_Y = softmax(\frac{Q_Y K_X^T}{\sqrt{d_{K_X}}}) V_X \end{cases} \tag{14}$$

where $d_{K_{X/Y}}$ is the dimension of the key vector. The $softmax$ operation converts raw attention scores into a normalized weight distribution whose elements sum to one, enabling the cross-modal attention mechanism to emphasize salient features while attenuating irrelevant or noisy inputs. When this mechanism is embedded within the DSSE framework, heterogeneous data sources are more effectively fused, inter-modal dependencies are fully exploited, and the estimator adapts more robustly to the complexities of power distribution environments—ultimately improving both the accuracy and the reliability of the state estimates.

3.3. The Overall Model for DSSE

As illustrated in Figure 1b, this paper first applies Cross-Modal Attention to fuse the noise vector and node attribute features, then feeds the resulting embedding into the GATNE backbone to estimate the state variable x. The model is trained end-to-end against the weighted least-squares (WLS) objective. To reduce complexity, this paper restricts the analysis to the positive-sequence network, aggregating generator and load injections directly at each bus.

Input features are organized according to the WLS framework, where every measurement is represented by its observed value and associated uncertainty. Voltage angles appear as optional inputs to support synchronized phasor measurements when available but remain non-mandatory for most distribution systems, as they lack phasor measurement units. All other inputs—including component impedances, capacities and connection statuses—encode the network topology (see Table 1). This paper further introduces the Boolean indicators $\mathbb{L}_z$, $\mathbb{L}_s$, and $\mathbb{L}_{cl}$, which denote zero-injection buses, slack buses, and closed-line statuses, respectively.

Table 1. Features and Parameters of GATNE.

	Buses	**Lines**
Topology param	Bus port i Zero-inj $\mathbb{L}_z$ Slack bool $\mathbb{L}_s$	Line ports $(i.j)$ Closed line bool $\mathbb{L}_{cl}$ Phase-shift φ_{ij}
Input features	Voltage magn $V_i - \sigma_{V_i}$ Voltage angle $\phi_i - \sigma_{\phi_i}$ Active Power in $P_i - \sigma_{P_i}$ Reactive Power inj $Q_i - \sigma_{Q_i}$	Activate PF $P_{ij} - \sigma_{P_{ij}}$ Reactivate PF $Q_{ij} - \sigma_{Q_{ij}}$ Current magn $I_{ij} - \sigma_{I_{ij}}$ Line admittance Y_{ij} Shunt admittance $Y_{s_{ij}}$
Output features	Voltage magn V_i Voltage angle ϕ_i	

Because real-world distribution networks often suffer from sparse observability, this paper augments the limited field measurements with pseudo-measurements derived from historical load and generation profiles. These pseudo-measurements take the form of active and reactive injections P_i and Q_i at otherwise unobserved buses, thereby restoring full system observability. By embedding cross-modal attention within the GATNE-WLS architecture, the estimator is able to reconcile noisy or missing data with physical constraints, resulting in faster convergence, improved numerical stability and higher accuracy under realistic operating conditions.

4. Case Study

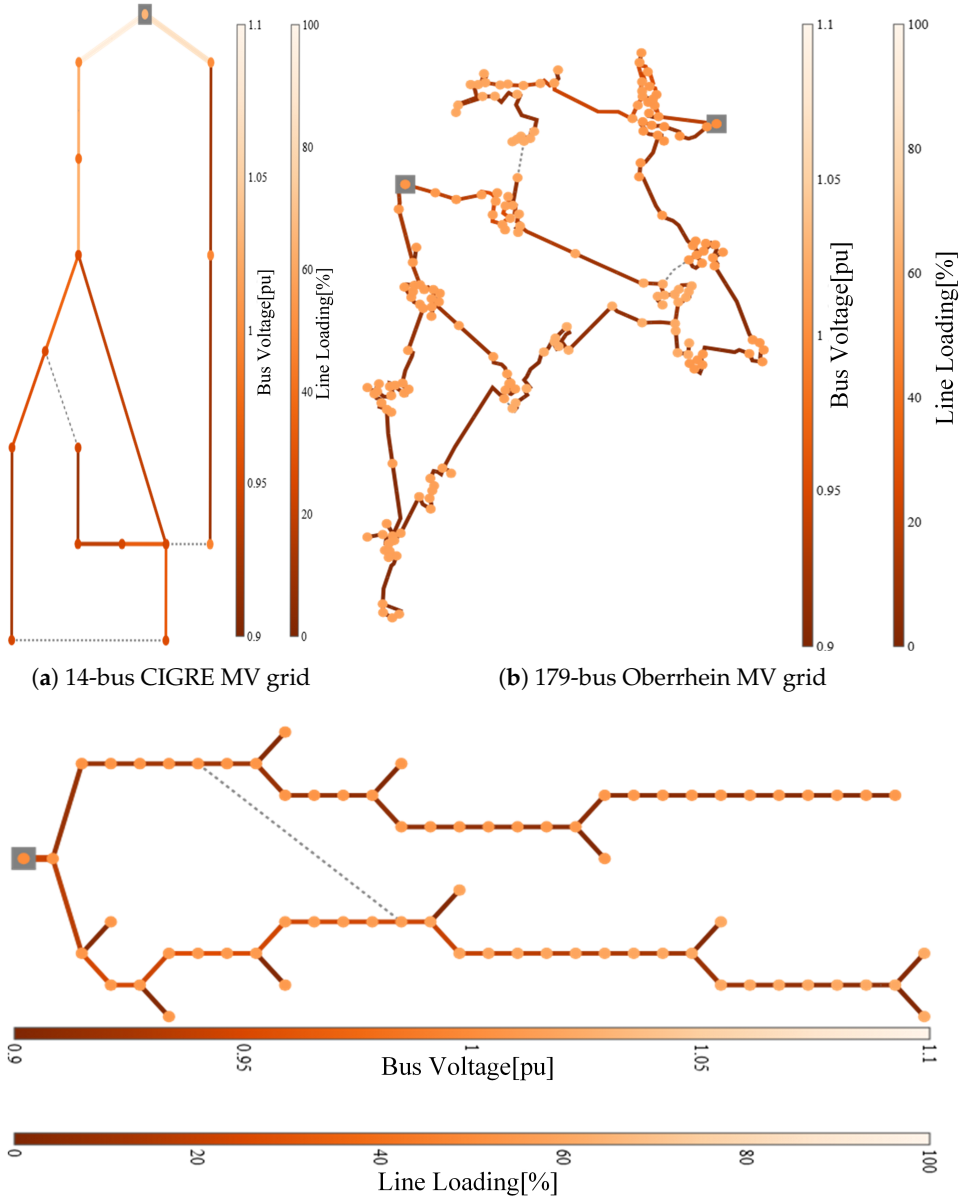
This paper will carry out tests on the 14-bus CIGRE MV distribution grid shown in Figure 4a activated with photovoltaic (PV) and wind distributed energy resources (DER) [39], the 179-bus Oberrhein grid (Figure 4b), and the 70-bus Oberrhein MV/LV sub-grid (Figure 4c) to provide information on the proposed method and evidence of its effectiveness [40]. A visualization of test cases is shown in Figure 4.

