Title
Zero-Shot Transfer Learning for Three-Phase Unbalanced Power Distribution Systems

Permalink

Journal
Proceedings of the PowerUp Conference

Authors
Wu, Tong
Scaglione, Anna
Miguel, Sandy
et al.

Publication Date
2026-10-01

Zero-Shot Transfer Learning for Three-Phase Unbalanced Power Distribution Systems

Tong Wu[1], Anna Scaglione[2], Sandy Miguel[1], Yanqing Tao[2], Daniel Arnold[3]

[1] *University of Central Florida* [2] *Cornell University* [3] *Lawrence Livermore National Laboratory*

Suggested Citation

T. Wu, A. Scaglione, S. Miguel, Y. Tao, and D. Arnold, "Zero-Shot Transfer Learning for Three-Phase Unbalanced Power Distribution Systems," in *Proceedings of the PowerUp Conference*, 2026.

BibTeX Entry

```
@inproceedings{wu2026zeroshot,
  author = {Wu, Tong and Scaglione, Anna and Miguel, Sandy and Tao, Yanqing and
      Arnold, Daniel},
  title = {Zero-Shot Transfer Learning for Three-Phase Unbalanced Power Distribution
       Systems},
  booktitle = {Proceedings of the PowerUp Conference},
  year = {2026},
}
```

Zero-Shot Transfer Learning Across Reconfigurations for Three-Phase Unbalanced Power Distribution Systems

Tong Wu, Anna Scaglione, Sandy Miguel, Yanqing Tao, and Daniel B. Arnold

***Abstract*—Distribution utilities increasingly rely on data-driven methods for state estimation and forecasting as large-scale renewable energy integration causes rapid state changes, yet most learning-based approaches remain tied to specific feeder topologies and require retraining when networks are reconfigured. In practice, retraining models for every feeder and operational configuration is infeasible due to limited data availability, modeling effort, and computational costs. This paper presents a physics-aware graph learning framework for three-phase unbalanced distribution networks that enables zero-shot transfer across unseen feeder reconfigurations, with potential extension to feeders sharing similar electrical and topological statistics. By combining scalar-weight graph convolutions grounded in power flow physics with adaptive graph pooling, the proposed model learns topology-invariant representations while accommodating variable network sizes. Case studies on the IEEE 123-bus unbalanced distribution feeder using AMI-based measurements demonstrate accurate state estimation and short-term forecasting under unseen reconfigurations, highlighting the promise of the approach for reconfigurable distribution systems.**

***Index Terms*—Zero-shot transferable learning, graph convolutional neural networks**

I. Introduction

Distribution networks are subject to frequent topology changes due to switching operations, fault isolation, maintenance, and incremental expansion. While Advanced Metering Infrastructures (AMIs) enabled the collection of fine-grained voltage and power measurements, the exploitation of these data streams for real-time state estimation remains challenging, particularly as distributed renewable energy resources introduce rapid and uncertain variations in network states. Classical distribution system state estimation (DSSE) relies on detailed feeder models and nonlinear optimization based on AC power flow equations [1]. Recently, learning-based approaches have shown promise in improving DSSE scalability and computational efficiency, as well accuracy when interpolate the state from sparse measurements [2], [3]. These benefits come from learning the posterior distribution of the state, going beyond what is observable based on the AC power flow equations.

A critical limitation of most machine learning approaches for DSSE is their lack of transferability across feeder reconfigurations [4], [5]. Even modest changes in topology alter both the dimension and the statistical distribution of the state variables, rendering pretrained models invalid. Existing transfer learning and fine-tuning strategies typically require labeled data or retraining for each new configuration [6], [7], which is operationally impractical for many utilities. In particular, retraining feeder-specific models using simulated data requires detailed and continuously updated feeder models, as well as significant computational resources. Recent work has explored graph-based learning as a natural way to incorporate network structure into power system inference tasks [3], [8]–[10]. However, most graph neural network (GNN) models remain tied to a fixed topology or assume identical input and output dimensions across systems, limiting their applicability in reconfigurable distribution networks [7], [11].

This paper proposes a transferable, physics-aware graph learning framework for DSSE in three-phase unbalanced power systems that achieves zero-shot generalization across unseen feeder reconfigurations without retraining. The framework integrates three technical innovations: (1) By modeling three-phase voltages as complex-valued graph signals coupled through the multiphase admittance matrix, scalar-weight graph convolutional operators exploit the locality and scale-invariance of power flow physics across all three phases simultaneously. (2) Adaptive graph pooling combined with parallel decoding and positional encoding handles variable network dimensions, enabling a single trained model to generalize across unseen reconfigurations and, in principle, feeders with similar electrical and topological statistics. (3) The framework's transferability is theoretically grounded in the spectral concentration of three-phase admittance matrices in tree-structured distribution networks, where the Schur complement converges to a statistically consistent structure despite phase imbalances.

II. Distribution Network Modeling and Graph Signal Processing Framework

Consider a collection of distribution feeder configurations indexed by $q \in \mathcal{Q}$, where each configuration represents a distinct network topology resulting from switching operations, reconfigurations, or different feeders. Each feeder configuration q is modeled as an undirected weighted graph $G_q = (V_q, E_q)$, where buses are nodes and distribution lines are edges, with $N_q := |V_q|$ denoting the number of buses in configuration q. The physical constraints of the system are governed by Kirchhoff's and Ohm's laws:

$$\mathbf{i}_t = \mathbf{Y}_q \mathbf{v}_t, \quad v_{n,t} = |v_{n,t}| e^{j\phi^v_{n,t}}, i_{n,t} = |i_{n,t}| e^{j\phi^i_{n,t}}, \ \forall n \in V_q \tag{1}$$

Tong Wu and Sandy Miguel are with the Department of Electrical and Computer Engineering, University of Central Florida, Orlando, FL, 32816 USA. Email: {tong.wu, sandy.miguel}@ucf.edu. Anna Scaglione and Yanqing Tao are with the Department of Electrical and Computer Engineering, Cornell Tech, Cornell University, New York City, NY, 10044 USA. Email: {as337, yt652}@cornell.edu. Daniel B. Arnold is with Lawrence Livermore National Laboratory. Email: arnold37@llnl.gov.

This work was supported in part by the National Science Foundation (NSF) under Grant NSF ECCS 2210012, and in part by the Cybersecurity, Energy Security, and Emergency Response (CESER), Risk Management and Tools and Technologies (RMT) Program of the U.S. Department of Energy through the "Mitigation via Analytics for Inverter-Grid Cybersecurity (MAGIC)" Project under contract DE-AC02-05CH11231.

where $\mathbf{v}_t \in \mathbb{C}^{N_q}$ and $\mathbf{i}_t \in \mathbb{C}^{N_q}$ denote the bus voltage and net current phasor vectors at time t, and $\mathbf{Y}_q \in \mathbb{C}^{N_q \times N_q}$ is the admittance matrix for configuration q. For any given apparent power injection vector $\mathbf{s}_t$, the AC power flow equations define the manifold on which the voltage phasors are constrained:

$$\mathbf{s}_t = \mathrm{diag}(\mathbf{v}_t)\mathbf{Y}_q^* \mathbf{v}_t^* = \mathrm{diag}^*(\mathbf{Y}_q \mathbf{v}_t \mathbf{v}_t^H) \tag{2}$$

While the geometry of the voltage phasor manifold is complex, simple reasoning reveals some algebraic constraints implied by (2). For instance, intermediate buses (those with zero net injection) impose the constraint that the voltage vector must lie in the null space of the corresponding rows of the admittance matrix $\mathbf{Y}_q$. Since intermediate buses typically constitute 20 to 30% of nodes in radial distribution systems, the voltage vector manifold effectively lies in a subspace of dimensionality 70 to 80% of the total number of buses.

