JOURNAL OF
Optical Communications
and Networking

Generalized few-shot transfer learning architecture for modeling the EDFA gain spectrum

Agastya Raj,[1,*] Zehao Wang,[2] Tingjun Chen,[2] Daniel C. Kilper,[3] and Marco Ruffini[1]

[1]CONNECT Centre, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland
[2]Duke University, Department of Electrical and Computer Engineering, Durham, North Carolina 27708, USA
[3]CONNECT Centre, School of Engineering, Trinity College Dublin, Dublin, Ireland
*rajag@tcd.ie

Received 3 March 2025; revised 13 May 2025; accepted 4 June 2025; published 25 July 2025

Accurate modeling of the gain spectrum in erbium-doped fiber amplifiers (EDFAs) is essential for optimizing optical network performance, particularly as networks evolve toward multi-vendor solutions. In this work, we propose a generalized few-shot transfer learning architecture based on a semi-supervised self-normalizing neural network (SS-NN) that leverages internal EDFA features—such as VOA input/output power and attenuation—to improve gain spectrum prediction. Our SS-NN model employs a two-phase training strategy comprising unsupervised pre-training with noise-augmented measurements and supervised fine-tuning with a custom-weighted MSE loss. Furthermore, we extend the framework with transfer learning (TL) techniques that enable both homogeneous (same-feature space) and heterogeneous (different-feature sets) model adaptation across booster, pre-amplifier, and ILA EDFAs. To address feature mismatches in heterogeneous TL, we incorporate a covariance matching loss to align second-order feature statistics between the source and target domains. Extensive experiments conducted across 26 EDFAs in the COSMOS and Open Ireland testbeds demonstrate that the proposed approach significantly reduces the number of measurement requirements on the system while achieving lower mean absolute errors and improved error distributions compared to benchmark methods.

https://doi.org/10.1364/JOCN.560987

1. INTRODUCTION

The relentless growth of global data traffic, driven by emerging technologies such as 5G/6G, smart cities, autonomous systems, and the Internet of Things (IoT), has placed increasing demands on optical networks to sustain higher bandwidth, lower latency, and enhanced reliability [1]. These requirements are further increased by the industry-wide shift toward access-metro-core convergence, an architectural framework that integrates disaggregated network layers, optimizes spectral and energy efficiency through software-defined control, and enables seamless scalability across the edge, metro, and long-haul domains [2]. A critical component that enables these advancements is the erbium-doped fiber amplifier (EDFA), which compensates for signal loss over long distances. This has an impactful role in optimizing optical network performance, as the output power of the EDFA determines the launch power level of each optical channel, which further affects the magnitude of nonlinear impairments in fiber. The noise figure of the EDFA, which quantifies the degradation of the signal-to-noise ratio (SNR) due to amplification, has a direct impact on key performance metrics such as the optical SNR (OSNR) and quality of service (QoT) [3,4]. The performance of EDFAs is inherently tied to the wavelength-dependent gain profile, which exhibits complex nonlinear behavior under varying operating conditions, including pump power, channel loading, and operating mode [5,6]. Recently, increasing demands for high-bandwidth, low-latency applications have led to the development of optical spectrum as a service (OSaaS), which allows multiple parties to utilize the same network infrastructure, thereby increasing spectral utilization. However, in such a dynamic scenario with multiple stakeholders, predicting the system behavior is necessary to optimize performance and minimize security risks [7,8]. Therefore, accurate modeling of the EDFA gain spectrum is critical for optimizing physical-layer control in dynamic networks, particularly as operators reduce design margins to improve spectral efficiency.

Traditional EDFA gain models rely on physics-based frameworks derived from the quantum-mechanical principles of

erbium-doped fiber (EDF) dynamics or on simplified expressions for the automatic gain control (AGC) mode, which require only a limited amount of measurements [9]. However, these models are constrained by oversimplified assumptions (e.g., idealized two-level erbium systems and a constant gain spectrum shape under all channel configurations) and neglect dynamic effects, manufacturing variability, and aging factors, reducing accuracy in practical deployments. Under dynamic channel loading conditions, the inaccuracy is further exacerbated for three reasons:

1. Critical EDFA parameters (e.g., Er^{3+} concentration and fiber length) are rarely available for commercial units and are difficult to measure accurately [10]
2. Commercial amplifiers exhibit complex multi-stage architectures that include gain-flattening filters, photodetectors, couplers, and variable optical attenuator (VOA) elements that are typically emitted in two-stage analytical models [11].
3. Physical models struggle to reproduce spectral hole burning (SHB) behavior [12].

Recently, machine learning (ML)-based EDFA models have been proposed to overcome these limitations by collecting measurements from EDFA devices [13]. Several methods have been proposed for predicting the EDFA gain spectrum under varying input power levels and channel loading configurations [14,15]. Despite these advancements, comparing reported ML models is difficult because the literature uses different experimental settings, measurement resolutions, and train–test splits. To mitigate the high measurement requirements of a purely data-driven approach, hybrid ML models have been proposed that combine physical EDFA models with ML techniques, achieving a favorable accuracy versus measurement trade-off [16]. Despite these innovations, even the best-performing ML techniques often require extensive labeled datasets that are impractical to collect on a live network and are typically trained on single-vendor devices, limiting their generalizability.

These data and generalization constraints become even more pronounced once individual amplifier models are stitched together to predict end-to-end system behavior. When per-component EDFA models are cascaded to form a multi-span optical link model, uncompensated error accumulation can severely degrade prediction accuracy [17,18]. Two broad adaptive strategies have been investigated: (1) *parameter refinement* methods that fine-tune physical model parameters after deployment but only partly correct the accumulated error [4] and (2) *cascaded learning* methods, which treat the entire link as an end-to-end differentiable chain of component models and use sparse link-level measurements to back-propagate corrections, thereby reducing the link error to the level of a single component [19,20]. Cascaded learning effectiveness thus relies on a high-fidelity component model of the EDFA gain profile, even though some margin for component-level error remains.

Building and maintaining such high-fidelity models for every amplifier in a large, heterogeneous network is, however, both computational and measurement-intensive. Transfer learning (TL) is a promising path to reduce data dependency by leveraging knowledge from source domains to accelerate learning in target domains [21]. Specifically, a well-characterized EDFA model can be used to model other devices with only limited additional measurements. Recent research showed that TL can be effective for modeling same-type EDFAs (e.g., booster-to-booster), achieving accurate predictions with just 0.5% of the target dataset [22]. However, the application of TL across different amplifier types and vendors (e.g., booster and pre-amplifier) remains under-explored. Particularly, in large networks comprising EDFAs from multiple vendors, TL can offer a viable means to characterize the power dynamics using minimal additional data. In addition, existing models only consider external features such as input power and gain spectra while overlooking internal telemetry that is readily available in modern commercial EDFAs and could enhance generalization.

In our prior work [23], we introduced a semi-supervised self-normalizing neural network (SS-NN) framework that integrates internal EDFA features for gain spectrum modeling, enabling transfer learning across in-built booster and preamp EDFAs in reconfigurable optical add-drop multiplexers (ROADMs). Building upon this foundation, the present work extends our previous contribution in several key ways:

- We extend our experimental validation by collecting gain spectrum measurements on four in-line amplifiers (ILAs) [without embedded optical channel monitors (OCMs)] and investigate the performance of the SS-NN model on different EDFA types.
- We investigate whether transfer learning can be applied across different EDFA types (such as boosters/preamps), as well as different manufacturers (such as Lumentum in-ROADM EDFAs/Juniper ILAs).
- We introduce a heterogeneous TL method that utilizes covariance matching (CORAL loss) to align the feature representations between the source and target domains, thereby addressing the challenges associated with feature mismatches across diverse EDFA types and vendors.

The contributions of this work can be summarized as follows:

1. We develop a novel SS-NN architecture that integrates both external and internal EDFA features. By combining unsupervised pre-training with supervised fine-tuning, our approach significantly reduces the number of measurements required for effective training.
2. We introduce a few-shot transfer learning mechanism that enables efficient knowledge transfer from one EDFA to another.
3. We incorporate a covariance matching technique (CORAL loss) for heterogeneous transfer learning, enhancing domain adaptability across EDFAs with differing feature sets.
4. We perform a comprehensive experimental evaluation on 26 commercial-grade EDFAs—including boosters, pre-amplifiers, and ILAs from the COSMOS and Open Ireland testbeds—demonstrating improved model performance and reduced measurement requirements.