4.1. Experiment Setups

To capture realistic demand dynamics, 8640 hourly load samples were collected over a representative period of one year. Each scenario comprises 24 consecutive hourly snapshots, which reflect diurnal load cycles. These scenarios were synthesized by Monte Carlo perturbation of standard load curves, incorporating a 15% uncertainty margin to emulate both forecasting and measurement errors [41].

All simulations assume steady-state operation, with AC power flows solved using PandaPower [39] under Python 3.10. Matlab has significant advantages in numerical computation and visualization, making it suitable for simulations of small-scale distribution networks. However, it is less flexible than Python in terms of training deep learning models; Spice focuses more on circuit-level simulations and has low adaptability for system-level state estimation of distribution networks. Based on the above considerations, this study ultimately chose to use Python 3.10/PyTorch 2.7.0 for experiments. The experimental

environment employed in this study consists of an Intel Core i7-13700KF CPU, a single NVIDIA RTX 4090 GPU, 32GB of RAM, a 2TB SSD, and the Windows 11 operating system. Distribution-system bus injections are often dominated by spurious measurements with accuracies below 50%, so we adopt a conservative 1% error bound for voltage readings. Additive zero-mean white noise is applied across all sensors [42], yielding deviations of 0.5–2.0% in voltage and current measurements and 1–5% in active and reactive power injections. These modeling choices ensure that our DSSE evaluation faithfully reflects the uncertainty levels encountered in real-world distribution networks.



(**a**) 14-bus CIGRE MV grid (**b**) 179-bus Oberrhein MV grid

(**c**) 70-bus Oberrhein MV sub-grid

Figure 4. There are three test cases. The blue dots represent buses; the yellow area denotes the external grid; the green lines are transformers; the gray lines are transmission lines; and the gray dashed lines indicate disconnected branches.

The dataset is divided into a training set, a validation set, and a test set in a ratio of 8:1:1. Here, z denotes the input at the measurement locations, and the complete state of the system is represented by the label y.

One metric of the experiment is Root Mean Squared Error (RMSE) [43]:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{15}$$

where y_i is the actual values and $\hat{y}_i$ is predicted values of the i-th observation, n is the total number of observations.

The comparative models selected in this paper include the standard SE WLS, a supervised ANN model, Message Passing Neural Network(MPNN) [44], and the GATNE algorithm proposed in this paper. All these methods are implemented in PyTorch. The hyperparameter penalty factors were set as follows: batch size, 64; dropout rate, 0.4; learning rate lr, 0.003; and soft limitation μ fixed. The AdamMax optimizer is adopted, the grid search range is α, the layer dimension is d_l, and the number of layers is set as $l \in [2, 3, 4, 5]$. During the GATNE process, the model initializes using the Xavier uniform distribution with the dimension set to 40 and takes information such as bus voltage magnitude and active/reactive power injections as input features. Additionally, the number of attention heads is set to four; for the message-aggregation part, neighborhood mean pooling is adopted along with a learnable matrix. The selected hyperparameters are shown in Table 2.

Table 2. The hyperparameter values of datasets.

Parameters	14-Bus CIGRE	70-Bus Oberrhein	179-Bus Oberrhein
Epochs	600	800	1000
λ	0.8	0.8	0.8
α	0.006	0.006	0.006
batch size	64	64	64
d_l	40	40	40
l	3	3	3
dropout rate	0.4	0.4	0.4
μ	0.001	0.001	0.001
lr	0.003	0.003	0.003

4.2. Comparison Experiments

To evaluate the scalability and robustness of our approach, this paper conducts case studies on the IEEE 14-bus, 70-bus and 179-bus systems. Table 3 summarizes these studies.

Table 3. The Comparison Results of GATNE in the 14-bus CIGRE dataset.