We argue in Section III-A that the effective admittance matrix obtained after eliminating these intermediate nodes, specifically the Schur complement of $\mathbf{Y}_q$, exhibits spectral concentration across different feeder configurations, providing the theoretical foundation for zero-shot transferability.

A. AMI-driven Nonlinear Coarse State Estimation

The objective of DSSE is to estimate the complex voltage phasors $\mathbf{v}_t \in \mathbb{C}^{N_q}$ form, typically, AMI measurements, which include typically active and reactive power as well as voltage and current magnitudes. The measurement model is therefore non-linear, i.e. of the type

$$\mathbf{m}_t = \mathbf{h}(\mathbf{v}_t) + \varepsilon_t, \tag{3}$$

where $\mathbf{h}(\cdot)$ maps bus voltages to measured quantities (e.g., power injections, power flows, voltage magnitudes) and ε_t is measurement noise. A representative component of $\mathbf{h}(\mathbf{v}_t)$ is the complex power injection $[\mathbf{s}_q]_m$ at a measurement bus.

Typically, AMI are sparsely deployed and, for the most part, are at the consumer points. A first level of adaptation in our framework is achieved by mapping the measurements into a coarse estimate of the state vector to be used as the input feature for our neural network model, so that the type of input of the neural network is homogeneous across systems. Utilities may have their own DSSE but for training purposes it will be necessary to emulate their operations using simulated values. For instance, a reasonably robust solution can be attained solving the following regularized least-square problem

$$\min_{\mathbf{v}_t} \left\| \mathbf{R}^{1/2} \left(\mathbf{m}_t - \mathbf{h}(\mathbf{v}_t) \right) \right\|_2^2 + \left\| \mathbf{\Lambda}^{1/2} \mathbf{v}_t \right\|_2^2, \tag{4}$$

where $\mathbf{R}$ accounts for measurement noise variances and $\mathbf{\Lambda}$ is the regularization matrix. To enable closed form solution, we linearize the measurement function $\mathbf{h}(\cdot)$ using the log3v transformation [12]. This linearization introduces an approximation error relative to the full nonlinear power flow model; in our pipeline, the resulting estimate is used only as a coarse physics-informed input, and the learned correction is evaluated end-to-end against nonlinear OpenDSS-generated states. Solving (4) yields an estimate $\mathbf{x}_{q,t}^{(0)} \in \mathbb{C}^{N_q}$ that serves as the input feature for our model. The training samples are therefore pairs $(\mathbf{x}_{q,t}^{(0)}, \mathbf{v}_{q,t})$, and the goal is to estimate the state given the temporal history $(\mathbf{x}_{q,t}^{(0)}, \ldots, \mathbf{x}_{q,t-T}^{(0)})$ and the feeder topology of configuration q. Because the learned filters operate on local Kirchhoff-consistent residuals, the architecture is naturally robust to moderate measurement noise and topology perturbations. The key challenge is that N_q varies across configurations due to feeder reconfigurations, outages, and expansions; hence both the input and output dimensions change with the specific grid, indexed by q.

B. Physics-Aware Graph Representation

To address the challenge of variable network dimensions while exploiting the underlying power flow physics, we adopt the graph signal processing (GSP) framework. This formulation provides a natural pathway to design neural network architectures that are inherently transferable across topologies.

For each feeder configuration q, the distribution network is represented by a graph $G_q = (V_q, E_q)$ with an associated graph shift operator (GSO) $\mathbf{S}_q \in \mathbb{C}^{N_q \times N_q}$. In distribution systems, $\mathbf{S}_q$ is constructed from the admittance matrix $\mathbf{Y}_q$ so as to preserve electrical connectivity and impedance-weighted locality, as in the physics-aware formulation of [3]. A graph signal is defined as a vector $\mathbf{x}_t \in \mathbb{C}^{N_q}$ indexed by buses. In this work, $\mathbf{x}_t$ represents the complex voltage phasors $\mathbf{v}_t$ defined in (1), though we use the notation $\mathbf{x}_t$ to emphasize the generality of the graph signal processing framework and its applicability to other network-indexed quantities such as features derived from DSSE estimates or current phasors. Linear shift-invariant graph filters are expressed as polynomials of the GSO,

$$\mathbf{H}(\mathbf{S}_q)\mathbf{x}_t = \sum_{k=0}^{K} h_k \, \mathbf{S}_q^k \mathbf{x}_t, \tag{5}$$

where $\{h_k\}_{k=0}^K$ are scalar filter coefficients. The coefficients h_k are independent of the graph size N_q, which enables the same filter to be applied across feeders with different numbers of buses, as long as one utilizes the correct GSO. This requirement means that the deployment setting considered here assumes an updated topology and admittance-based GSO for the target configuration; robustness to incomplete or inaccurate admittance information is an important practical extension. To capture temporal dynamics, the graph filter in (5) is extended across time lags, yielding the spatio-temporal graph filter

$$\mathbf{z}_t = \mathbf{H}(\mathbf{S}_q)\mathbf{x}_t = \sum_{\tau=0}^{T-1} \sum_{k=0}^{K} h_{k,\tau} \, \mathbf{S}_q^k \mathbf{x}_{t-\tau}, \tag{6}$$

where $h_{k,\tau}$ are spatio-temporal scalar parameters. This formulation models the propagation of electrical interactions over both the network topology and time.

The spatial graph frequencies are the eigenvectors $\mathbf{U}_q$ of the matrix $\mathbf{S}_q$: the spatio-temporal Graph Fourier Transform (GFT) of a signal is $\tilde{\mathbf{Z}}(e^{-j\omega}) := \sum_{t=\infty}^{+\infty} \mathbf{U}_q^{-1} \mathbf{z}_t e^{-j\omega t}$. Let $H_k(e^{-j\omega}) = \sum_{t=\infty}^{+\infty} h_{k,\tau} e^{-j\omega t}$ and:

$$\mathrm{diag}(\tilde{\mathbf{H}}(e^{-j\omega})) = \sum_{k=0}^{K} H_k(e^{-j\omega}) \mathbf{\Lambda}_q^k. \tag{7}$$

It is easy to show that for the signal in (6) (i.e. the output of a graph filter) it holds:

$$\tilde{\mathbf{Z}}(e^{-j\omega}) = \mathrm{diag}(\tilde{\mathbf{H}}(e^{-j\omega}))\tilde{\mathbf{X}}(e^{-j\omega}) \tag{8}$$

Hence, *when graphs have the same eigenvalues for their GSO, in GFT domain graph filters with equal coefficients perform identically, even their topology is radically different* motivating us in Section III-A to study the asymptotic convergence of $\mathbf{Y}_q$.