The remainder of this paper is organized as follows: Section 2 describes the experimental setup and data collection methodology for EDFA gain spectrum measurements

across the COSMOS and Open Ireland testbeds. Section 3 describes the proposed SS-NN model architecture, the training process, and the results for directly trained models. Section 4 discusses the homogeneous and heterogeneous TL methods and results for model transfer across different EDFA types. Finally, Section 5 summarizes our findings.

2. EXPERIMENTAL SETUP AND DATA COLLECTION

This section outlines the experimental setup and data collection methodology for EDFAs in the Open Ireland and PAWR COSMOS testbeds. The Open Ireland testbed [24] is an open reconfigurable optical-wireless testbed in Dublin, Ireland, while the PAWR COSMOS testbed [19,25] is a city-scale programmable testbed deployed in Manhattan, USA.

A. Experimental Setup for Booster/Preamp

Gain spectrum measurements were conducted across multiple C-band wavelengths using commercial-grade Lumentum ROADM-20 units—three deployed in the Open Ireland testbed and eight in PAWR COSMOS. Each unit contains two EDFAs, yielding data from 11 booster and 11 pre-amplifier EDFAs. Figure 1 illustrates the experimental topology. A broadband source generates 95 × 50 GHz wavelength division multiplexing (WDM) channels in the C-band following the International Telecommunication Union (ITU) dense wavelength division multiplexing (DWDM) 50 GHz grid specification. Data collection procedures were standardized across both testbeds [10].

For booster measurements, the MUX wavelength selective switch (WSS) flattens channels and controls power and loading configurations. In pre-amplifier measurements, the comb source connects to an auxiliary ROADM's line-in port, with its DEMUX managing power and channel loading before forwarding the signal to the EDFA under test. Input and output power spectra for all 95 channels are recorded via built-in OCMs, while total input/output power through the EDFAs is captured via integrated power monitors (PMs). Additionally, we collect the internal VOA input/output power (P^V_{in}, P^V_{out}) and attenuation (P^V_{attn}) measurements.

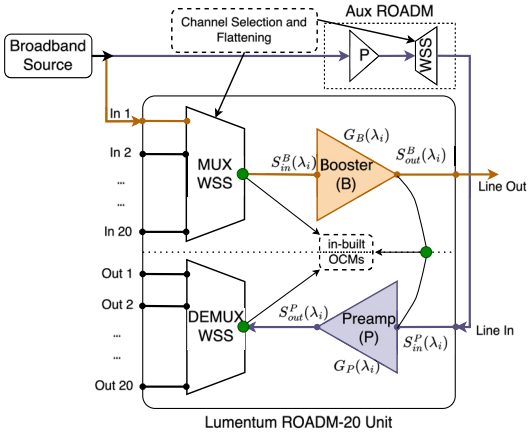

Fig. 1. Experimental setup for characterization of booster/pre-amplifier EDFAs of Lumentum ROADMs in the COSMOS and Open Ireland testbeds.

B. Experimental Setup for ILA

We collect gain spectrum measurements from two commercial-grade, bidirectional Juniper TCX-1000 ILA units deployed in the Open Ireland testbed. The data are collected for both forward and backward directions (typically denoted as AB and BA). Although housed on the same equipment, the forward and backward direction amplifiers function independently.

Figure 2 shows the experimental topology; as in previous setups for boosters/preamps, a broadband source is used to generate 95 × 50 GHz WDM channels in the C-band. Unlike ROADMs, ILAs do not have a built-in OCM, which makes data collection difficult. Therefore, we employ two auxiliary ROADMs to collect gain spectrum measurements from both EDFAs in a single ILA. For the forward direction (AB), the input and output power spectra are collected using built-in OCM at the DEMUX output of the first auxiliary ROADM and at the MUX input of the second auxiliary ROADM, and vice versa for the backward direction EDFA (BA). The total input/output power through the ILAs is collected through built-in PM in the ILAs. It should be noted that there will be insertion loss between the actual input/output power spectra between auxiliary ROADMs and the input/output of the ILAs. To compensate for these losses, a scaling factor (σ) is applied to reflect the actual input/output power spectrum values at the ILA input/output, respectively. For each measurement, the scaling factor (σ) is calculated as follows:

$$\sigma_{\text{in}} = P^{\text{Aux}}_{\text{out}} / P^{\text{ILA}}_{\text{in}}, \tag{1}$$

and

$$\sigma_{\text{out}} = P^{\text{ILA}}_{\text{out}} / P^{\text{Aux}}_{\text{in}}, \tag{2}$$

where $P^{\text{Aux}}_{\text{in}}$ and $P^{\text{Aux}}_{\text{out}}$ are the total input and output powers of the input and output ROADMs in milliwatts (mW), respectively. $P^{\text{ILA}}_{\text{in}}$ and $P^{\text{ILA}}_{\text{out}}$ are the total input and output powers of the ILA EDFA under test. The total power readings are measured through the in-built PMs in both ROADMs and ILAs. From Eqs. (1) and (2), the normalized power readings for each channel i are calculated as follows:

$$P_{\text{in}}(\lambda_i) = \sigma_{\text{in}} \cdot P^{\text{Aux}}_{\text{in}}(\lambda_i), \tag{3}$$

and

$$P_{\text{out}}(\lambda_i) = \sigma_{\text{out}} \cdot P^{\text{Aux}}_{\text{out}}(\lambda_i), \tag{4}$$

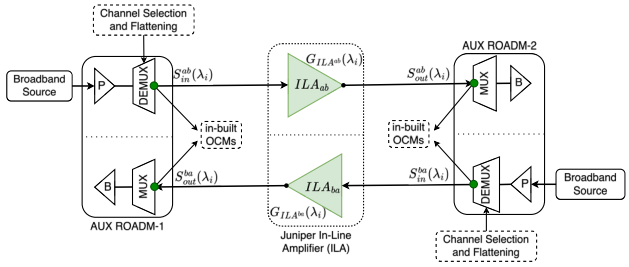

Fig. 2. Experimental setup for characterization of the Juniper TCX-1000 ILAs in the Open Ireland testbed.

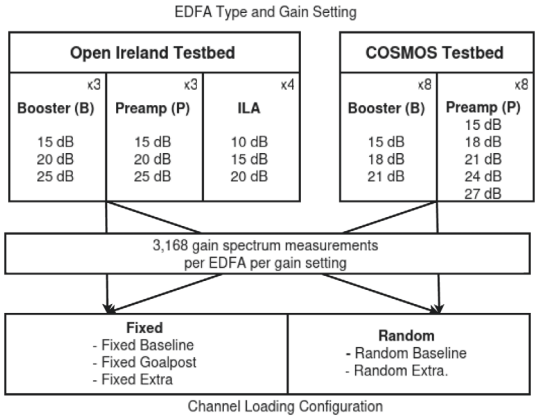

Fig. 3. Complete set of measurements collected from EDFAs across the COSMOS and Open Ireland testbeds.

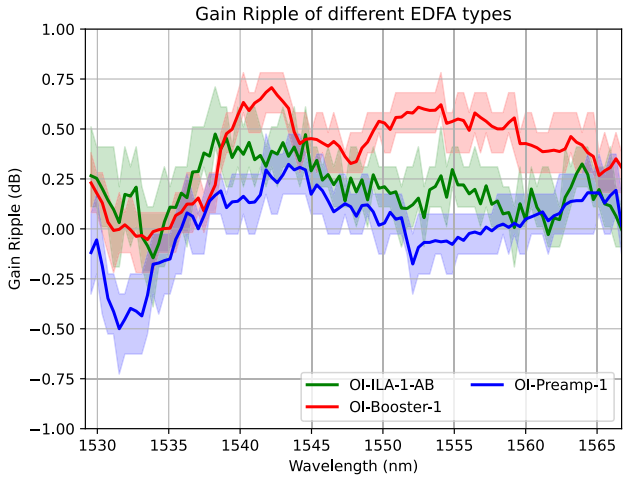

Fig. 4. Measured gain ripple for different EDFA types (ILA, booster, and preamp) at the same target gain setting of 15 dB. One device of each EDFA type used in the Open Ireland testbed is shown. Solid lines show the mean gain ripple, while the shaded areas represent the min/max range of gain ripple values.

where $P_{\text{in}}^{\text{Aux}}(\lambda_i)$ and $P_{\text{out}}^{\text{Aux}}(\lambda_i)$ are the measured OCM power readings at the auxiliary ROADMs for each channel i in the 95×50 GHz DWDM channels. It should be noted that, unlike boosters/pre-amplifiers, the internal VOA measurements are not exposed for ILAs.

