Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Meta-Learning-Based Adaptive Hyperparameter Tuning for Nonlinear Compensation in Optical Transmission

Lei Deng

denglei_hust@mail.hust.edu.cn

Huazhong University of Science and Technology https://orcid.org/0000-0003-2647-4610

Zicai Cao

Huazhong University of Science and Technology

Ziheng Zhang

Huazhong University of Science and Technology

Xiangyu Xie

Huazhong University of Science and Technology

Ming Tang

Huazhong University of Science and Technology https://orcid.org/0000-0001-8669-4186

Liu Deming

Huazhong University of Science and Technology

Mengfan Cheng

Huazhong University of Science & Technology

Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10362749/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is no conflict of interest

Meta-Learning-Based Adaptive Hyperparameter Tuning for Nonlinear Compensation in Optical Transmission

ZICAI CAO[1], ZIHENG ZHANG[1], XIANGYU XIE[1], MING TANG[1], DEMING LIU[1,2], MENGFAN CHENG[1,2,*], AND LEI DENG[1,2,*]

1 School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan, 430074, China

2 Shenzhen R&D Center of Huazhong University of Science and Technology, Shenzhen, 518000, China

**Corresponding author E-mail address: chengmf@hust.edu.cn or denglei_hust@mail.hust.edu.cn*

Abstract

To break the nonlinear Shannon capacity bottleneck, next-generation coherent systems are widely adopting probabilistic constellation shaping (PCS) and ultra-wideband multiplexing. However, these critically exacerbate nonlinear impairments. While neural network-based nonlinear equalizers (NNLEs) offer a promising mitigation path, their practical deployment is fundamentally crippled by the prohibitive manual overhead and fragile convergence of hyperparameter tuning (HT). We introduce a meta-learning-based adaptive HT framework (MAHT) that breaks this barrier by establishing a differentiable co-optimization across training dynamics and data rebalancing, enabling NNLEs to self-adapt without human intervention. We experimentally demonstrate an S+C+L-band DP-PCS-128QAM transmission over 150 km fiber, attaining an achievable information rate (AIR) of 215.35 Tbps. Moreover, MAHT delivers exceptional weight efficiency, outperforming conventional methods by nearly 150%, and automatically optimizes across multiple channels. This work establishes a self-optimizing, resource-efficient paradigm that bridges machine learning with practical high-speed optical networks, paving the way for truly autonomous transceiver systems.

Introduction

Fueled by the explosive growth of artificial intelligence and large-scale cloud computing, global data traffic is undergoing an unprecedented surge, placing immense pressure on the optical backbone infrastructure that interconnects data centers and forms the internet's physical foundation [1]. To sustain this relentless capacity scaling, next-generation coherent transmission systems are widely adopting probabilistic constellation shaping (PCS) and ultra-wideband multiplexing. These advanced formats, however, critically exacerbate fiber Kerr nonlinearity and hardware impairments, distortions that are inherently time-varying, multidimensional, and defy conventional analytical models [2]. A wide range of critical optoelectronic components in coherent systems can act as sources of nonlinear impairments, including digital-to-analog converters (DACs), electrical power amplifiers (PAs), optical modulators, and optical fiber, degrading signal quality and limiting achievable data rates [3]. The data-fitting nonlinear equalizers (NLEs) [4-12] at the receiver provide a solution for modeling and compensating for the combined effects of these impairments. For currently available digital signal processing (DSP) resources, neural network-based nonlinear equalizers (NNLEs) have emerged as the most compelling solution to mitigate these coupled impairments [13], as they can approximate arbitrary nonlinear functions without explicit higher-order term enumeration [14]. They exhibit broad applicability in probabilistic constellation shaping (PCS) systems [15,16], wavelength division multiplexing systems [17,18], low-resolution DACs [19], and complex fiber-induced non-linearity [20,21]. Furthermore, NNLEs demonstrate superior scalability, enabling integration with performance enhancement frameworks, including feature engineering [22] and perturbation terms mapping [23-25] to address scenario-specific demands.