Metrics	WLS	ANN	MPNN	GATNE
Voltage RMSE [10^{-3}]	9.53 (0.45)	2.75 (0.21)	6.78 (0.10)	4.32 (0.07)
Line Loading RMSE [%]	3.41 (0.06)	41.38 (0.76)	11.25 (0.19)	7.86 (0.37)
Line & Trafos Loading RMSE [%]	4.57 (0.05)	38.71 (0.51)	9.71 (0.05)	9.28 (0.23)
Minimum Voltage [p.u.]	0.941	0.945	0.948	0.957
Total Power Loss [MW]	0.168	0.161	0.157	0.154
Convergence [%]	100	100	100	100
Computation Time [s]	85 (36.71)	4.72 (1.03)	5.49 (0.97)	6.28 (1.22)

As shown in Table 3, on the 14-bus CIGRE dataset, in small-scale power grids, the voltage RMSE of GATNE is 4.32×10^{-3}, while that of the traditional WLS algorithm is 9.53×10^{-3}. GATNE's voltage RMSE is only 45.3% of WLS's. Compared with MPNN, which has a voltage RMSE of 6.78×10^{-3}, GATNE achieves better performance. This may be because MPNN treats edge features homogeneously without differentiation; in contrast, GATNE can better represent network heterogeneity, thereby achieving superior results. The

artificially designed ANN uses multi-layer MLPs for prediction, which is essentially the superposition of multiple linear fitting functions. With a sufficient number of layers, ANNs can more easily achieve better performance in this metric. However, in other indicators such as line-loading RMSE, the ANN scores 41.38%. This is likely due to the fact that purely data-driven ANNs do not incorporate physical constraints, leading to deviations of branch power estimation from actual operating laws. In contrast, GATNE integrates soft power-flow residual regularization, which significantly reduces the deviation in load estimation. To present the data in more detail, Figure 5a displays the voltage RMSE for each bus and the RMSE loading for each line. It can be observed that the voltage RMSE of GATNE is below the green dashed line (0.5) for all buses, indicating that GATNE meets the qualified performance standard.

Additionally, Figure 5b shows the load RMSE for different lines. In comparison, GATNE is better able to learn coupled data than the WLS and the ANN, which may be because cross-modal attention represents coupled data more effectively than other models. However, the RMSE estimation increases on lines indices 12 and 13, likely due to oversimplified modeling of transformers, which results in accuracy loss.

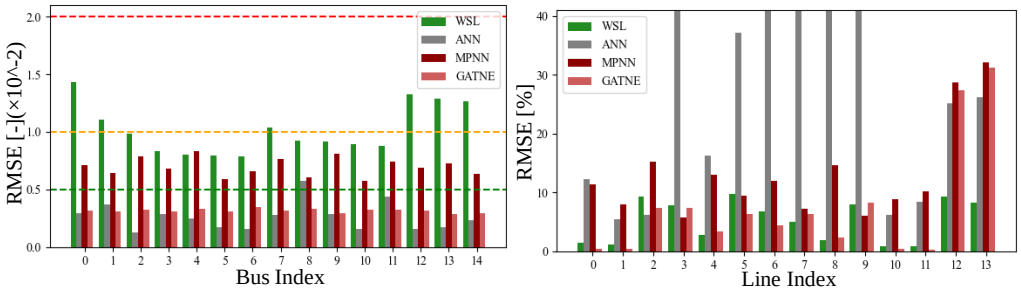


Figure 5. (**a**) The voltage level in the 14-bus network. The dashed lines show the acceptable values. Values above the red line are unacceptable, and those under the green line are acceptable. (**b**) The loading level in the 14-bus network.

In addition to the above metrics, convergence speed, accuracy, and computation time are also critical indicators, as shown in Table 4. In the 70-bus system, the voltage RMSE of GATNE (2.22×10^{-3}) is 92.6% lower than that of WLS (30.21×10^{-3}) and 49.1% lower than that of MPNN (4.36×10^{-3}). In the 179-bus system, the voltage RMSE of GATNE (2.97×10^{-3}) is 29.9% that of WLS (9.91×10^{-3}) and 41.9% lower than that of MPNN (5.11×10^{-3}). In large-scale networks, GATNE's cross-modal attention mechanism can fuse multi-source heterogeneous data such as distributed generation output and load fluctuations more efficiently. In contrast, MPNN relies solely on node message passing, making it difficult to handle information redundancy and noise interference in complex networks. In addition, in the comparison of the two indicators, computational efficiency and convergence rate, the WLS becomes less capable of full coverage as the network scale expands, easily leading to iterative divergence. The computational complexity of GATNE's graph-embedding process exhibits a linear relationship with the number of nodes, making it more suitable for use in modern distribution-network architectures.

The proposed GATNE algorithm outperforms WLS in all metrics. To validate the accuracy of the GATNE algorithm, this paper selects measured buses 24 and 85 in the 70-bus system to estimate voltage levels under normal sampling conditions using both WLS and GATNE, as shown in Figure 6.

As shown in Tables 3 and 4, the proposed GATNE reduces the computation time by 13 times, 4 times, and 24 times compared to WLS on the 14-bus, 70-bus, and 179-bus systems, respectively. The GATNE model outperforms both WLS and MPNN on large-scale

grid datasets (70-bus and 179-bus) in terms of various metrics. Specifically, the Voltage RMSE for GATNE is 2.91% in the 70-bus data set and 3.28% in the 179-bus data set, while MPNN achieves 4.36% and 5.11%, respectively. Both neural network models demonstrate superior performance compared to the traditional algorithms. This could be due to the ability of neural networks to effectively fit the data and produce better results. In contrast, the WLS algorithm, which is sensitive to redundancy and noise, performs poorly.

Table 4. The comparative results of GATNE in large datasets.