C. Measurement Configuration

The measurement schema for data collection is shown in Fig. 3. In the Open Ireland testbed, all booster and preamp EDFAs were characterized at target gain settings of 15/20/25 dB, while in the COSMOS testbed, the target gains were set to 15/18/21 dB for boosters and 15/18/21/24/27 dB for pre-amplifiers in high gain mode. All ILAs were measured in the Open Ireland testbed at target gains of 10/15/20 dB in low gain mode. All measurements were taken with a 0 dB gain tilt. The use of varying gain configurations across different EDFA types and testbeds is intended to replicate the operational diversity encountered in different network environments. The dataset comprises 3168 gain measurements (collected across multiple wavelengths) per EDFA for each designated target gain setting. Overall, the COSMOS testbed yielded 202,752 measurements across 11 boosters and 11 preamps. In the Open Ireland testbed, a total of 57,024 measurements were taken in three boosters and three preamps, while 38,016 measurements were taken from four EDFAs in two ILAs (in both the forward and backward directions). All measurements were collected under three different channel loading configurations:

1. *Fixed*: This includes fully loaded (WDM), half-loaded (lower/upper spectrum and even/odd numbered channels), adjacent single/double channel loads, as well as a complete set of single/double channel configurations.
2. *Random allocation*: This configuration comprises random channel configurations from small-scale single-channel loads to fully loaded setups.
3. *Goalpost allocation*: This includes structured channel-loading configurations across different wavelength bands (short, medium, and long wavelengths), with both balanced and unbalanced loads across the considered bands.

Figure 4 shows the measured gain ripple at a target gain setting of 15 dB for a booster, pre-amplifier, and ILA in the Open Ireland testbed. It can be seen that different EDFA types exhibit different gain ripples. Figure 5 shows the gain ripples of different devices measured across the booster, pre-amplifier, and ILA. It can be seen that even devices from the same manufacturer exhibit different gain ripples at the same target gain setting.

3. MODEL ARCHITECTURE

The SS-NN model is designed to address two challenges in EDFA gain spectrum characterization: limited training data and poor transferability across different EDFA devices. The proposed model integrates internal features of the EDFA, a self-normalizing activation function, and a two-stage training process, which enables few-shot learning and improved generalization capabilities. In the following sections, we provide a detailed explanation of the model architecture and training methodology.

A. SS-NN Model

Existing ML approaches for EDFA modeling often treat the amplifier as a black box, relying solely on input/output spectra. For boosters and preamps, we incorporate three key features derived from the internal VOA: input/output power (P_{in}^{V}, P_{out}^{V}) and attenuation (P_{attn}^{V}). VOAs are an internal component of EDFAs, which indirectly influence the shape of the gain profile by acting on the signal's input powers. This is done to ensure the EDFA operates in its design average inversion for a flat spectrum gain profile, which matches the gain flattening filter (GFF) attenuation [26]. The VOA attenuation is controlled automatically in the EDFA based on the model's gain dynamic range, and it grants intrinsic information on the operation of each EDFA. In addition to the VOA parameters, the full input feature set includes

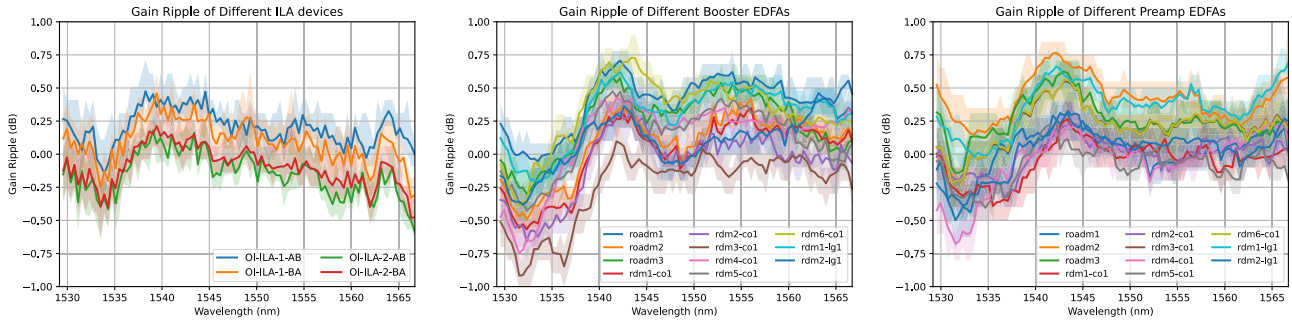

Fig. 5. Measured gain ripple of the same EDFA type, across different devices at the same target gain setting. Solid lines show the mean gain ripple, while the shaded areas represent the min/max range of gain ripple values.

the target gain (G_0) with constant-gain configuration, total EDFA input/output power (P_{in}, P_{out}), input power spectrum ($\bar{P}(\lambda_i) = [P(\lambda_1), P(\lambda_2), P(\lambda_3), \ldots, P(\lambda_{95})]$), and a binary channel-loading vector $\bar{C} = [c_i]_{i=1}^{95}$, where

$$c_i = \begin{cases} 1, & \text{if the } i\text{th wavelength channel is active,} \\ 0, & \text{otherwise.} \end{cases} \quad \textbf{(5)}$$

Figure 6 shows the overall architecture of the SS-NN model, which utilizes a four-hidden layer topology with 200/200/100/100 neurons. The input layer consists of 196 neurons, while the output layer consists of 95 neurons in all cases, predicting the gain spectrum at 95 wavelengths, $G(\lambda_i) = [G(\lambda_1), \ldots, G(\lambda_{95})]$. The final model architecture was determined through a random search over the subspace of (i) the number of hidden layers $\in \{1, 2, 4, 8, 10\}$ and (ii) the number of neurons in each hidden layer $\in \{50, 95, 100, 150, 200, 300\}$. The final model architecture with four hidden layers and a neuron count of (200–200–100–100) was chosen optimally on the basis of model accuracy as well as pre-training computational cost. The model is based on self-normalizing neural networks (SNNs) with a scaled exponential linear unit (SELU) as the activation function, where the SELU function is defined as

$$\text{selu}(x) = \lambda x \cdot I(x > 0) + \alpha(e^x - 1) \cdot I(x \leq 0), \quad \textbf{(6)}$$

where $\alpha = 1.673$ and $\lambda = 1.050$, and I denotes the indicator function. In ILAs, the missing internal VOA features are imputed with an extreme value of -999, in order to saturate the output of input neurons corresponding to the missing input features. In neurons with the SELU activation function, the gradient becomes saturated at highly negative inputs such as -999, essentially acting as a dead neuron.

Typically, batch normalization is commonly applied to stabilize training by normalizing the outputs of hidden layers [22]. However, batch normalization tends to perform poorly when training with limited data, as it relies on the estimation of batch statistics that can be unreliable in such scenarios [27]. In contrast, SNNs with the SELU [28] activation function ensure stable variance propagation across layers, even with less data. This self-normalizing property is a key factor enabling the SS-NN architecture to effectively perform one-shot training and achieve cross-device transferability.

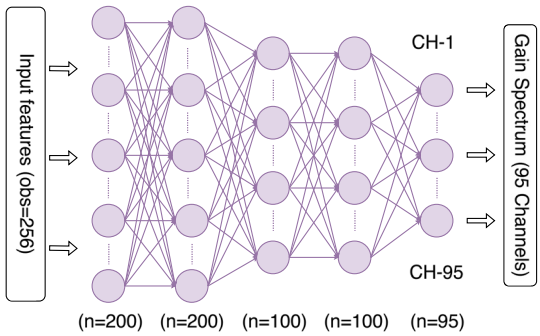

Fig. 6. SS-NN model structure with five layers.

B. Training Process

The proposed model employs a two-phase training methodology comprising unsupervised pre-training [29,30] followed by supervised fine-tuning [31]. This approach is driven by two considerations:

1. Unsupervised pre-training leverages readily available unlabeled input power spectrum measurements. These data points are more accessible and can also be simulated in flat spectrum scenarios, which significantly mitigates the need for resource-intensive labeled datasets.
2. The pre-training process helps achieve a better weight initialization compared to random initialization and captures more complex dependencies between parameters [30]. Furthermore, it introduces implicit regularization properties that improve generalization, particularly beneficial when transferring knowledge between different EDFA configurations and units [32].

Figure 7 illustrates the training procedure. During unsupervised pre-training, weights are initialized progressively using a layer-wise denoising strategy, where each layer is trained on 512 noise-augmented measurements per gain setting. Gaussian noise is added to the inputs, and an autoencoder structure, which matches the dimensionality of the original feature set, is tasked with reconstructing denoised outputs. The training employs mean squared error (MSE) loss to evaluate reconstruction fidelity under noise corruption. Each layer undergoes sequential training for 1800 epochs (learning rate $= 1e-03$), with its weights frozen upon initialization to preserve hierarchical feature representations during subsequent training stages.