In general, another reason why it is useful to interpret the states vector as a graph signal is that it provides intuition on the manifold that is defined by the AC powerflow equations and it directly allows to reason about an equivalent weight-sharing approach in the design of the perceptron structure. More specifically, the notion of graph signal allows to draw an equivalence with learning tasks that pertain machine vision task performed through images and Convolutional Neural Networks. Images are a special case of graph signals defined over a regular graph that represents the spatial adjacency matrix of the pixels. Videos are the counterpart of a spatio-temporal graph signal for vision tasks. For classification and inference problems that process labeled videos as inputs it is well known that convolutional layers outperform the basic perceptron structure; furthermore, transferability to different resolution images is standard as lower resolutions can be obtained by pooling pixels with a low-pass filter while higher resolutions (of course at a loss of details) can be obtained with a process of interpolation. Voltage phasors are a graph signal defined over the irregular support of the grid topology, and the question in designing what we name a Universal Graph Convolutional Neural Network (UGCN) lies in adapting not only the notion of graph convolution but the well known mechanism of pooling and interpolation to the case of grid data, as well as understanding how the neighborhood structure shapes the manifold of the data, ensuring that the training samples are rich enough to sample in the neighborhood of the state manifolds for unseen systems projected in a uniform dimension thanks to the pooling operation.

III. Transferable Physics-Aware Graph Learning

This section describes the proposed design of a neural network transferable across feeder configurations with varying topologies and dimensions. We first establish the theory by proving spectral concentration in distribution network admittance matrices, then introduce three architectural components: scalar-weight graph convolutions for topology-invariant features, adaptive pooling for dimension reduction, and parallel decoding for variable-size outputs in Fig. 1.

A. Structural Universality and Decaying Entropy

The theoretical foundation for the transferability of the proposed architecture lies in the statistical regularity of distribution topologies. We expect the embedding subspace to remain similar across systems whose reduced admittance matrices have comparable spectral statistics, feeder-depth profiles, and load-to-boundary-node ratios. Power systems exhibit a geometric degree distribution. Previous study on the topological and electrical characteristics of power networks have identified clear emerging patterns in parameters such as node degree, cable length, and load distribution [13]. Specifically, for distribution networks many of these properties are functionally linked to a node's distance, measured in hops, from the primary substation [13]. For systems with similar connectivity statistics and a consistent ratio of load-to-generation nodes, the empirical spectral distribution of the admittance-based GSO converges toward a deterministic density, analogous to the Marchenko-Pastur distribution in random matrix theory. Such spectral concentration suggests that in the Graph Fourier Transform (GFT) domain, the optimal filter parameters for feature extraction remain similar across different feeders, as the filters operate on a universal spectral density. While the GFT basis is tied to the specific topology, the underlying feature extraction logic is consistent.

Distribution systems simplified tree topologies (the motifs induced by the three-phase connections are also repetitive) are not only top heavy but exhibit a decaying mean degree, and thus a decaying entropy, for each successive generation from the root. As the branching process moves toward the leaves (customers), the local structural variance decreases. Consequently, the intermediate rows of $\mathbf{Y}_q$, which enforce the homogeneous constraints $[\mathbf{Y}_q\mathbf{v}_t]_i = 0$, represent a collection of local Kirchhoff operators sampled from a narrowing set of motifs. In Appendix (see Appendix B) we argue that for feeders encompassing thousands of nodes, this concentration of measure implies that the span of the intermediate rows acts as a uniform smoothing operator. If the branching statistics are identical, the effective resistance and the resulting voltage manifolds are equivalent up to a permutation of the node indices. This structural universality supports the hypothesis that a model trained on a subset of configurations can generalize to unseen topologies, as the underlying physical constraints are associated to system matrices whose spectrum is asymptotically converging to the same empirical distributions for the eigenvalues. Hence, simply adapting the eigenvectors (i.e. the graph frequencies) to the target topology is sufficient, as the Graph Frequency Transform (GFT) response that is optimum for the task remains the same.

B. Scalar-Weight Graph Convolutions

We now specialize the GSP formulation to the scalar-weight graph convolutional layers used in the proposed architecture [10]. Let $\mathbf{X}_t^{(\ell)} \in \mathbb{C}^{N_q \times F_\ell}$ denote the matrix of node features at layer ℓ and time t, where each row corresponds to a bus and F_ℓ is the number of feature channels. The input features are constructed from DSSE-based voltage phasor estimates and their temporal history. A scalar-weight spatio-temporal graph convolutional layer computes the output features as follows

$$\mathbf{X}_t^{(\ell+1)} = \sigma\left(\sum_{\tau=0}^{T-1}\sum_{k=0}^{K} h_{k,\tau}^{(\ell)}\,\mathbf{S}_q^k\mathbf{X}_{t-\tau}^{(\ell)}\right), \tag{9}$$

where $h_{k,\tau}^{(\ell)} \in \mathbb{C}$ are learnable scalar parameters, $\mathbf{S}_q$ is the physics-aware graph shift operator derived from the feeder admittance matrix, and $\sigma(\cdot)$ is a pointwise nonlinearity. Equation (9) is a direct neural realization of the spatio-temporal graph filter in (6).

Importantly, the parameters $h_{k,\tau}^{(\ell)}$ are independent of the feeder size N_q and of the specific topology of configuration q. As a result, the same convolutional filters can be applied to feeders with different numbers of buses and to previously unseen reconfigurations. This scale invariance is the key mechanism enabling zero-shot transfer across feeder topologies.

C. Adaptive Graph Pooling

While scalar-weight GCNs handle variable input dimensions, downstream layers require fixed-size representations. To bridge this gap, we introduce adaptive graph pooling that maps variable-size node embeddings to a fixed-dimensional latent space. Pooling is performed using application-oriented clustering of buses, followed by average and max aggregation. This operation preserves salient electrical features while enabling a unified hidden representation across feeders. To this end, let $\mathbf{X}_q^{(L)} \in \mathbb{C}^{N_q \times F_L}$ denote the output of the final graph convolutional layer for configuration q. Buses are partitioned into N_p clusters $\{\mathcal{C}_i\}_{i=1}^{N_p}$ based on electrical distance and feeder depth. The pooling operation maps node-level features to cluster-level features through

$$\mathbf{Z}_q = \mathbf{A}_q \mathbf{X}_q^{(L)}, \qquad \mathbf{A}_q \in \mathbb{R}^{N_p \times N_q}, \tag{10}$$

where the pooling matrix $\mathbf{A}_q$ is configuration-dependent but structured. Specifically, denoting by $\|$ feature concatenation, for each cluster $\mathcal{C}_i$, average and max pooling are applied:

$$[\mathbf{Z}_q]_{i,:} = \Big[\frac{1}{|\mathcal{C}_i|}\sum_{n\in\mathcal{C}_i} \mathbf{X}_{q,n,:}^{(L)} \;\Big\|\; \max_{n\in\mathcal{C}_i} \mathbf{X}_{q,n,:}^{(L)}\Big]. \tag{11}$$

This construction preserves dominant voltage patterns within electrically coherent regions while producing a fixed-size latent representation $\mathbf{Z}_q \in \mathbb{C}^{N_p \times 2F_L}$.