Metrics	70-Bus Oberrhein			179-Bus Oberrhein		
	WLS	**MPNN**	**GATNE**	**WLS**	**MPNN**	**GATNE**
Voltage RMSE [10^{-3}]	30.21 (0.97)	4.36 (0.06)	2.91 (0.03)	9.91 (0.30)	5.11 (0.12)	3.28 (0.04)
Line Loading RMSE [%]	16.73 (0.84)	5.29 (0.03)	2.22 (0.02)	5.87 (0.41)	5.90 (0.19)	2.97 (0.02)
Line & Trafos Loading RMSE [%]	38.96 (2.64)	5.77 (0.04)	2.46 (0.03)	4.14 (0.16)	3.56 (0.04)	3.48 (0.03)
Minimum Voltage [p.u.]	0.932	0.963	0.978	0.946	0.958	0.964
Total Power Loss [MW]	1.87	1.67	1.22	2.54	2.23	2.05
Convergence [%]	25	100	100	53	100	100
Computation Time [s]	122 (30.25)	29 (5.21)	32 (7.88)	1134 (356)	56 (12.13)	47 (10.38)

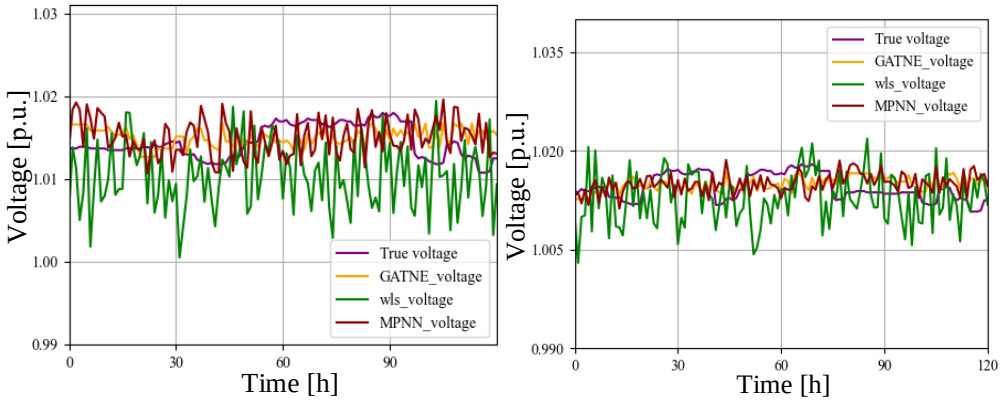


Figure 6. Voltage estimation at bus 24 (**a**) and bus 85 (**b**) in the 70-bus network under normal operating conditions.

In addition to the RMSE metrics, the convergence rate is also an important indicator of whether the algorithm can converge to the optimal solution within a given time. As shown in the table, GATNE achieves a convergence rate of 100% on both datasets, meaning that GATNE can quickly and stably find the optimal solution in each run, ensuring accuracy and stability in computation.

In contrast, the WLS method has a convergence rate of only 25% on the 70-bus Oberrhein dataset, indicating significant convergence issues when handling large-scale datasets. This likely occurs because, during the iteration process, it struggles to compute useful information, leading to a decline in performance. This can be attributed to the fact that for larger systems, WLS using the Newton–Raphson iterative method requires more iterations to converge. Although MPNN also achieves a convergence rate of 100%, GATNE demonstrates more stable convergence performance and is overall superior.

Minimum voltage and total power loss are core metrics for evaluating system operational safety and economic efficiency. Because the largest estimation errors occur in the WLS method, its calculated minimum voltage is the lowest across all three datasets, and it simultaneously yields the highest total power loss. The typical voltage safety threshold is 0.95 p.u., yet WLS results consistently fall below this critical value in all three cases.

In contrast, the GATNE-based method proposed in this work delivers state estimates closest to the true system state. Consequently, it achieves the highest minimum voltage values, aligning with physical expectations. Furthermore, WLS generates the highest total power-loss estimates, which may mislead dispatchers during optimization decision-making. Conversely, GATNE computes the lowest power loss values, significantly enhancing operational reliability.

As the network scale expands and noise inputs increase, WLS becomes more difficult to iterate, leading to a significant rise in computation time. Ultimately, owing to its superior accuracyin state estimation, the GATNE method derives more precise and robust operational metrics, demonstrating clear engineering advantages.

4.3. Noise Experiment

To verify the robustness of the model proposed in this paper, this paper conducts a comparison of measurement performance under noise interference on a 70-bus network. This paper directly adds Gaussian noise with a standard deviation of σ to the measured values and divides it into three different noise levels. The default noise refers to applying 1% noise to the voltage and current and adding 2% noise to the active and reactive power measurements; the low-level noise is 0.5% and 1%, and the high-level noise is 3% and 5%. Under the three different conditions, this paper presents a comparison between the traditional WLS algorithm and the GATNE algorithm. The comparison between the evaluation values of WLS and GATNE at bus 24 is shown in the Figure 7.

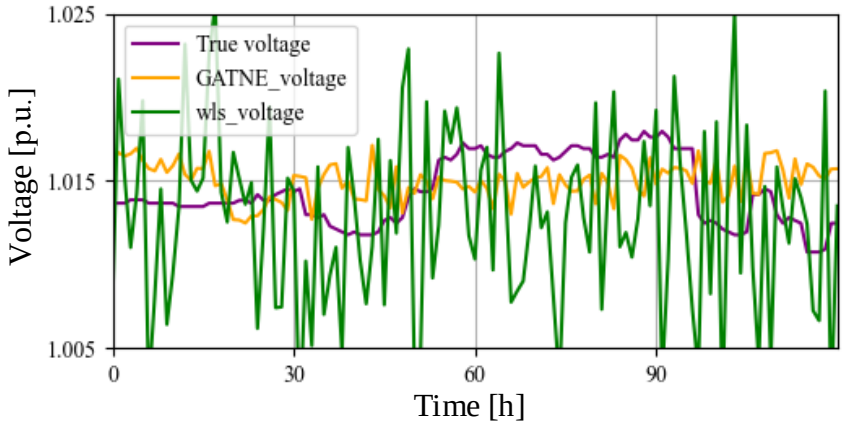
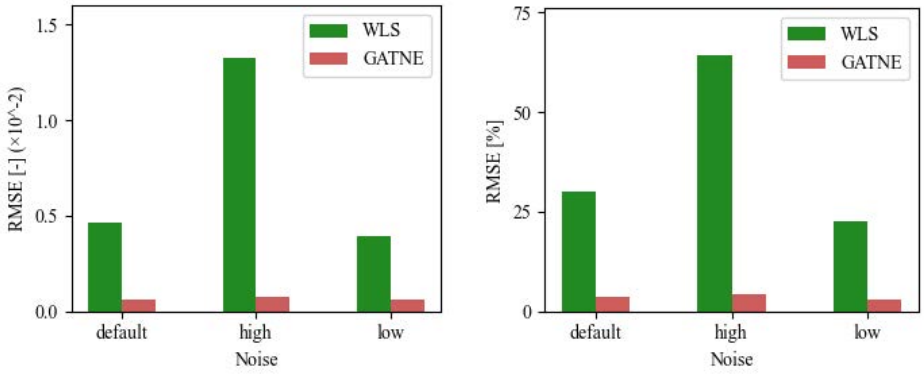


Figure 7. The voltage-measurement values of different algorithms at bus 24 in the 70-bus network under high-level noise.