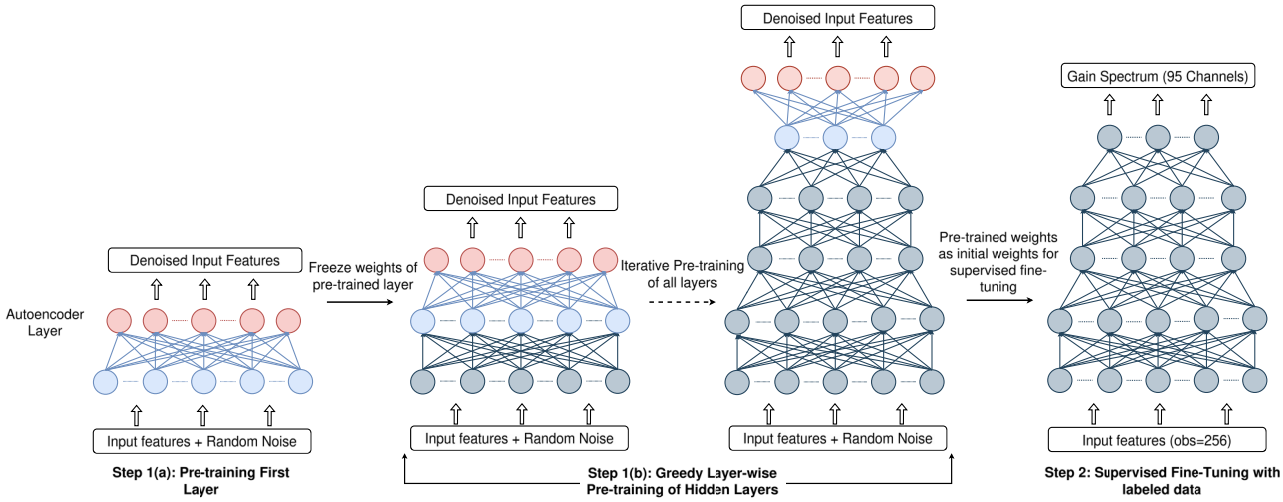

Fig. 7. SS-NN model training framework. Steps 1(a) and 1(b) show the greedy layer-wise pretraining of hidden layers using unsupervised pre-training. This pre-trained model forms the basis for Step 2, where supervised fine-tuning is performed with 256 labeled measurements.

Once all layers are pre-trained, the finalized model serves as the foundation for supervised fine-tuning.

For supervised fine-tuning, the model is optimized using a labeled dataset of 512 measurements containing fully loaded and random channel configurations. We utilize a custom weighted MSE loss function, and for the kth measurement, it is defined as

$$\mathrm{MSE}_k = \frac{1}{\sum_{i=1}^{95} c_i^k} \cdot \sum_{i=1}^{95} c_i^k \cdot \left[g_{\mathrm{pred}}^k(\lambda_i) - g_{\mathrm{meas}}^k(\lambda_i) \right]^2. \quad (7)$$

In this phase, the model is fine-tuned using the Adam optimizer with a learning rate of 1e − 03, over 1200 epochs with a batch size of 32. To account for the wide dynamic range of gain-spectrum measurements obtained across different power levels and channel configurations, we apply gradient clipping [33] with a threshold of 1.0 to stabilize training and prevent divergence. During the entire training process, hyperparameter optimization was performed using a random search to fine-tune the learning rate, the number of layers, the number of neurons in each layer, and the batch size.

C. Training and Test Sets

We compare the SS-NN model with a benchmark state-of-the-art method [10,22]. For equivalent comparison, we follow the same dataset selection criteria. For each gain setting, we split the dataset into a training/test set ratio of 0.86/0.14. The test set contains 436 gain spectrum measurements per gain setting. This test set contains a mixture of random and goalpost channel loading measurements, which represent a diverse set of channel loading configurations. Note that, although the SS-NN model uses less data for training, we allocate a larger portion of training data for the benchmark model, which uses 2732 measurements per gain setting.

D. Model Performance

We compare the SS-NN model with the benchmark model using the same set of features to highlight the benefits of our approach. Additionally, for boosters/preamps, we demonstrate the advantage of incorporating internal EDFA features by comparing the results of the SS-NN model with and without including these additional features.

Figure 8 shows the distribution of absolute errors of the gain spectrum predicted by the benchmark model, the SS-NN model using the same set of features, and the SS-NN model with additional internal VOA features. The errors are calculated across 11 boosters, 11 pre-amplifiers, and 4 ILA EDFAs in the Open Ireland and COSMOS testbeds on the test set with random and goalpost channel configurations. For boosters, the SS-NN model achieves a mean absolute error of 0.05 and 0.07 dB under the random and goalpost channel configurations. This is comparable to the performance of the benchmark model, which uses a considerably higher number of measurements (8196 measurements), compared to a total of 1792 measurements utilized by the SS-NN model. Importantly, SS-NN models exhibit a superior error distribution, with a narrow inter-quartile range, and a 95th percentile error of 0.15/0.24 dB, compared to 0.38/0.16 dB by the benchmark model, across the goalpost/random test sets.

For preamps, the SS-NN model achieves a mean absolute error of 0.08/0.05 dB using the same set of features and 0.07/0.05 dB using additional internal features across goalpost/random channel configurations. This is marginally better than the benchmark model, which achieves a 0.09/0.05 dB error across goalpost/random test sets. Additionally, the distribution of errors for SS-NN models is more stable, with a narrow inter-quartile range and a 95th percentile error within 0.3 dB across both channel configurations, showing that the SS-NN model generalizes well to unseen channel configurations even when trained with reduced measurements. It should be noted that using additional internal features when directly training EDFA models, i.e., training on the source EDFA's

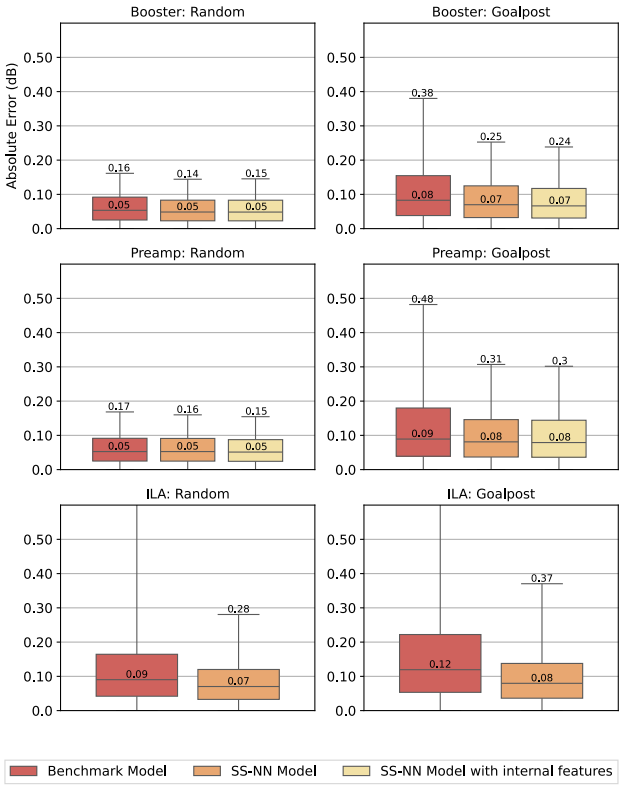

Fig. 8. Boxplot distribution of absolute errors across all 11 booster, 11 pre-amplifier, and 4 ILA EDFAs for goalpost and random channel loading. The boxes denote the inter-quartile range, and the whiskers denote the min/95th percentile.

measurements without TL, does not provide additional performance.

For ILAs, we only develop the SS-NN model against the benchmark model due to the non-availability of internal VOA measurements in Juniper ILAs. Moreover, because of the measurement setup involving external ROADMs, the ILA input and output spectra exhibit more noise compared to booster/preamps—which are in-ROADM EDFAs. In this case, the SS-NN model performs much better than the benchmark model both in terms of mean absolute error (MAE) and distribution. The model achieves 0.07/0.08 dB error on goalpost datasets, with a 95th percentile error within 0.4 dB across both the channel configurations.

Table 1. Computational Overhead for the SS-NN Model with the Largest Number of Features (Including Internal Features)[a]

Parameter	Value
Trainable parameters	119,395
Batch size	32
Total training time	1180 s
Inference latency	1.1 ms
Inference floating point operations (FLOPs)	238,095

[a]Latency times are averaged over the entire dataset and approximated to the first decimal place.