The important observation here is that, although $\mathbf{A}_q$ depends on the feeder configuration, the pooling dimension N_p is fixed across all feeders. This allows the downstream layers to operate on a configuration-invariant representation while retaining sensitivity to local electrical structure. The pooling operation is physically justified by the convergence of the Schur complement discussed in the previous section and studied in Appendix B. Since the branching process exhibits decaying entropy, the effective electrical distance between nodes concentrates around a deterministic profile. Consequently, it makes sense to partition systems into N_p clusters $\{\mathcal{C}_i\}_{i=1}^{N_p}$ based on this electrical distance and feeder depth. Because the RDE solution for the tree's impedance converges to a stable distribution, the cluster centroids N_p represent universal *electrical zones* that remain invariant across different feeder realizations. This allows the downstream layers to operate on a configuration-invariant representation while retaining sensitivity to the underlying manifold of the voltage phasors.

D. Training Across Feeder Reconfigurations

Let $\mathcal{Q}$ denote the set of training configurations and $\mathbf{S}_q$ the corresponding admittance matrix/GSO, $\mathbf{y}_q$ be the ground truth labels for system q, $\ell(\cdot)$ is the loss function for individual systems, and $\mathbf{H} = \{\mathbf{H}_{k,\tau}^{(\ell)}\}_{\ell=1}^{L}$ the collection of all learnable parameters across layers. The UGCN training objective is minimizing the estimation loss across all Q power systems:

$$\min_{\mathbf{H}} \mathcal{L}(\mathbf{H}) := \frac{1}{Q}\sum_{q=1}^{Q} \ell(\mathbf{H}; \mathbf{S}_q, \mathbf{X}_q^{(0)}, \mathbf{y}_q). \tag{12}$$

By jointly training across configurations, the topology variation is treated as an intrinsic learning dimension rather than a deployment-time disturbance. In Section III-E we further discuss how to generate via simulations a suitable sample set that leverages the concentration phenomena we highlighted while sampling appropriately the manifolds so that it is possible to connect them across systems.

This said, there is another problem to solve: in the state estimation setting considered in this paper ($H = 0$), the loss $\mathcal{L}(\cdot)$ is defined directly on the complex voltage phasors, i.e. $\mathbf{y}_q = (\mathbf{v}_{q,1}, \ldots, \mathbf{v}_{q,T})$ and as

$$\mathcal{L} = \sum_{t=1}^{T} \left\| \mathbf{v}_{q,t} - \Phi(\mathbf{H}; \mathbf{S}_q, \mathbf{x}_{q,t}^{(0)}) \right\|_2^2, \tag{13}$$

where $\mathbf{x}_{q,t}^{(0)}$ denotes the solution of (4) $\mathbf{v}_q(t)$ is the ground-truth system state. This implies that for systems with different number of buses, the output of the neural network will need to change size. The next subsections focus on these challenges.

E. Training Set Generation via Structural Continua

A critical challenge in training a transferable GCN is ensuring that the feature representations remain aligned across disparate topological manifolds. To accomplish this, we construct the training set $\mathcal{D} = \{(\mathbf{Y}_q, \mathbf{v}_{q,t}, \mathbf{s}_{q,t})\}$ not as a collection of independent snapshots, but as a sequence of gradual structural and loading perturbations. The objective is to densely sample the space of voltage manifolds $\mathcal{M}_q$ by smoothly interpolating between different feeder configurations. By constructing the sample set through these gradual transitions, the scalar-weight filters learn to associate the local Kirchhoff residuals with the global voltage profile, regardless of the permutation of the node indices in any single realization. Specifically, the training samples are the result of a stochastic process $\mathcal{P}(\mathbf{Y}, \mathbf{s})$ that evolves in two dimensions:

a) Admittance Perturbations:: Rather than switching lines discretely, we generate intermediate "virtual" configurations by varying the admittance of tie-lines and sectionalizing switches $\mathbf{Y}_q(\alpha) = (1-\alpha)\mathbf{Y}_A + \alpha\mathbf{Y}_B$ for $\alpha \in [0,1]$. This ensures the null space of the internal Laplacian evolves continuously, allowing the GCN to track the transformation of the voltage manifold.

b) Loading Redistribution:: We redistribute the aggregate feeder load $\mathbf{S}_{total}$ along the tree using a decaying entropy profile consistent with the IGW process discussed in Appendix B. By varying the *branchiness* of the load (e.g., shifting load from the trunk to the fringes), we expose the model to the full range of the Schur complement's spectral concentration.

F. Adaptive Parallel Decoder Output

While the preceding sections introduced adaptivity through pooling in graph convolutional networks, we now extend this

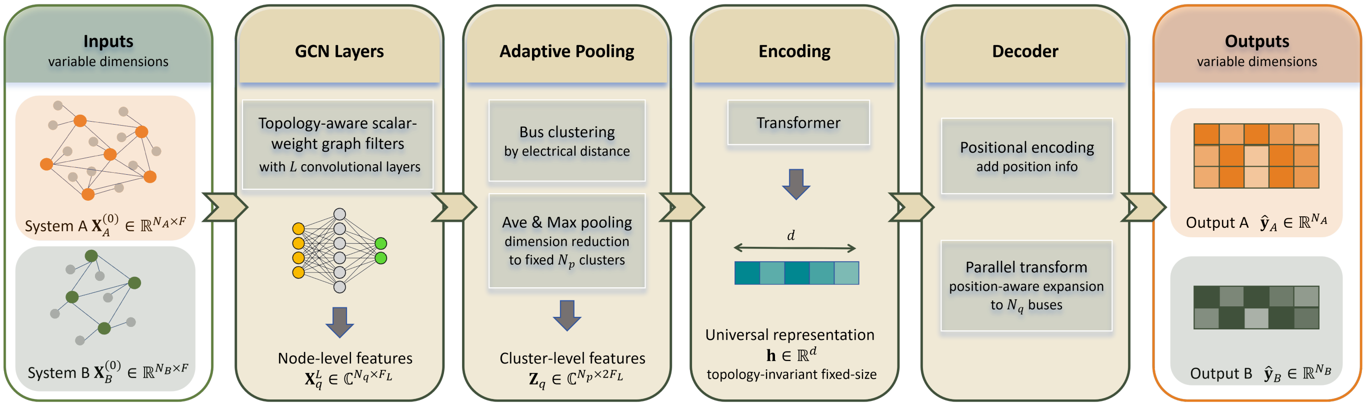

Fig. 1: Universal Graph Convolutional Neural Network (UGCN) architecture for zero-shot transfer across variable-sized distribution networks. The framework processes inputs through: (1) GCN layers with scalar-weight filters on graph-structured data, (2) adaptive pooling that clusters buses by electrical distance, (3) encoding to a fixed-size universal representation, (4) parallel decoding with positional encoding, and (5) variable-sized outputs matching the input dimensions.

principle to the output layer. The proposed adaptive parallel decoder produces outputs whose dimensionality varies with the network size via fully parallel computation, offering increased flexibility and improved computational efficiency relative to conventional autoregressive decoding schemes and allowing to estimate state vectors of different sizes seamlessly.