Despite this promise, the practical deployment of NNLEs remains fundamentally obstructed by a persistent and often underestimated bottleneck: hyperparameter tuning (HT) [26]. This bottleneck is particularly acute for NNLEs. Constrained by available DSP resources, NNLEs are intentionally kept to minimal architectural scales. Multiple methods are employed to design efficient models, such as complex-valued NN [6,27], structured pruning [23-25], and Kolmogorov-Arnold networks [28,29]. Notable progress has also been made toward hardware implementations [30-33]. However, this very compactness makes their training outcomes exquisitely sensitive to hyperparameter choices, as any mis-specified learning rate or regularization factor disproportionately affects the few available degrees of freedom [37]. The challenge manifests at

three interconnected levels. At the data level, PCS signals exhibit a severely imbalanced radial energy distribution, where low-probability outer-ring symbols are vastly underrepresented, rendering conventional oversampling strategies either ineffective or noise-amplifying. At the training level, the coupled dynamics among learning rate, regularization, and initialization make gradient-based optimization highly prone to non-stationary convergence or local traps, a fragility that grid-search [34], staged [35,36], or decreasing learning rates [25] schedules consistently fail to address. At the system level, when scaled to multi-channel environments, this tuning burden multiplies prohibitively, as each channel's unique loss profile and amplifier condition demand per-channel re-tuning, leading to insurmountable manual effort and computational overhead (OH). Moreover, during single-channel NNLE training, the intricate coupling of multi-dimensional training hyperparameters, such as model architecture, learning rate, and regularization factor, makes the model training highly susceptible to convergence failure, training non-stationarity, or suboptimal solutions. A minimal model size heightens the criticality of the HT design for NNLEs.

Here, extending on our previous work [38], we introduce a meta-learning-based adaptive hyperparameter tuning (MAHT) framework to tackle these challenges. Rather than treating data balancing, learning rate scheduling, and regularization as independent knobs, MAHT establishes a differentiable linkage between them and the NNLE's performance, enabling gradient-based joint optimization. This is achieved through three synergistic mechanisms: a ring-decomposition resampling strategy with constrained oversampling that mitigates PCS radial energy imbalance; a context-aware dynamic scheduler that adjusts hyperparameters in real time in response to training states (loss value, epoch, and loss change rate), stabilizing convergence; and a multi-task meta-learning architecture that extracts generalizable strategies across channels, eliminating per-channel re-tuning overhead. To this end, we experimentally validate MAHT in a full S+C+L-band dual-polarization (DP)-PCS-128 quadrature amplitude modulation (QAM) transmission over 150 km of G654.E fiber, achieving a high AIR of 215.35 Tbps. The results demonstrate that MAHT not only pushes performance ceilings, outperforming conventional HT approaches by nearly 150% in weight efficiency, but also substantially reduces the engineering overhead of NNLE deployment, providing a self-adaptive and resource-efficient solution for next-generation ultra-wideband optical networks.

Results

In this work, the MAHT is designed to conduct a trilevel adaptive optimization of the learning rate, regularization factor, and PCS data balancing strategy. Figure 1 illustrates the methodology of the proposed MAHT.

Neural network-based NLE (NNLE). The NNLE used in this work is based on a standard feedforward NN (FFN) under complexity-constrained scenarios. The basic layer structure of the NN model is shown in Fig. 1 (b). In a typical NNLE architecture, considering that both inter-symbol interference (ISI) and nonlinear effects cause the current output to depend on past and future inputs, the operation objects are the received equalization target symbols (ETS) sequences composed of the ETS and their multiple adjacent symbols. Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_M\}$ represents the received sequences and $\boldsymbol{y} = \{y_1, y_2, \ldots, y_M\}$ represents the transmitted sequences. The ETS sequence sets $\mathcal{D}_E$ can be expressed as $\mathcal{D}_E = \{(s_i, y_i) | s_i = \{x_{i-T}, x_{i-T+1}, \ldots, x_i, \ldots, x_{i+T-1}, x_{i+T}\}\}$, where $i = T+1, T+2, \ldots, M-T$. Subsequently, as shown in Fig. 1 (b), the ETS sequence sets undergo I/Q separation and standard normalization to yield the sample sets $\mathcal{D}$. The forward output of each layer of an FFN model is described by [41]:

$$\boldsymbol{h}^{(l)} = \sigma(\boldsymbol{W}^{(l)}\boldsymbol{h}^{(l-1)} + b^{(l)}), \tag{1}$$