Table 1 summarizes the computational overhead for the SS-NN model, including the total number of trainable parameters, per-inference FLOPs, and approximate training times. All experiments were conducted on an Nvidia RTX 4090 GPU in the Open Ireland testbed, with 24 GB of VRAM and a peak 32 bit floating-point throughput of 82.58 tera floating-point operations per second (TFLOPS) [34]. Training latency is substantially higher than inference latency because of a smaller batch size of 32. This was driven by our finding that lower batch size improved convergence when training with a limited set of measurements in the fine-tuning step. The inference time per observation is ≈1.1 ms, which is considerably faster than the 6 s temporal resolution of in-device OCM measurements on the Lumentum ROADM-20 module [10].

4. TRANSFER LEARNING

Transfer learning improves performance on a target task by leveraging information from a related source domain [21]. In the context of ML algorithms, TL involves fine-tuning a model trained on a source domain (D_S) using new feature representations from a target domain (D_T). This approach is particularly useful for modeling the gain spectrum in EDFAs, as it can reduce the required measurement time.

For a source domain D_S with n features, the feature representation for an instance i can be expressed as

$$X_s^i = \{x_{s1}^i, x_{s2}^i, \ldots, x_{sn}^i\}.$$

Similarly, for a target domain D_T with m features, an instance j is represented by

$$X_t^j = \{x_{t1}^j, x_{t2}^j, \ldots, x_{tm}^j\}.$$

Given a model f_s trained on D_S, TL is performed by augmenting f_s with additional instances from D_T. When $n = m$, TL is considered *homogeneous*, whereas when $n \neq m$, it is classified as *heterogeneous*. In the latter case, differences in feature space dimensionality can lead to information loss and degraded model performance. Furthermore, if the source domain contains more information than the target ($n > m$), negative transfer may occur [35,36], where the source knowledge adversely affects the target model. Although training both models on only shared features could improve performance, it would compromise the source model's accuracy.

In the case of EDFA TL, a similar trade-off is evident. As discussed previously, incorporating additional internal EDFA features enhances model performance. However, a source model trained with these extra features cannot be directly transferred to a target EDFA (e.g., an ILA) that lacks the ability to measure them. Conversely, training a source model without these features would enable direct transfer but require maintaining dual models or accepting suboptimal performance in the source domain. Based on these considerations, we categorize EDFA TL as follows:

1. **Homogeneous TL:** This applies to TL between EDFAs with identical feature spaces, including
   (a) *same-type transfers*: $B \leftrightarrow B$, $P \leftrightarrow P$, $ILA \leftrightarrow ILA$,
   (b) *cross-type transfers*: $B \leftrightarrow P$.

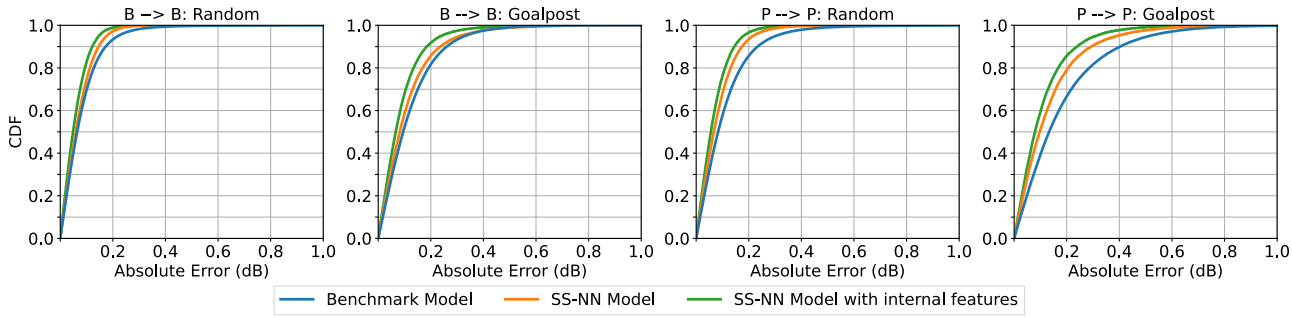

Fig. 9. CDF plots of absolute errors for TL across same-type homogeneous TL for (i) booster-to-booster TL: random channel loading, (ii) booster-to-booster TL: goalpost channel loading, (iii) preamp-to-preamp TL: random channel loading, and (iv) preamp-to-preamp TL: goalpost channel loading.

2. **Heterogeneous TL:** This covers transfers between models trained on boosters/pre-amplifiers (with the full feature set) and ILAs (which lack internal features), i.e., $B/P \leftrightarrow ILA$.

Note that same-type TL and cross-type TL between boosters and preamps are categorized as homogeneous TL, as are transfers among ILAs, given their consistent feature space. In contrast, TL between an ILA and a booster/preamp is heterogeneous due to the missing internal VOA features in ILAs.

In this section, we demonstrate that TL for SS-NN models can effectively model the gain spectrum across different EDFAs with minimal additional data. Section 4.A details the TL techniques and results for same-type and cross-type transfers, while Section 4.B investigates transfers between EDFAs with differing feature sets and presents an updated TL approach for heterogeneous scenarios.

A. Homogeneous Transfer Learning

To transfer an existing model from a source EDFA to a target EDFA, we fine-tune the source model using a single fully loaded measurement per target gain setting. The model is fine-tuned using the Adam optimizer for 10,000 epochs, employing the MSE loss function as in Eq. (7) and a gradient clipping threshold of 1.0 to ensure stable training. Instead of using a uniform learning rate, a differential learning rate is applied across layers. Rather than completely freezing initial layers, we adopt an exponentially decaying, layer-wise differential learning rate strategy [37,38]. The layer-specific learning rate α_l for layer l in a neural network with a total of L layers can be given as

$$\alpha_l = \alpha_0 \cdot 10^{\theta \cdot (L-l)},$$

where α_0 is a constant denoting the base learning rate, and θ is the exponential multiplier. Empirically, we found the values of $\alpha_0 = 10^{-3}$ and $\theta = -1$ to work well with convergence. This strategy enables the output layer to adapt rapidly to the target EDFA's specific characteristics, while the lower layers are fine-tuned more conservatively to prevent overfitting and promote robust generalization.

1. Same-Type Transfer Learning

Figure 9 presents the cumulative distribution function (CDF) of absolute errors for same-type TL among booster/preamp EDFAs under both random and goalpost channel loading configurations. Error distributions are provided for all possible booster-to-booster (B → B) transfers across 11 booster EDFAs and preamp-to-preamp (P → P) transfers across 11 preamp EDFAs from the COSMOS and Open Ireland testbeds. We compare the SS-NN model with (i) the benchmark TL technique [10], which employs the same NN architecture and (ii) the SS-NN model trained without internal EDFA features. This comparison highlights the benefit of incorporating additional internal variables along with the proposed modeling technique. Importantly, the benchmark model requires 13 additional measurements per target gain setting for TL, whereas the SS-NN model requires only 1.

The error distribution—particularly in the 90th percentile—is critical since neural network (NN) models are prone to overfitting on data subsets. The SS-NN models, however, exhibit a more favorable error distribution for both B → B and P → P transfers under both channel loading configurations. While training directly with internal EDFA features does not yield a substantial performance benefit, these variables enhance TL performance, indicating that they contain distinctive information about the EDFAs under test.

Figure 10 shows the CDF of absolute errors for TL across four ILAs under random and goalpost channel loading configurations in the Open Ireland testbed. The SS-NN model demonstrates a superior error distribution compared to the

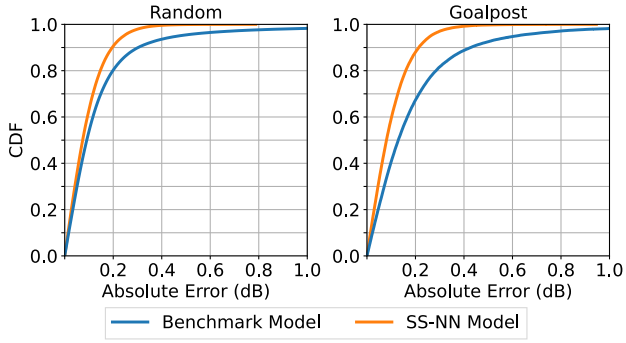

Fig. 10. CDF plots of absolute errors for TL across ILAs in the Open Ireland testbed.