For a power network comprising N_q buses, the decoder takes as input the pooled feature matrix $\mathbf{X}_{pool} \in \mathbb{R}^{N_p \times F_L}$ generated by the adaptive pooling layer and maps it to a common latent representation:

$$\mathbf{h} = \sigma(\mathbf{W}_{enc}\text{vec}(\mathbf{X}_{pool}) + \mathbf{b}_{enc}) \in \mathbb{R}^d, \tag{14}$$

where $\text{vec}(\mathbf{X}_{pool}) \in \mathbb{R}^{N_p \cdot F_L}$ denotes vectorization of the pooled features, d is the latent dimension, and $\{\mathbf{W}_{enc}, \mathbf{b}_{enc}\}$ are learnable encoding parameters.

A central element of the architecture is a parallel positional encoding scheme that constructs embeddings for all N_q output locations simultaneously. This operation realizes the inverse of the structural mapping described in Sections III-B and III-C, whereby the shared voltage manifold representation is projected back onto bus-specific phasors:

$$\mathbf{p}_{N_q} = \left[0, \frac{1}{N_q - 1}, \dots, 1\right]^T \in \mathbb{R}^{N_q}, \tag{15}$$

$$\mathbf{E}_{pos} = \tanh(\mathbf{p}_{N_q}\mathbf{W}_{pos}^T + \mathbf{b}_{pos}) \in \mathbb{R}^{N_q \times d}, \tag{16}$$

where $\mathbf{W}_{pos} \in \mathbb{R}^{d \times 1}$ and $\mathbf{b}_{pos} \in \mathbb{R}^d$ parameterize a continuous spatial embedding. The latent representation is replicated across all output positions and combined with the positional embeddings to form the final voltage estimates:

$$\mathbf{C} = \mathbf{1}_{N_q}\mathbf{h}^T + \mathbf{E}_{pos} \in \mathbb{R}^{N_q \times d}, \tag{17}$$

$$\mathbf{T} = \sigma(\mathbf{C}\mathbf{W}_T + \mathbf{1}_{N_q}\mathbf{b}_T^T) \in \mathbb{R}^{N_q \times d}, \tag{18}$$

$$\mathbf{y} = \mathbf{T}\mathbf{w}_{out} + b_{out} \in \mathbb{R}^{N_q}. \tag{19}$$

Here, $\mathbf{W}_T \in \mathbb{R}^{d \times d}$ and $\mathbf{w}_{out} \in \mathbb{R}^d$ are shared parameters that capture topology-invariant power flow structure.

This parallel decoding strategy avoids the parameter overwriting effects associated with sequential output generation by retaining simultaneous access to configuration-dependent output spaces through distinct positional encodings $\mathbf{p}_{N_q}$. For a given feeder with N_q buses, the corresponding position vector deterministically selects the appropriate N_q-dimensional output space without affecting representations associated with other network sizes. Because the RDE-based pooling mechanism in Section III-C enforces a topology-consistent latent representation $\mathbf{h}$, the shared decoder parameters effectively learn a universal mapping from the voltage manifold to node-level estimates. In other words, adaptive pooling standardizes the latent representation, while the parallel decoder re-expands it in a topology-aware manner; the two operations address distinct sources of dimensional mismatch.

Unlike fine-tuning–based transfer methods, which still require feeder-specific labeled data, the proposed approach requires no post-deployment adaptation for the reconfiguration setting considered in this study.

IV. Case Studies

A. Simulation Settings

Experiments are conducted on the IEEE 123-node three-phase unbalanced radial distribution test feeder, simulated using OpenDSS. Time series active and reactive power demand profiles are drawn from the Austin (AUS) region of the SMART-DS dataset hosted on the Open Energy Data Initiative (OEDI) [14]. The network structure is subject to significant variability through common reconfiguration events in radial distribution systems, including feeder disconnections, new feeder installations, link failures, subtree mergers, and line parameter variations. Multiple reconfiguration events may occur randomly and concurrently, resulting in substantial topological changes that alter both network size and admittance matrix structure. For the power system state forecasting (PSSF) task, we generate 1000 distinct distribution system reconfigurations for training and 200 previously unseen reconfigurations for testing. These counts were selected to give broad empirical coverage of the sampled reconfiguration family and stable validation trends in the present IEEE 123-bus study, rather than as a universal sample-complexity guarantee. The UGCN architecture comprises two scalar-weight graph convolutional layers for feature extraction, followed by two fully connected layers with 256 neurons each. The input to UGCN consists of Advanced Metering Infrastructure (AMI) measurements including voltage magnitudes and apparent power injections available at a subset of buses. We assume that AMI devices are installed at all feeder endpoints, providing approximately 30% observability for estimating global system states. The testing dataset includes both new topological configurations and future time series data unseen during training, evaluating UGCN's zero-shot transferability within the sampled IEEE 123-bus reconfiguration family.

Our baseline algorithms include **FNN** (four-layer fully-connected network with 512 neurons per layer), **CplxFNN** (complex-valued FNN with identical architecture), **GRU** (GNN layer [8] with three fully-connected layers), **Transformer** [15] (three encoder/decoder sub-layers with 512 features), **GTN** [16] (graph transformer with 10 channels and three fully-connected layers), **RNN** [17], **CNN** [18], and **LSTM** [19]. All methods use voltage phasors as primary inputs, with RNN additionally incorporating power injections and flows.

B. Manifold Visualization

To visualize cross-topology alignment, we compute a low-dimensional embedding by applying Isomap dimensionality reduction to the feature representations after adaptive pooling. These visualizations represent simplified projections from high-dimensional complex-valued feature spaces to three-dimensional representations, providing a topology-agnostic view of the learned voltage manifold and demonstrating whether pooled features form continuous trajectories across distinct feeder reconfigurations.