In Figure 7, it can be seen that GATNE can effectively remove noise in the presence of noise, while WLS is vulnerable to noise, resulting in a large value for the measurement of the voltage deviation. In addition, the performance of voltage RMSE and line-loading RMSE for GATNE and WLS are shown in Figure 8. It is clearly evident from the figure that GATNE is more robust in the presence of noise.

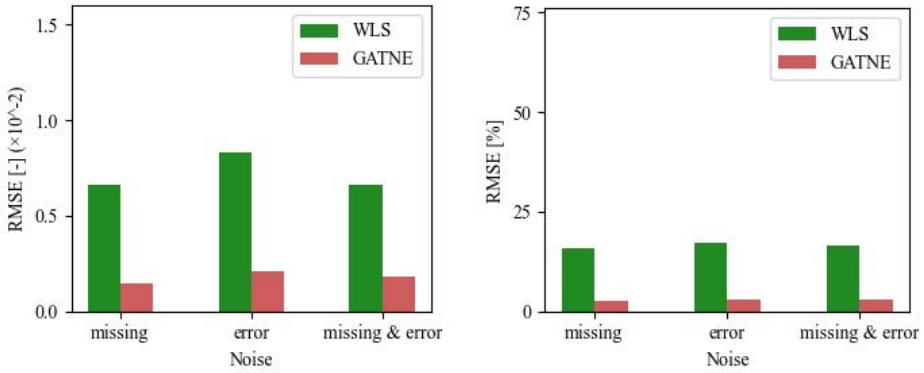
4.4. Missing Values and Error Measurements Experiments

In order to further verify the robustness of GATNE, this paper studies the stability of different algorithms when missing values and error measurements occur in the 70-bus network. The experimental steps in this paper are derived from the literature [1]. Under the same settings, as shown in Figure 9, GATNE performs better than WLS, which fully demonstrates that GATNE is more robust and less affected by noise, missing values, and incorrect values. This may be because the graph-aggregation algorithm of GATNE can better analyze the network topology and the existence of the attention mechanism can automatically assign weights, resulting in more robustness for GATNE.



(**a**) (**b**)

Figure 8. Comparison of RMSE between GATNE and WLS under different noise conditions, where (**a**) represents the voltage RMSE and (**b**) represents the line-loading RMSE.



(**a**) (**b**)

Figure 9. Comparison of RMSE between GATNE and WLS under conditions of missing values and error measurements, where (**a**) represents the voltage RMSE and (**b**) represents the line-loading RMSE.

4.5. Hyperparameters Analysis

In this section, this paper analyzed the hyperparameters mentioned in Section 2.3, which mainly include penalty term hyperparameters λ and the accuracy deviation μ of power-flow equations.

This paper sets the physical penalty term as follows: $\lambda = \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5$. This is because such a setup aims to strike a balance among physical consistency, computational efficiency, and model stability. If there are significant differences between different hyperparameters, certain constraints may be over-emphasized or under-emphasized, thereby disrupting the inherent balance between physical laws. For instance, if positive voltage constraints $\max(0, V - 1.05)$ are set to be far stronger than negative $\max(0, 0.95 - V)$, that could lead to distorted model results. Additionally, in large-scale systems like the 70-bus and 179-bus networks, individual hyperparameters would significantly increase computational costs.

Then, λ_0 is set to 1. As shown in Table 5, the performance of voltage RMSE on the 14-bus CIGRE dataset under different hyperparameters is presented. In general, adopting a value of $\lambda = 0.8$ yields the best results. For more detailed optimization, a grid-search method should be employed to select hyperparameter combinations.

Table 5. The different λ values of 14-CIGRE validation.

λ Value	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Voltage RMSE [10^{-3}]	6.47	6.04	5.93	6.04	5.74	5.32	4.96	4.58	4.32	4.51	4.79
Line Loading RMSE [%]	14.35	15.62	13.01	11.32	10.84	9.28	8.89	8.09	7.86	8.13	8.92

Additionally, to validate the role of the soft constraints proposed in this paper, we preform validation. Specifically, Table 6 presents the performance of GATNE on the 14-bus CIGRE dataset under different values of μ. As shown in the table, the value of μ should neither be too high nor too low. When μ is too low, the constraints on the model are weak, leading to larger result deviations. When μ is too high, the constraints on the model become overly strict, which also causes a decline in performance.

Table 6. The different μ values of 14-CIGRE validation.

μ Value	0	0.001	0.005	0.01	0.5	1
Voltage RMSE [10^{-3}]	5.31	4.32	5.77	6.42	7.63	8.69
Line Loading RMSE [%]	8.23	7.86	9.63	10.28	11.51	13.22

4.6. Ablation Studies

To validate that the proposed model, adopted modules, and loss function have positive effects, this paper conducts ablation studies on each component. The baseline is a model built with GNN, and the specific data are shown in Table 7.

Table 7. The Ablation Study on the Proposed GATNE.