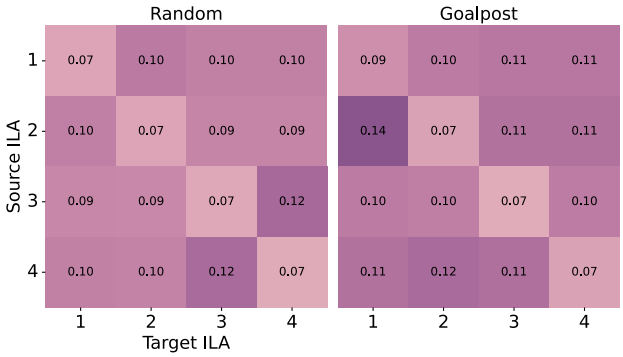

Fig. 11. Transfer learning MAE matrix of the SS-NN model for four ILAs in the Open Ireland testbed, shown across random and goalpost channel loading configurations. The (i, j) entry corresponds to the TL-based ILA model, where the ith and jth EDFA serve as the source and target models, respectively.

benchmark model, particularly in the 95th percentile where the absolute error is $\leq$0.2 dB in both test sets. Figure 11 displays the MAE matrices (in dB) for the SS-NN model across four ILAs under both channel loading configurations. In each matrix, the diagonal entry (i, i) corresponds to a directly trained model (without TL), while the off-diagonal entry (i, j) corresponds to a transferred EDFA model with the ith and jth EDFAs serving as the source and target, respectively. These results indicate that the SS-NN model generalizes well to other ILAs even when TL is performed with only a single measurement per target gain setting. Specifically, the SS-NN-based TL model achieves a per-EDFA MAE below 0.14 dB across both test sets.

2. Cross-Type Transfer Learning

Cross-type TL in EDFAs is more challenging than same-type TL due to significant differences in the feature space. Boosters and preamps are designed for different purposes: preamps are low-noise, high-gain EDFAs that can substantially improve a transceiver's SNR, whereas boosters are optimized for operation in the gain saturation range to enable long-distance transmission [26].

Figure 12 shows the CDF plots of absolute errors for cross-type TL between boosters and preamps, evaluated under both random and goalpost channel loading configurations. The results include error distributions for all possible combinations of 11 boosters and 11 preamps from the COSMOS and Open Ireland testbeds. As before, comparisons are made with the benchmark TL model and the SS-NN model trained without internal EDFA features. Although cross-type TL performance is slightly inferior to same-type TL, the SS-NN model still demonstrates a more favorable error distribution than the benchmark. In particular, under random channel loading, the SS-NN model achieves an absolute error of $\leq$0.2 dB for 95% of the measurements and outperforms the benchmark in goalpost configurations. Enhancing the error distribution in goalpost scenarios remains an area for future work.

B. Heterogeneous Transfer Learning

Transferring a source model trained on booster/preamp EDFAs (which incorporate internal EDFA features) to an ILA EDFA is non-trivial for the following reasons:

1. Booster/preamp EDFAs are in-ROADM devices, where gain measurements are obtained via built-in OCMs. In contrast, ILAs typically lack integrated OCMs; their gain spectrum measurements are collected using external OCMs. Although these external variables are accounted for during data processing, they introduce additional noise.
2. SS-NN models for booster/preamp EDFAs achieve optimal performance when trained with additional internal features. However, these features may not be available for ILAs due to vendor constraints. While it is possible to transfer a source model trained without these extra features, doing so would require maintaining dual models or sacrificing performance on the source EDFA.

TL in these cases suffers from non-significant differences in source and target feature distributions, as well as a feature mismatch. This often leads to significant changes in the model's input representations, requiring more measurements for domain adaptation. We utilize a covariance matching technique for TL—correlation alignment for deep domain adaptation (CORAL) [39,40]—in order to align the feature representations between the source and target models. Essentially, CORAL loss aligns the second-order statistics (i.e., covariance matrices) of the source and the target domain

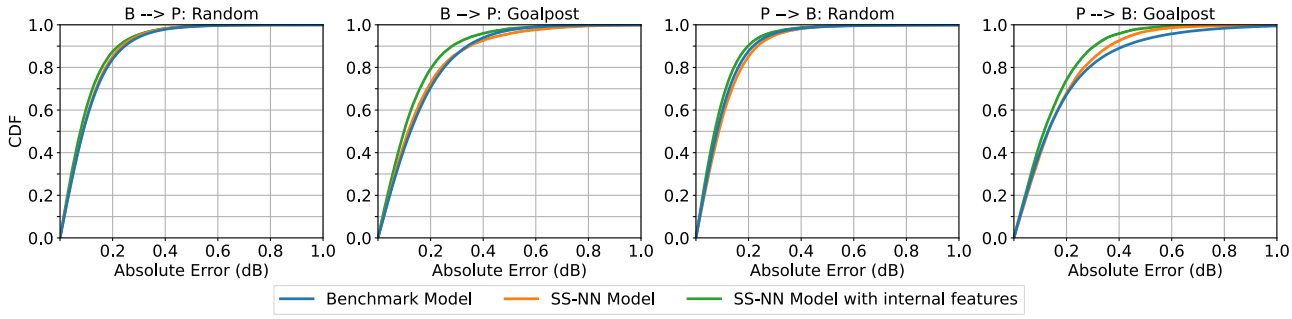

Fig. 12. CDF plots of absolute errors for TL across cross-type homogeneous TL for (i) booster-to-preamp TL: random channel loading, (ii) booster-to-preamp TL: goalpost channel loading, (iii) preamp-to-booster TL: random channel loading, and (iv) preamp-to-booster TL: goalpost channel loading.

for increased domain adaptability. The heterogeneous TL process is implemented as follows:

1. **Reference covariance computation:** From the training set of the source EDFA model, 128 randomly selected measurements are sampled. The output of the last hidden layer (with $d = 100$ neurons) is recorded; let this be represented by the matrix $\mathbf{F}_S \in \mathbb{R}^{N\times d}$, where $N = 128$. The covariance matrix $\mathbf{C}_S$ is then computed as

$$\mathbf{C}_S = \frac{1}{N-1}\left(\mathbf{F}_S - \bar{\mathbf{F}}_S\right)^{\mathrm{T}}\left(\mathbf{F}_S - \bar{\mathbf{F}}_S\right), \tag{8}$$

   where $\bar{\mathbf{F}}_S$ is the mean feature vector over the batch. This matrix is saved as a fixed reference.
2. **Feature imputation:** For the target EDFA, any missing internal VOA features are imputed with -999 to ensure consistency in the model architecture.
3. **Differential fine-tuning:** The source EDFA model is fine-tuned using m additional measurements from the target EDFA. A differential learning rate is applied: the output layer is updated with a larger learning rate of 1×10^{-2}, while each hidden layer is assigned a learning rate that is 10% of that of the subsequent layer. Moreover, the learning rate for each layer is halved every 2000 epochs.
4. **Loss function update:** During fine-tuning, for each batch, the target model's last hidden layer features $\mathbf{F}_T \in \mathbb{R}^{N\times d}$ are obtained and their covariance matrix $\mathbf{C}_T$ is computed similarly. From Eq. (7), the overall loss function $\mathcal{L}_k$ for the kth measurement is defined as

$$\mathcal{L}_k = \frac{1}{\sum_{i=1}^{95} c_i^k}\sum_{i=1}^{95} c_i^k\left[g_{\mathrm{pred}}^k(\lambda_i) - g_{\mathrm{meas}}^k(\lambda_i)\right]^2 + \lambda\cdot\frac{1}{4d^2}\left\|\mathbf{C}_S - \mathbf{C}_T\right\|_F^2, \tag{9}$$

   where $\|\cdot\|_F$ denotes the Frobenius norm, and λ is the weighting factor for the CORAL loss.
5. **Training:** Finally, the target EDFA model is trained with the updated loss function for 10,000 epochs, using $\lambda = 0.4$. A base learning rate (α_l) of $\alpha_l = 1e-02$ was found to perform better in this case, as compared to $\alpha_l = 1e-03$ used in homogeneous TL.

1. Model Performance

We first investigate the number of additional measurements required to effectively transfer a source EDFA model to a target EDFA. The additional measurements for fine-tuning are randomly loaded and fully loaded channel configurations. To evaluate this, we compare the TL performance of our proposed SS-NN model incorporating the CORAL loss with two alternative TL configurations:

1. **TL with a source model trained without additional features:** In this case, the source model is trained using only the basic features, excluding the internal EDFA measurements. This configuration serves as a baseline for TL when additional internal features are not available.
2. **TL with a source model trained with additional features (MSE loss):** Here, the source model is trained with the complete set of internal features, and TL is performed using the standard MSE loss.