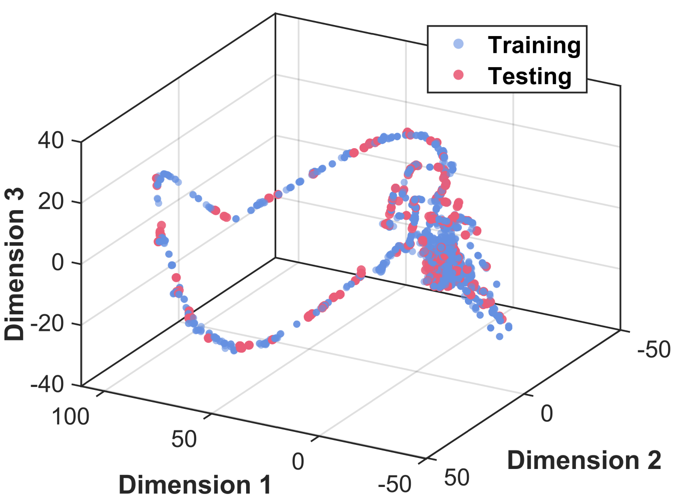

Fig. 2: Isomap embedding of complex voltage phasors across multiple IEEE 123-bus feeder reconfigurations. Training samples (blue) and testing samples (red) form continuous clusters in the embedded space, demonstrating that the learned representations capture topology-invariant voltage manifold structure despite significant variations in network configuration.

C. Power Distribution System State Forecasting

1) Experimental Setup: All experiments use a 10-hour historical window ($T = 10$) to forecast voltage phasors for the next 1–5 hours ($H = 1, 2, 3, 4, 5$). When $H = 0$, the problem reduces to power system state estimation. We use AMI measurements on 30% of buses in (4). Estimated voltage phasors serve as forecasting inputs for both scenarios.

2) Results on IEEE 123-Bus System: We evaluate performance under various reconfigurations comparing UGCN against baseline models (FNN, CplxFNN, GRU, Transformer, GTN, RNN, LSTM) and their padded extensions.

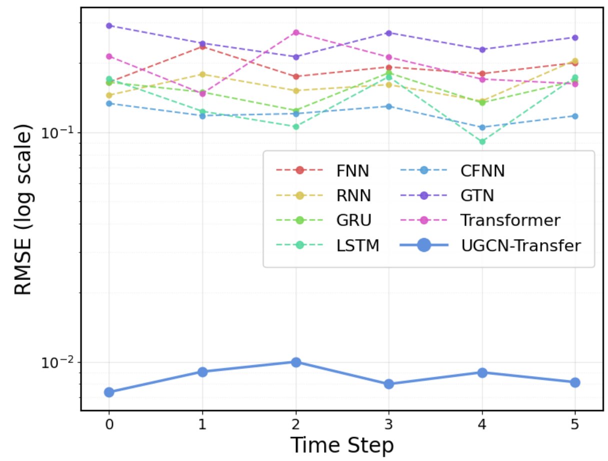

Fig. 3: UGCN transferability for power system state estimation and forecasting across unseen IEEE 123-bus reconfigurations.

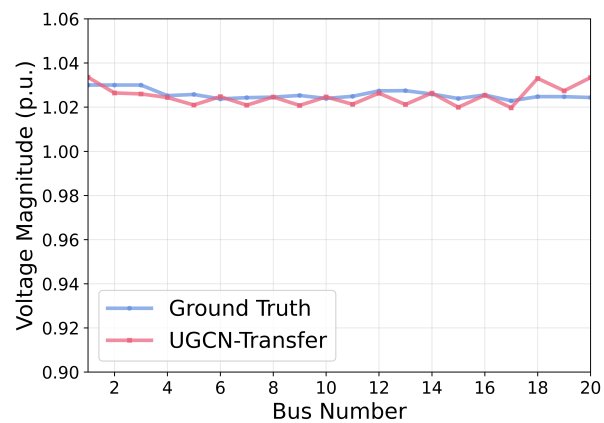

Fig. 4: Zero-shot transfer example: UGCN voltage magnitude estimation ($H = 1$) performance on unseen IEEE 123-bus system reconfiguration. Results are shown for the first 20 buses.

As shown in Figure 3, UGCN-Transfer achieves the lowest RMSE of 7.38×10^{-3} p.u. at $H = 0$, significantly outperforming all benchmark methods. At this horizon, baseline RMSEs range from 1.33×10^{-1} p.u. (CplxFNN) to 2.91×10^{-1} p.u. (GTN), corresponding to an improvement of approximately $18.1\times$ over the best competing method.

Figure 4 provides detailed visualization of voltage magnitude estimation, showing close agreement between UGCN predictions and ground-truth values for state estimation ($H = 0$) on IEEE 123-bus system reconfigurations, with a mean absolute error of 4.1×10^{-3} p.u. These results demonstrate that UGCN maintains superior state estimation and forecasting accuracy across different configurations in the tested feeder family, supporting strong generalization capabilities under reconfiguration. The consistent order-of-magnitude improvements highlight UGCN's ability to capture fundamental power system physics that remain invariant across the sampled network topologies.

V. Conclusion

This paper presented a transferable, physics-aware graph learning framework for distribution network state estimation and forecasting. Unlike conventional black-box models, our approach leverages the structural universality inherent in distribution topologies. We have shown, through the lens of Recursive Distributional Equations (RDEs), that the decaying entropy and degree distribution of these networks cause the internal voltage manifold to concentrate around a statistically deterministic subspace. By combining scalar-weight GCNs with adaptive graph pooling, the proposed model effectively learns the "universal interpolation" rules of the feeder. This enables zero-shot transfer across major reconfigurations without the need for retraining, and suggests a path toward transfer across feeders with similar reduced admittance statistics. Our results demonstrate that by aligning the learning architecture with the underlying spectral properties of the admittance matrix, we can achieve a level of scalability and robustness useful for reconfigurable distribution grids. The present experiments are limited to synthetic reconfigurations of the IEEE 123-bus feeder with an available admittance-based GSO. Future work will evaluate larger and electrically diverse feeders, quantify sensitivity to imperfect topology and admittance information, and isolate the contribution of individual architectural components through ablation studies.

Appendix

A. Acknowledgements

AI assistance was used in this work as follows: AI tools were used for spell-checking and grammar correction to improve the clarity and readability of the manuscript.

B. Convergence of the Schur Complement in Decaying Trees

Consider a sequence of distribution feeders q modeled as Inhomogeneous Galton-Watson (IGW) trees (see e.g. [20], [21]). Let $\mathcal{V}_{int}$ denote the set of intermediate nodes and $\mathcal{V}_{ext}$ the boundary nodes (root and leaves). The Schur complement of $\mathbf{Y}_q$ is $\mathbf{Y}_{red} = \mathbf{Y}_{ee} - \mathbf{Y}_{ei}\mathbf{Y}_{ii}^{-1}\mathbf{Y}_{ie}$, where $\mathbf{Y}_q$ is:

$$\mathbf{Y}_q = \begin{bmatrix} \mathbf{Y}_{ee} & \mathbf{Y}_{ei} \\ \mathbf{Y}_{ie} & \mathbf{Y}_{ii} \end{bmatrix} \tag{20}$$

Theorem 1 (Spectral Concentration). *Given a decaying mean degree μ_d and decaying entropy $H(d) \to 0$ as depth $d \to \infty$, the Schur complement $\mathbf{Y}_{red}$ converges in the spectral sense to a deterministic operator $\bar{\mathbf{Y}}$ as $N_q \to \infty$.*

Proof Sketch.