Ablation Settings			Voltage RMSE [10^{-3}]	Line Loading RMSE [%]
GATNE	Physical Soft Constraints	Cross-Modal Attention		
			8.59	14.60
✓			6.16	11.23
	✓		7.82	9.47
		✓	6.65	10.25
✓	✓		5.87	8.06
✓		✓	4.95	8.61
	✓	✓	5.14	7.98
✓	✓	✓	4.32	7.86

The baseline model (with all components disabled) achieves a voltage RMSE of 8.59×10^{-3} and a line-loading RMSE of 14.60%, reflecting the estimation accuracy when one relies solely on the basic GNN architecture. When the GATNE framework is enabled individually (the second row), voltage RMSE and line-loading RMSE decrease to 6.16×10^{-3} and 11.23%, respectively, indicating that GATNE optimizes information aggregation between nodes through graph attention mechanisms, significantly enhancing the model's ability to model the graph structure of power grids. Further introduction of physical soft constraints (the third row) reduces line-loading RMSE substantially to 9.47%, while voltage RMSE drops to 7.82×10^{-3}. This demonstrates that physical constraints, by embedding prior knowledge such as Kirchhoff's laws and power conservation, effectively regulate the physical rationality of model predictions, particularly yielding more pronounced optimization for system-level constrained indicators like line loading.

When cross-modal attention is introduced individually (the fourth row), the voltage RMSE decreases to 6.65×10^{-3} and the line load RMSE drops to 10.25%, demonstrating the

advantage of this module in the integration of multimodal data features such as voltage, current, and line parameters. This validates that cross-modal interaction plays a more critical role in improving the estimation accuracy of node-level states. When the GATNE framework is combined with physical soft constraints (the fifth row), line-loading RMSE further decreases to 8.06%, showing that the integration of data-driven graph modeling and physical priors forms a complementarity at the system-level constraint dimension. Conversely, the combination of GATNE and cross-modal attention (the sixth row) reduces voltage RMSE to 4.95×10^{-3}, indicating that graph structure modeling and multi-source feature fusion achieve synergistic optimization for node-state estimation.

Notably, even without enabling the GATNE framework, the combination of physical soft constraints and cross-modal attention (the seventh row) still reduces both metrics, though performance lags behind configurations including GATNE, highlighting that GATNE serves as the foundational architecture supporting the effectiveness of other modules. When all components are enabled (the eighth row), the voltage RMSE and the line-loading RMSE reach 4.32×10^{-3} and 7.86%. This validates the synergistic enhancement of the GATNE framework, physical soft constraints, and cross-modal attention. Specifically, GATNE enables efficient graph-structure representation learning, physical constraints ensure that predictions remain consistent with grid operation laws, and cross-modal attention strengthens the interaction between input z and node attributes $Attr$, collectively yielding a high-precision and robust state estimation model.

In summary, the ablation experiments demonstrate that each component makes an indispensable positive contribution to model performance and that their combination achieves optimal estimation through mechanistic complementarity, thereby providing a reliable modular design basis for state estimation in complex power grid environments.

5. Conclusions

This paper proposes a distribution system state estimation (DSSE) method using GATNE and physical constraints. As a graph-structured dedicated embedding model, GATNE can automatically encode the topological relationships of the distribution network into vectors, eliminating the need for manual definition of topology matrices. Furthermore, the method we proposed leverages the capability for weighted information aggregation inherent to GATNE: in the case of sparse measurement data, it can infer the states of unmeasured nodes using a small amount of high-value measurement data, thereby significantly improving the estimation accuracy. The method is validated on the 14-bus CIGRE, 70-bus Oberrhein, and 179-bus Oberrhein datasets based on Pandapower, and ablation experiments confirm the effectiveness of the adopted components. GATNE demonstrates higher efficiency in graph modeling than traditional GNN models, which allows it to perform better in large-scale networks. Additionally, through the incorporation of power-flow equations as constraints, the model's generalizability is improved, and the method aligns more closely with the physical characteristics of distribution systems.

Although this approach offers advantages, it still has limitations. The current method primarily focuses on node aggregation learning, and attribute relationships are modeled only via cross-modal attention for interactive learning, leaving edge features incompletely captured. Furthermore, while the method proposed in this paper has incorporated certain considerations of uncertainties in the system, it has not fully covered their relevant impacts, and at the same time, has not comprehensively integrated the inherent temporal characteristics of modern distribution networks. In future work, we will further strengthen the treatment of uncertainty characteristics, fully incorporate relevant temporal effects, and explore more robust and generalizable DSSE modeling methods. We also plan to develop a graph modeling system with stronger dynamic topology adaptation to address

unbalanced modeling tasks, thereby enhancing the model's generalizability, effectiveness, and robustness. Meanwhile, cross-industry studies have also demonstrated that multi-energy coordination and supply chain optimization play a crucial role in achieving carbon reduction goals. For example, explorations in ammonia industry chain planning [45] and low-carbon pathways in the steel sector [46] both highlight the deep coupling between the power system and other energy systems. In the future, our approach can be integrated with such multi-energy system planning research to support the safe, efficient, and low-carbon operation of new power systems under the "dual carbon" strategy.

Author Contributions: Conceptualization, S.L. and Z.T.; methodology, S.L.; software, Z.T.; validation, Z.Z. and S.L.; formal analysis, S.L. and Z.T.; investigation, S.L.; resources, B.C.; data curation, S.L.; writing—original draft preparation, S.L.; writing—review and editing, S.L., Z.Z. and B.C.; visualization, S.L.; supervision, B.C.; project administration, B.C.; funding acquisition, B.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was Supported by the National Key Research and Development Program of China (2022YFB2404200).

Data Availability Statement: The data presented in this study are available on request from the corresponding author. Due to privacy, the data cannot be accessed publicly.