It should be noted that the direction of transfer introduces fundamentally different challenges. In the (booster/preamp) → ILA TL scenario, there is an *information loss* problem: the source booster/preamp models are trained with additional internal features, which are missing in the target ILA dataset. In contrast, the ILA → (booster/preamp) TL represents a case of *information gain*—the source ILA model is trained with imputed additional features and is then transferred to a dataset with a complete feature set.

Figure 13 shows the MAE of TL models as a function of the number of additional measurements per target gain setting used during training. The results, aggregated across all combinations of EDFAs in the COSMOS and Open Ireland testbeds under both random and goalpost channel configurations, indicate that heterogeneous TL requires more data points than homogeneous TL, which needs only a single fully loaded gain spectrum measurement. Moreover, the performance of a full-feature set SS-NN model degrades even with a large number of additional measurements, exhibiting negative TL. In contrast, models employing the CORAL loss achieve performance comparable to those trained with a reduced feature set. Notably, the MAE is higher for (booster/preamp) → ILA TL than for ILA → (booster/preamp) TL. We attribute this asymmetry to the greater difficulty of *unlearning* the pre-trained source representations under a conservative training mechanism; optimizing this directional model asymmetry will be a focus of our future work.

Based on these results, we selected 32 additional measurements per target gain setting for ILA → (booster/preamp) TL and 48 additional measurements for (booster/preamp) → ILA TL. Figure 14 presents the CDFs for both TL scenarios across all EDFAs in the COSMOS and Open Ireland testbeds under random and goalpost allocations. Models employing the modified loss function with the CORAL loss perform similarly—and in some cases better—than the benchmark models. Specifically, for ILA → (booster/preamp) TL, the CORAL loss yields superior performance with an absolute error of $\leq$0.22 dB across all test sets. For (booster/preamp) → ILA TL, while models using CORAL loss outperform the benchmarks, overall performance is lower compared to other TL combinations, particularly under goalpost channel configurations.

These results demonstrate that incorporating CORAL loss into the SS-NN framework can enhance TL performance across EDFAs with different feature sets—eliminating the need to train and maintain multiple versions of the same model. Homogeneous TL is achieved with minimal additional measurements, while heterogeneous TL—particularly in the (booster/preamp) → ILA scenario—requires a larger dataset to overcome information loss and feature mismatches. The proposed approach not only mitigates negative transfer effects but also yields consistent performance as evidenced by improved error distributions and lower MAE values. These findings provide a basis for further investigation into

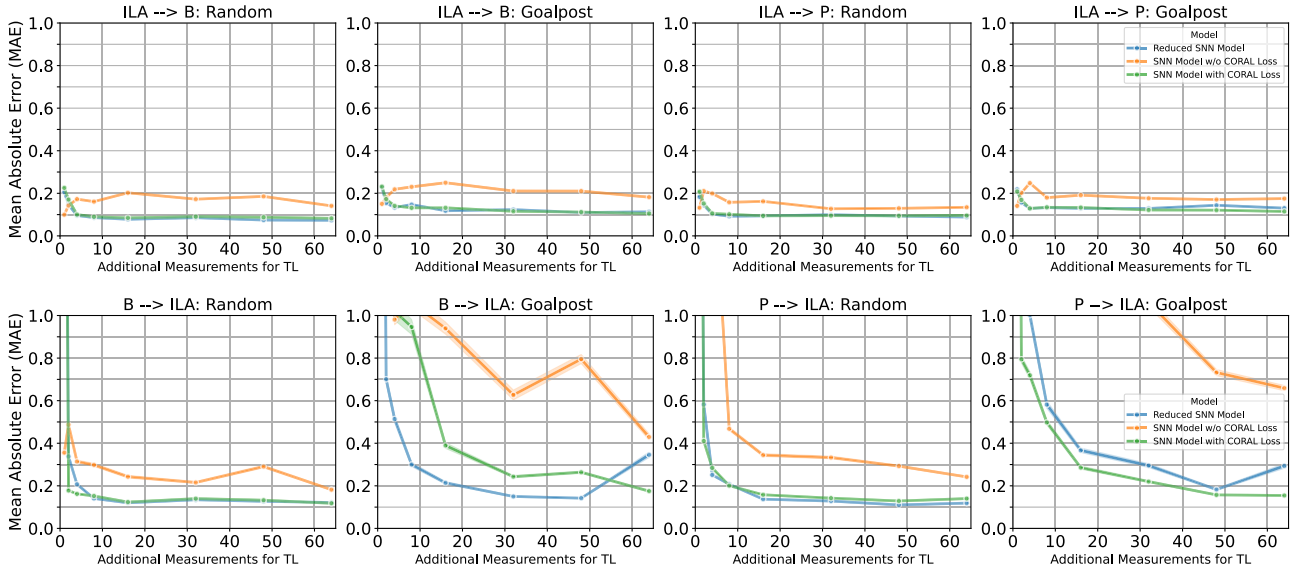

Fig. 13. Mean absolute error for TL across boosters/preamps and ILAs, shown across the number of additional measurements per target-gain setting required for TL. Results are shown for (i) the source model trained without internal EDFA features (reduced SNN model), (ii) the source model trained with internal EDFA features but TL without CORAL loss, and (iii) the source model trained with additional features and TL incorporating CORAL loss.

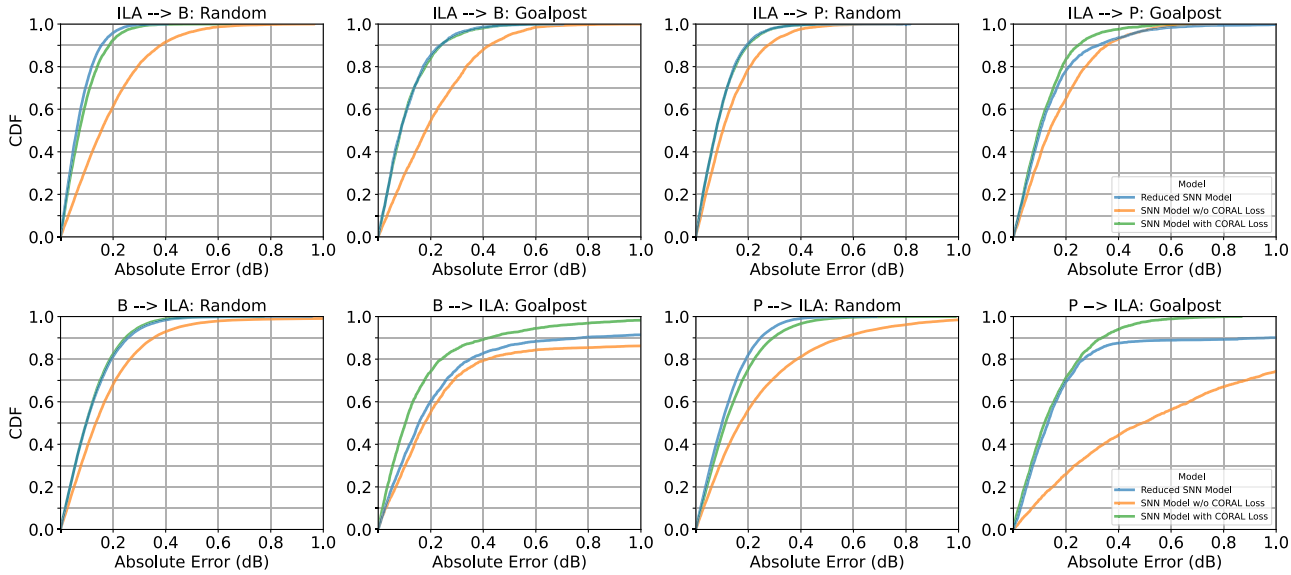

Fig. 14. CDF plots for TL across boosters/preamps and ILAs for random and goalpost channel loading configurations. (i) For TL from ILAs → booster/preamp, the target model is fine-tuned with 32 additional measurements per target-gain setting for each target EDFA, while for TL from booster/preamp → ILAs, 48 such additional measurements are required. Results are shown for (i) the source model trained without internal EDFA features (reduced SNN model), (ii) the source model trained with additional features but TL without CORAL loss, and (iii) the source model trained with additional features and TL incorporating CORAL loss.

covariance alignment techniques for transfer learning in optical networks, especially in multi-span field networks where multiple vendor technologies are used.