1) *Recursive Distributional Equations (RDE) Formulation:* For trees, the marginals of the Gaussian Markov Random Field associated with $\mathbf{Y}$ (which determine $\mathbf{Y}_{ii}^{-1}$) can be solved via belief propagation. The effective impedances $[\mathbf{Y}_{ii}^{-1}]_{p,c}$ seen from a node k toward its descendants n satisfy the RDE:
$$Z_{parent} = f(\{Z_{children}, \mathbf{y}_{edges}\}) \tag{21}$$
where y are line weights. Under the condition $\mu_d \leq 1$ for large d, this RDE has a unique stable fixed point in the space of distributions [22].
2) *Entropy Collapse:* The decay in degree entropy $H(d) \to 0$ implies that the local topology at depth d becomes increasingly deterministic. This "structural crystallization" ensures that the sum of contributions from the thousands of internal branches to the boundary nodes concentrates around its expected value.
3) *Permutation Equivalence:* Because the rows of $\mathbf{Y}_{ii}$ are samples from a narrowing set of motifs, any two feeders with the same statistics will yield row spans that are equivalent up to a permutation $\mathbf{P}$. Thus, the resulting embedding subspace for the voltage manifold is topology-invariant for sufficiently large N_q.

References

[1] J. Zhu, *Optimization of power system operation*. John Wiley & Sons, 2015.

[2] M. Khodayar, G. Liu, J. Wang, and M. E. Khodayar, "Deep learning in power systems research: A review," *CSEE J. Power Energy Syst.*, vol. 7, no. 2, pp. 209–220, 2020.

[3] T. Wu, I. L. Carreño, A. Scaglione, and D. Arnold, "Spatio-temporal graph convolutional neural networks for physics-aware grid learning algorithms," *IEEE Trans. Smart Grid*, 2023.

[4] Q. Li *et al.*, "A transfer learning framework for power system event identification," *IEEE Transactions on Power Systems*, 2022.

[5] M. Hijazi, P. Dehghanian, and S. Wang, "Transfer learning for transient stability predictions in modern power systems under enduring topological changes," *IEEE Trans. Autom. Sci. Eng.*, 2023.

[6] S. Niu, Y. Liu, J. Wang, and H. Song, "A decade survey of transfer learning (2010–2020)," *IEEE Trans. Artif. Intell.*, vol. 1, no. 2, pp. 151–166, 2021.

[7] C. Ren and Y. Xu, "Transfer learning-based power system online dynamic security assessment: Using one model to assess many unlearned faults," *IEEE Transactions on Power Systems*, vol. 35, no. 1, pp. 821–824, 2019.

[8] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[9] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio, *et al.*, "Graph attention networks," *stat*, vol. 1050, no. 20, pp. 10–48550, 2017.

[10] T. Wu, A. Scaglione, and D. Arnold, "Complex-value spatio-temporal graph convolutional neural networks and its applications to electric power systems ai," *IEEE Trans. Smart Grid*, 2023.

[11] Y. Li *et al.*, "Pmu measurements-based short-term voltage stability assessment via deep transfer learning," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[12] I. L. Carreño, A. Scaglione, S. S. Saha, D. Arnold, S.-T. Ngo, and C. Roberts, "Log (v) 3lpf: A linear power flow formulation for unbalanced three-phase distribution systems," *IEEE Transactions on Power Systems*, vol. 38, no. 1, pp. 100–113, 2022.

[13] E. Schweitzer, A. Scaglione, A. Monti, and G. A. Pagani, "Automated generation algorithm for synthetic medium voltage radial distribution systems," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 7, no. 2, pp. 271–284, 2017.

[14] B. Palmintier, C. Mateo Domingo, F. E. Postigo Marcos, T. Gomez San Roman, F. de Cuadra, N. Gensollen, T. Elgindy, and P. Duenas, "Smart-ds synthetic electrical network data opendss models for sfo, gso, and aus." Open Energy Data Initiative (OEDI), National Renewable Energy Laboratory (NREL), 2020. Accessed: 2026-01-29.

[15] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[16] L. Rampášek, M. Galkin, V. P. Dwivedi, A. T. Luu, G. Wolf, and D. Beaini, "Recipe for a General, Powerful, Scalable Graph Transformer," *Adv. Neural Inf. Process. Syst.*, vol. 35, 2022.

[17] L. Zhang, G. Wang, and G. B. Giannakis, “Power system state forecasting via deep recurrent neural networks,” in *IEEE ICASSP*, 2019.

[18] S. Wang, S. Bi, and Y.-J. A. Zhang, “Locational detection of the false data injection attack in a smart grid: A multilabel classification approach,” *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8218–8227, 2020.

[19] Y. Wang, Z. Zhang, J. Ma, and Q. Jin, “Kfrnn: an effective false data injection attack detection in smart grid based on kalman filter and recurrent neural network,” *IEEE Internet Things J.*, vol. 9, 2021.

[20] R. Lyons, “Random walks, capacity and percolation on trees,” *The Annals of Probability*, pp. 2043–2088, 1992.

[21] R. Lyons and Y. Peres, “Probability on trees and networks,” vol. 42, 2017.

[22] D. J. Aldous and J. M. Steele, “The objective method: Probabilistic combinatorial optimization and local weak convergence,” *Probability on discrete structures*, vol. 110, pp. 1–72, 2005.

I. REVIEW #52A

A. Paper summary

The paper develops a GNN-based ML model for SE of three-phase unbalanced power distribution systems that is applicable over a wide range of topologies. The key idea is to support and apply transfer learning to achieve that goal.

B. Comments for authors

1) Authors support and prove the hypothesis that few learned parameters h_k apply to SE over a wide range of topologies.
2) Authors clearly motivate that transfer learning is a practical approach for distribution grids (in this case) as re-training is not always viable due to limited data.
3) Authors state that: "transfer learning is limited to re-configurations of the trained feeder and "similar" feeder." Similar is not clearly defined yet.
4) In the results section, the network size is not representative of scale necessary for real-world analysis
5) "For the power system state forecasting (PSSF) task, we generate 1000 distinct distribution system reconfigurations for training and 200 previously unseen reconfigurations for testing" - How do you know this is sufficient training data?
6) Authors state: "To enable closed form solution, we linearize the measurement function $h(\cdot)$ using the log3v transformation [12]." Given there are errors associated with linearizing the power flow model, how does this affect the accuracy of overall algorithm?
7) The manuscript observes that because of the tree structure of power networks, the reduced form of different distribution network has similar structural properties, driving the underlying use of transfer learning. The observation is backed by a theorem and proof sketch in the appendix.

C. Summarize your recommendation in one to two sentences

I enjoyed reading the paper. I really like the observation that demonstrates that different reconfigurations of distribution grid feeder has some structural similarities, when Kron reduced.