where $\boldsymbol{h}^{(0)} = v(s_i)$ is the input feature vector, $v(\cdot)$ represents the pre-processing into the NN, including normalization or other designed mapping. $\boldsymbol{W}^{(l)}, b^{(l)}$ are the weights and biases of the l-th layer, respectively. σ is the activation function for nonlinear mapping. The model training is mathematically framed as a parameter optimization to train a parametric predictive model $\theta^* = \underset{\theta}{argmin}\mathcal{L}(\mathcal{D}; \theta, \alpha, \eta)$, $\theta = \{\boldsymbol{W}^{(1)}, b^{(1)}, \ldots, \boldsymbol{W}^{(l)}, b^{(l)}\}$ represents all the trainable parameters, and α, η represents the untrainable learning rate and regularization factor. $\mathcal{L}(\cdot)$ represents the loss function to calculate the error between the model prediction and the labels. The loss for each training iteration is described as:

$$\mathcal{L}_B(\theta) = \frac{1}{N_b}\sum_{i\in\mathcal{D}_B}\left[\mathcal{F}(\boldsymbol{s}_i; \theta) - y_i\right]^2, \tag{2}$$

$\mathcal{F}(\boldsymbol{s}_i; \theta)$ denotes the prediction of the model parameterized by θ for an input $\boldsymbol{s}_i$. $\mathcal{D}_B \in \mathcal{D}$ is a mini-batch with a batch size of N_b. The θ is iteratively optimized through gradient backpropagation (BP) as:

$$\theta \leftarrow \theta - \alpha\frac{2}{N_b}\sum_{i\in\mathcal{D}_B}\left(\mathcal{F}(\boldsymbol{s}_i, \theta) - y_i\right)\nabla_\theta\mathcal{F}(\boldsymbol{s}_i, \theta), \tag{3}$$

where α is the learning rate. The total sample set $\mathcal{D}$ is divided into a training set $\mathcal{D}^{train}$ for the gradient descent process, a validation set $\mathcal{D}^{val}$ for model performance evaluation after each iteration, and a test set $\mathcal{D}^{test}$. In the inference stage, $\mathcal{D}^{test}$ is used as an estimation of the generalization error of the model. It can be expressed as:

$$\mathbb{E}\left[\mathcal{L}(\mathcal{F}_\theta(\boldsymbol{s}), y)\right] \approx \frac{1}{N^{test}}\sum_{k\in\mathcal{D}^{test}}\mathcal{L}(\mathcal{F}(\boldsymbol{s}_k), y_k). \tag{4}$$

Meta-learning-based adaptive HT framework (MAHT). The selection of the hyperparameters can greatly affect the learning process and the performance metrics of the trained model. Conventional NNLEs are specific to a single task and dataset with no specific HT methods, which means the hyperparameters are always pre-specified. In MAHT, adaptive hyperparameter optimization is designed for learning across multiple tasks. For each task, the dataset is divided into support datasets $\mathcal{D}_{sup}^{(N)}(S)$ and query datasets $\mathcal{D}_{que}^{(N)}(Q)$. $\mathcal{D}_{sup}^{(N)}(S)$ are used to train the NNLE corresponding to each task, while $\mathcal{D}_{que}^{(N)}(Q)$ are used to evaluate the meta-learning performance under each task and compute the meta gradient [38]. After preparing the datasets, MAHT implements a simultaneous