Conflicts of Interest: Authors Siyan Liu, Zhuang Tang and Bo Chai were employed by the company China Electric Power Research Institute Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Habib, B.; Isufi, E.; van Breda, W.; Jongepier, A.; Cremer, J.L. Deep statistical solver for distribution system state estimation. *IEEE Trans. Power Syst.* **2023**, *39*, 4039–4050. [CrossRef]
2. Lourenço, E.M.; London, J.B.A. (Eds.) *Power Distribution System State Estimation*; IET: Stevenage, UK , 2022; Volume 183.
3. Zhai, B.; Yang, D.; Zhou, B.; Li, G. Distribution System State Estimation Based on Power Flow-Guided GraphSAGE. *Energies* **2024**, *17*, 4317. [CrossRef]
4. Della Giustina, D.; Pau, M.; Pegoraro, P.A.; Ponci, F.; Sulis, S. Electrical distribution system state estimation: Measurement issues and challenges. *IEEE Instrum. Meas. Mag.* **2014**, *17*, 36–42. [CrossRef]
5. Cao, D.; Zhao, J.; Hu, W.; Yu, N.; Hu, J.; Chen, Z. Physics-informed graphical learning and Bayesian averaging for robust distribution state estimation. *IEEE Trans. Power Syst.* **2023**, *39*, 2879–2892. [CrossRef]
6. Ali, M.; Dimitrovski, A.; Qu, Z.; Sun, W. A voltage inference framework for real-time observability in active distribution grids. In Proceedings of the 2023 IEEE Power & Energy Society General Meeting (PESGM), Orlando, FL, USA, 16–20 July 2023; pp. 1–5.
7. Lin, S.; Zhu, H. Enhancing the spatio-temporal observability of grid-edge resources in distribution grids. *IEEE Trans. Smart Grid* **2021**, *12*, 5434–5443. [CrossRef]
8. Fotopoulou, M.; Petridis, S.; Karachalios, I.; Rakopoulos, D. A review on distribution system state estimation algorithms. *Appl. Sci.* **2022**, *12*, 11073. [CrossRef]
9. Carvalho, B.; Bretas, N. Analysis of the largest normalized residual test robustness for measurements gross errors processing in the WLS state estimator. *J. Syst. Cybern. Inform.* **2013**, *11*, 1–6.
10. Roux, E.; Cauhope, M.; Bonnet, M.P.; Calmant, S.; Vauchel, P.; Seyler, F. Daily water stage estimated from satellite altimetric data for large river basin monitoring/Estimation de hauteurs d'eau journalières a partir de données d'altimétrie radar pour la surveillance des grands basins fluviaux. *Hydrol. Sci. J.* **2008**, *53*, 81–99. [CrossRef]
11. Mili, L.; Phaniraj, V.; Rousseeuw, P.J. Least median of squares estimation in power systems. *IEEE Trans. Power Syst.* **2002**, *6*, 511–523. [CrossRef]
12. Amini, M.; Roozbeh, M. Least trimmed squares ridge estimation in partially linear regression models. *J. Stat. Comput. Simul.* **2016**, *86*, 2766–2780. [CrossRef]
13. Jin, X.B.; Jeremiah, R.J.R.; Su, T.L.; Bai, Y.T.; Kong, J.L. The new trend of state estimation: From model-driven to hybrid-driven methods. *Sensors* **2021**, *21*, 2085. [CrossRef]
14. Kumari, N.; Kulkarni, R.; Ahmed, M.R.; Kumar, N. Use of kalman filter and its variants in state estimation: A review. In *Artificial Intelligence for a Sustainable Industry 4.0*; Springer: Cham, Switzerland, 2021; pp. 213–230.