— Anonymous reviewer

II. REVIEW #52B

A. Paper summary

The authors propose a "Universal Graph Convolutional Neural Network" (UGCN). Zero-shot transfer means the model can be trained on a set of grid configurations and then flawlessly operate on a brand new, unseen grid configuration without needing any retraining.

B. Comments for authors

Strengths:

- Highly Relevant and Novel Approach: The theoretical grounding of combining physics-aware graph learning with zero-shot transferability is highly interesting and tackles a major real-world bottleneck for utility companies.
- Smart Architectural Pipeline: The overarching idea of compressing variable-sized graphs into a fixed latent space and expanding them out using positional encoding is structurally clever.

Weaknesses:

- The title and introduction heavily boast about "zero-shot transfer across different feeders." However, the experiments only test reconfigurations of a single, specific grid (the IEEE 123-bus system). I feel testing different switch states on one feeder is not the same as transferring the model to a completely different feeder. The core claim of the paper is left entirely unvalidated by the experiments.
- The authors introduce several complex custom components (Scalar-weight GCNs, specific Graph Pooling, specific dense layers) but provide no tests/ reasoning isolating them. Without ablation studies (e.g., swapping out their GCN for a standard one, or testing different pooling strategies), it is impossible to know if their complex math is actually responsible for the good results, or if it's just an arbitrary architectural choice.
- The authors claim to use a "Transformer-based" architecture for their decoder. However, looking at Equations 14–19, the math contains absolutely none of the components that define a Transformer (no multi-head attention, no Query/Key/Value mechanisms, no softmax, no layer normalization).
- One big limitation is that the zero-shot capability relies entirely on the model being fed the exact Graph Shift Operator which requires a perfect admittance matrix. If a utility company reconfigures a grid, they often don't have an perfectly accurate, updated electrical map right away. If the exact admittance matrix is missing, the model simply cannot run, limiting real-world usefulness. (knowing topology is different than knowing full Admittance matrix)

C. Summarize your recommendation in one to two sentences

Although the theoretical grounding is interesting, there is a critical disconnect between the paper's overarching claims and its experimental validation. Furthermore, significant architectural claims require accurate re-labeling and rigorous justification.

— *Parikshit Pareek*

III. Review #52C

A. Paper summary

The paper proposes a Universal Graph Convolutional Network (UGCN) for distribution system state estimation (DSSE) and short-term forecasting on three-phase unbalanced feeders under network reconfiguration. The main claim is zero-shot generalization across feeders with different topologies and sizes without retraining. The architecture combines scalar-weight graph filters independent of graph size, an adaptive pooling mechanism that produces a fixed-size latent representation, and a parallel decoder that reconstructs variable-length voltage phasor outputs.

Experiments are conducted on OpenDSS simulations of the IEEE-123 feeder with synthetic AMI measurements and multiple reconfiguration scenarios. The proposed model is compared with several neural baselines and reports lower RMSE for both estimation and forecasting tasks.

B. Comments for authors

Strengths:

- The paper proposes a structured architecture combining scalar-weight GCN filters, electrically informed graph pooling, and a parallel decoder capable of producing variable-length outputs across different feeder topologies. The design is supported by a physics-aware graph representation, which improves transparency compared to other black-box GNN architectures.
- The paper targets two relevant operational challenges, namely state estimation and forecasting under frequent feeder reconfigurations. To enable cross-topology zero-shot transfer, the authors design graph filters whose parameters are independent of network size.

Weaknesses:

- The experimental validation is limited to a single feeder with synthetic reconfiguration scenarios. No experiments are reported on larger feeders (e.g., 8500-node systems) or on multiple feeders with different electrical characteristics.
- The dataset generation procedure relies on random concurrent events and synthetic topology changes, but the distributions, random seeds, and exact scenario counts are not specified. In addition, the paper does not report model sizes, training times, or variance across runs.
- Beyond the illustrative plots in Figs. 3–4, the paper does not report tabulated performance metrics. In particular, there are no statistics such as mean, median, or percentile errors for voltage magnitudes and angles (in p.u.), nor feasibility metrics across reconfigurations. As a result, it is difficult to assess statistical variability, robustness to topology changes, or whether the predicted states are physically consistent.

C. Summarize your recommendation in one to two sentences

The paper proposes a zero-shot transfer learning architecture for DSSE and forecasting under feeder reconfigurations. However, the experimental validation and numerical reporting are limited, making it difficult to assess the robustness and generality of the claimed zero-shot transfer capability.

— *Anonymous reviewer*

IV. Brief response to reviewers (optional)

We thank the reviewers for their careful reading and constructive feedback. We appreciate the positive comments on the motivation, theoretical observation, and architecture of the proposed UGCN framework. In the final version, we clarified the intended scope of the zero-shot transfer claim, emphasizing that the experimental validation focuses on unseen reconfigurations of the IEEE 123-bus feeder, while broader transfer across substantially different feeders remains an important direction for future work. In response to the reviewers' comments, we refined the wording in the title, abstract, introduction, results, and conclusion to avoid overstating cross-feeder generalization. We also clarified what is meant by "similar" feeders in terms of comparable reduced-admittance spectral statistics, feeder-depth profiles, and load-to-boundary-node ratios. In addition, we revised the decoder description to avoid inaccurately characterizing it as a standard Transformer architecture, and instead describe it as a parallel decoder with positional encoding. We further added discussion of practical assumptions and limitations, including the need for an updated topology/admittance-based graph shift operator, the role of the log3v linearization as a coarse physics-informed input, and the fact that the selected number of training/testing reconfigurations provides empirical coverage for the present IEEE 123-bus study rather than a universal sample-complexity guarantee. We also noted future work on larger and more diverse feeders, imperfect admittance information, and ablation studies.

V. Brief summary of changes made to the final paper

We made targeted revisions to improve clarity and align the paper's claims with the presented experimental evidence. First, we revised the title, abstract, introduction, results discussion, and conclusion to describe the demonstrated zero-shot transfer setting more precisely as transfer across unseen feeder reconfigurations. We retained discussion of transfer to similar feeders as a possible extension, but clarified that the current experiments focus on the IEEE 123-bus reconfiguration family. Second, we clarified the meaning of "similar" feeders by relating it to reduced-admittance spectral statistics, feeder-depth profiles, and load-to-boundary-node ratios. We also added a note that the proposed method assumes an updated topology and admittance-based graph shift operator for the target configuration, and that robustness to incomplete or inaccurate admittance information is an important future direction. Third, we revised the architectural description of the output module. The previous wording referred to a Transformer-based decoder; the final version describes this component more accurately as an adaptive parallel decoder with positional encoding. We also added clarification that the log3v transformation introduces an approximation relative to nonlinear power flow, but is used only to generate a coarse physics-informed input before end-to-end learned correction. Finally, we added brief limitation-oriented discussion in the experimental and conclusion sections. These additions note that the 1000 training and 200 testing reconfigurations provide empirical coverage for the present study rather than a general sample-complexity guarantee, and identify future work on larger feeders, diverse electrical characteristics, imperfect network models, and ablation studies.