5. CONCLUSION AND FUTURE WORK

In this paper, we introduce a generalized few-shot transfer learning architecture for EDFA gain spectrum modeling that integrates internal amplifier features within a SS-NN. The proposed two-phase training process, combining unsupervised pre-training with supervised fine-tuning, enables the SS-NN model to achieve high accuracy with limited labeled data. Furthermore, by incorporating a covariance matching (CORAL) loss, our approach effectively addresses domain discrepancies in heterogeneous transfer learning scenarios, particularly between booster/pre-amplifier EDFAs and ILAs. Experimental results on the COSMOS and Open Ireland testbeds show that the proposed framework not only reduces

the required measurement overhead but also yields improved error distributions and lower mean absolute errors compared to state-of-the-art benchmarks.

Building on these findings, we aim to pursue several open challenges in the future. First, although an initial 10-month follow-up suggests that gain spectrum drift in the COSMOS EDFAs is below 0.1 dB [10], a multi-year dataset is required to quantify the impact of EDFA aging effects on model performance. Additionally, systematic exploration in the SHB region, especially under extreme channel configurations, could improve model reliability in power excursion scenarios. Finally, with the rise of flex-grid systems such as OSaaS [7], our future work will also focus on adapting these fixed-grid models to flex-grid operation.

Funding. Science Foundation Ireland (12/RC/2276_p2, 18/RI/5721, 22/FFP-A/10598); National Science Foundation (CNS-1827923, OAC-2029295, CNS-2112562, CNS-2330333).

REFERENCES

1. A. Fayad, T. Cinkler, and J. Rak, "Toward 6G optical fronthaul: a survey on enabling technologies and research perspectives," IEEE Commun. Surv. Tutorials **27**, 629–666 (2025).
2. M. Ruffini, M. Achouche, A. Arbelaez, *et al.*, "Access and metro network convergence for flexible end-to-end network design," J. Opt. Commun. Netw. **9**, 524–535 (2017).
3. A. Mahajan, K. Christodoulopoulos, R. Martínez, *et al.*, "Modeling EDFA gain ripple and filter penalties with machine learning for accurate QOT estimation," J. Lightwave Technol. **38**, 2616–2629 (2020).
4. M. P. Yankov, P. M. Kaminski, H. E. Hansen, *et al.*, "SNR optimization of multi-span fiber optic communication systems employing EDFAs with non-flat gain and noise figure," J. Lightwave Technol. **39**, 6824–6832 (2021).
5. R. T. Jones, K. R. H. Bottrill, N. Taengnoi, *et al.*, "Spectral power profile optimization of a field-deployed wavelength-division multiplexing network enabled by remote EDFA modeling," J. Opt. Commun. Netw. **15**, C192–C202 (2023).
6. Y. Liu, X. Liu, L. Liu, *et al.*, "Modeling EDFA gain: approaches and challenges," Photonics **8**, 417 (2021).
7. A. Raj, Z. Wang, F. Slyne, *et al.*, "Interference identification in multi-user optical spectrum as a service using convolutional neural networks," in *European Conference on Optical Communication (ECOC)* (2024).
8. A. Raj, D. C. Kilper, and M. Ruffini, "Interference detection in spectrum-blind multi-user optical spectrum as a service," J. Opt. Commun. Netw. **17**, C117–C126 (2025).
9. G. P. Agrawal, *Fiber-Optic Communication Systems* (Wiley, 2012).
10. Z. Wang, D. Kilper, and T. Chen, "An open EDFA gain spectrum dataset and its applications in data-driven EDFA gain modeling," J. Opt. Commun. Netw. **15**, 588–599 (2023).
11. "Lumentum terminal amplifier graybox," 2021, https://www.lumentum.com/en/products/dci-terminal-amplifier-graybox.
12. J. T. de Araujo, A. C. Meseguer, and J.-C. Antona, "Experimental model of EDFA spectral hole burning for WDM transmissions systems," in *Optical Fiber Communication Conference (OFC)* (Optica Publishing Group, 2023), paper M2E.1.
13. A. Raj, D. Kilper, and M. Ruffini, *Digital Twins for Optical Networks* (IET, 2024), Chap. 7, pp. 171–202.
14. S. Zhu, C. L. Gutterman, W. Mo, *et al.*, "Machine learning based prediction of erbium-doped fiber WDM line amplifier gain spectra," in *European Conference on Optical Communication (ECOC)* (IEEE, 2018).
15. F. da Ros, U. C. de Moura, and M. P. Yankov, "Machine learning-based EDFA gain model generalizable to multiple physical devices," in *European Conference on Optical Communications (ECOC)* (2020).
16. S. Zhu, C. Gutterman, A. D. Montiel, *et al.*, "Hybrid machine learning EDFA model," in *Optical Fiber Communication Conference (OFC)* (Optica Publishing Group, 2020), paper T4B.4.
17. Z. Wang, E. Akinrintoyo, D. Kilper, *et al.*, "Optical signal spectrum prediction using machine learning and in-line channel monitors in a multi-span ROADM system," in *European Conference and Exhibition on Optical Communication* (Optica Publishing Group, 2022), paper We3B.5.
18. Y. You, Z. Jiang, and C. Janz, "OSNR prediction using machine learning-based EDFA models," in *45th European Conference on Optical Communication (ECOC)* (IET, 2019).
19. T. Chen, J. Yu, A. Minakhmetov, *et al.*, "A software-defined programmable testbed for beyond 5G optical-wireless experimentation at city-scale," IEEE Netw. **36**, 90–99 (2022).
20. A. Raj, Z. Wang, F. Slyne, *et al.*, "Multi-span optical power spectrum evolution modeling using ML-based multi-decoder attention framework," in *50th European Conference on Optical Communication (ECOC)* (VDE, 2024), pp. 1037–1040.
21. F. Zhuang, Z. Qi, K. Duan, *et al.*, "A comprehensive survey on transfer learning," Proc. IEEE **109**, 43–76 (2021).
22. Z. Wang, D. Kilper, and T. Chen, "Transfer learning-based ROADM EDFA wavelength dependent gain prediction using minimized data collection," in *Optical Fiber Communication Conference (OFC)* (Optica Publishing Group, 2023), paper Th2A.1.
23. A. Raj, D. Kilper, and M. Ruffini, "Transfer learning for EDFA gain modeling: a semi-supervised approach using internal amplifier features," in *IEEE Future Networks World Forum (FNWF)* (2024).
24. CONNECT, "Open Ireland testbed, funded by Science Foundation Ireland" (2022).
25. D. Raychaudhuri, I. Seskar, G. Zussman, *et al.*, "Challenge: COSMOS: a city-scale programmable testbed for experimentation with advanced wireless," in *ACM International Conference on Mobile Computing and Networking (MobiCom)* (2020).
26. J. Zyskind and A. Srivastava, *Optically Amplified WDM Networks*, 1st ed. (Elsevier/Academic, 2011).
27. S. Ioffe, "Batch renormalization: towards reducing minibatch dependence in batch-normalized models," in *Advances in Neural Information Processing Systems* (Curran Associates, 2017), Vol. **30**.
28. G. Klambauer, T. Unterthiner, A. Mayr, *et al.*, "Self-normalizing neural networks," arXiv (2017).
29. Y. Bengio, P. Lamblin, D. Popovici, *et al.*, "Greedy layer-wise training of deep networks," in *Advances in Neural Information Processing Systems* (MIT, 2006), Vol. **19**.
30. J. Ge, S. Tang, J. Fan, *et al.*, "On the provable advantage of unsupervised pretraining," arXiv (2023).
31. N. J. Prottasha, A. A. Sami, M. Kowsher, *et al.*, "Transfer learning for sentiment analysis using BERT based supervised fine-tuning," Sensors **22**, 4157 (2022).
32. D. Erhan, A. Courville, Y. Bengio, *et al.*, "Why does unsupervised pre-training help deep learning?" in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics* (2010), pp. 201–208.
33. R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," arXiv (2012).
34. "Nvidia Ada GPU architecture," https://images.nvidia.com/aem-dam/Solutions/geforce/ada/nvidia-ada-gpu-architecture.pdf.
35. K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," J. Big Data **3**, 9 (2016).
36. S. Orouji, M. C. Liu, T. Korem, *et al.*, "Domain adaptation in small-scale and heterogeneous biological datasets," Sci. Adv. **10**, eadp6040 (2024).
37. J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," arXiv (2018).
38. J. Yosinski, J. Clune, Y. Bengio, *et al.*, "How transferable are features in deep neural networks?" arXiv (2014).
39. B. Sun and K. Saenko, "Deep coral: correlation alignment for deep domain adaptation," in *Computer Vision—ECCV 2016 Workshops* (Springer, 2016), pp. 443–450.
40. L. Li and Z. Zhang, "Semi-supervised domain adaptation by covariance matching," IEEE Trans. Pattern Anal. Mach. Intell. **41**, 2724–2739 (2019).