15. Pegoraro, P.A.; Angioni, A.; Pau, M.; Monti, A.; Muscas, C.; Ponci, F.; Sulis, S. Bayesian approach for distribution system state estimation with non-Gaussian uncertainty models. *IEEE Trans. Instrum. Meas.* **2017**, *66*, 2957–2966. [CrossRef]
16. Gao, M.; Zhou, S.; Gu, W.; Wu, Z.; Liu, H.; Zhou, A.; Wang, X. MMGPT4LF: Leveraging an optimized pre-trained GPT-2 model with multi-modal cross-attention for load forecasting. *Applied Energy* **2025**, *392*, 125965. [CrossRef]
17. Raghuvamsi, Y.; Teeparthi, K. A review on distribution system state estimation uncertainty issues using deep learning approaches. *Renew. Sustain. Energy Rev.* **2023**, *187*, 113752. [CrossRef]
18. Yarlagadda, R.; Kosana, V.; Teeparthi, K. Power system state estimation and forecasting using cnn based hybrid deep learning models. In Proceedings of the 2021 IEEE International Conference on Technology, Research, and Innovation for Betterment of Society (TRIBES), Raipur, India, 17–19 December 2021; pp. 1–6.
19. Zamzam, A.S.; Fu, X.; Sidiropoulos, N.D. Data-driven learning-based optimization for distribution system state estimation. *IEEE Trans. Power Syst.* **2019**, *34*, 4796–4805. [CrossRef]
20. He, Y.; Chai, S.; Xu, Z. A novel approach for state estimation using generative adversarial network. In Proceedings of the 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, 6–9 October 2019; pp. 2248–2253.
21. Mestav, K.R.; Luengo-Rozas, J.; Tong, L. Bayesian state estimation for unobservable distribution systems via deep learning. *IEEE Trans. Power Syst.* **2019**, *34*, 4910–4920. [CrossRef]
22. Kundacina, O.; Cosovic, M.; Vukobratovic, D. State estimation in electric power systems leveraging graph neural networks. *arXiv* **2022**, arXiv:2201.04056. [CrossRef]
23. Zamzam, A.S.; Sidiropoulos, N.D. Physics-aware neural networks for distribution system state estimation. *IEEE Trans. Power Syst.* **2020**, *35*, 4347–4356. [CrossRef]
24. Zhang, L.; Wang, G.; Giannakis, G.B. Real-time power system state estimation and forecasting via deep unrolled neural networks. *IEEE Trans. Signal Process.* **2019**, *67*, 4069–4077. [CrossRef]
25. Kumar, D.M.V.; Srivastava, S.C.; Shah, S.; Mathur, S. Topology processing and static state estimation using artificial neural networks. *IEE Proc.-Gener. Transm. Distrib.* **1996**, *143*, 99–105. [CrossRef]
26. Oliveira, R.; Bessa, R.; Iranda, V.M. Identifying topology in power networks in the absence of breaker status sensor signals. In Proceedings of the 2018 19th IEEE Mediterranean Electrotechnical Conference (MELECON), Marrakech, Morocco, 2–7 May 2018; pp. 160–165.
27. Duan, N.; Stewart, E.M. Deep-learning-based power distribution network switch action identification leveraging dynamic features of distributed energy resources. *IET Gener. Transm. Distrib.* **2019**, *13*, 3139–3147. [CrossRef]
28. Gotti, D.; Amaris, H.; Larrea, P.L. A deep neural network approach for online topology identification in state estimation. *IEEE Trans. Power Syst.* **2021**, *36*, 5824–5833. [CrossRef]
29. Ostrometzky, J.; Berestizshevsky, K.; Bernstein, A.; Zussman, G. Physics-informed deep neural network method for limited observability state estimation. *arXiv* **2019**, arXiv:1910.06401.
30. Wang, L.; Zhou, Q.; Jin, S. Physics-guided deep learning for power system state estimation. *J. Mod. Power Syst. Clean Energy* **2020**, *8*, 607–615. [CrossRef]
31. Battaglia, P.W.; Hamrick, J.B.; Bapst, V.; Sanchez-Gonzalez, A.; Zambaldi, V.; Malinowski, M.; Tacchetti, A.; Raposo, D.; Santoro, A.; Faulkner, A.; et al. Relational inductive biases, deep learning, and graph networks. *arXiv* **2018**, arXiv:1806.01261. [CrossRef]
32. Hadou, S.; Kanatsoulis, C.I.; Ribeiro, A. Space-time graph neural networks with stochastic graph perturbations. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
33. Lin, H.; Sun, Y. ElegNN: Electrical-model-guided graph neural networks for power distribution system state estimation. In Proceedings of the GLOBECOM 2022—2022 IEEE Global Communications Conference, Rio de Janeiro, Brazil, 4–8 December 2022; pp. 5292–5298.
34. Madbhavi, R.; Natarajan, B.; Srinivasan, B. Graph neural network-based distribution system state estimators. *IEEE Trans. Ind. Inform.* **2023**, *19*, 11630–11639. [CrossRef]
35. Ngo, Q.H.; Nguyen, B.L.; Vu, T.V.; Zhang, J.; Ngo, T. Physics-informed graphical neural network for power system state estimation. *Appl. Energy* **2024**, *358*, 122602. [CrossRef]
36. Cao, D.; Zhao, J.; Hu, W.; Liao, Q.; Huang, Q.; Chen, Z. Topology change aware data-driven probabilistic distribution state estimation based on Gaussian process. *IEEE Trans. Smart Grid* **2022**, *14*, 1317–1320. [CrossRef]
37. Cen, Y.; Zou, X.; Zhang, J.; Yang, H.; Zhou, J.; Tang, J. Representation learning for attributed multiplex heterogeneous network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 1358–1368.
38. Abur, A.; Exposito, A.G. *Power System State Estimation: Theory and Implementation*; CRC Press: Boca Raton, FL, USA, 2004.
39. Force, C.T. Benchmark systems for network integration of renewable and distributed energy resources. *E-cigre* **2014**, *63*.

40. Thurner, L.; Scheidler, A.; Schäfer, F.; Menke, J.H.; Dollichon, J.; Meier, F.; Meinecke, S.; Braun, M. pandapower—An open-source python tool for convenient modeling, analysis, and optimization of electric power systems. *IEEE Trans. Power Syst.* **2018**, *33*, 6510–6521. [CrossRef]
41. Rudion, K.; Orths, A.; Styczynski, Z.A.; Strunz, K. Design of benchmark of medium voltage distribution network for investigation of DG integration. In Proceedings of the 2006 IEEE Power Engineering Society General Meeting, Montreal, QC, Canada, 18-22 June 2006; p. 6.
42. Fan, J.; Zhuang, W.; Xia, M.; Fang, W.; Liu, J. Optimizing attention in a transformer for multihorizon, multienergy load forecasting in integrated energy systems. *IEEE Trans. Ind. Inform.* **2024**, *20*, 10238–10248. [CrossRef]
43. Zhuang, W.; Fan, J.; Xia, M.; Zhu, K. A multi-scale spatial–temporal graph neural network-based method of multienergy load forecasting in integrated energy system. *IEEE Trans. Smart Grid* **2023**, *15*, 2652–2666. [CrossRef]
44. Gilmer, J.; Schoenholz, S.S.; Riley, P.F.; Vinyals, O.; Dahl, G.E. Neural message passing for quantum chemistry. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1263–1272.
45. Liu, H.; Zhou, S.; Gu, W.; Zhuang, W.; Gao, M.; Chan, C.C.; Zhang, X. Coordinated planning model for multi-regional ammonia industries leveraging hydrogen supply chain and power grid integration: A case study of Shandong. *Applied Energy* **2025**, *377*, 124456. [CrossRef]
46. Yang, F.; Meng, F.; Qiu, Y.; Zhou, S.; Zhuang, W.; Liu, H.; Gu, W.; Yang, Y. Multi-dimensional assessment of decarbonization technologies and pathways in China's iron and steel industry: An energy-process chain perspective. *Energy Strategy Reviews* **2025**, *61*, 101810. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.