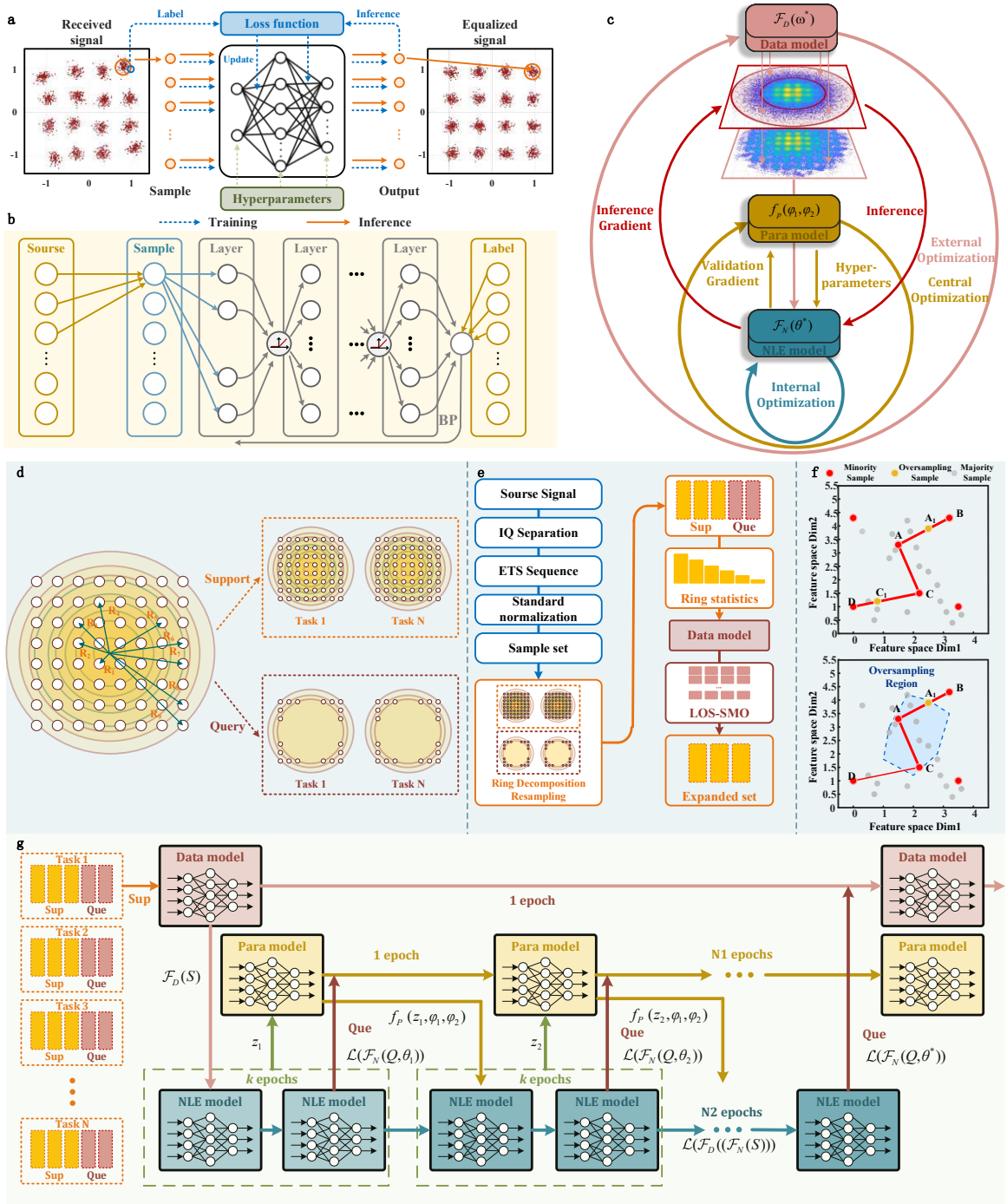


Fig. 1 Schematic of the proposed MAHT. a The general equalization process of the neural network-based NLE. **b** The basis layer structure of the NN model in the proposed MAHT. **c** Principle of the trilevel optimization. **d** Principle of ring decomposition resampling. **e** Data processing workflow from the source signals to the Sup/Que dataset into the NN model. **f** Principle of limited oversampling space synthetic minority oversampling. **g** Detailed data flow of trilevel optimization dissected from the epoch perspective.

optimization of the learning rate, regularization factor, and PCS data balancing strategy through a trilevel optimization. It consists of three models: the internal NNLE model $\mathcal{F}_N(\theta)$, the central para model $f_p(\varphi_1, \varphi_2)$, and the external data model $\mathcal{F}_D(\omega)$, where $f_p(\varphi_1, \varphi_2)$ and $\mathcal{F}_D(\omega)$ are standard feedforward models with extremely low parameters of about 1000 multipliers and accumulation (MACs). In MAHT, $\mathcal{F}_N(\theta)$ is trained under the datasets expanded by $\mathcal{F}_D(\omega)$ and hyperparameters from $f_p(\varphi_1, \varphi_2)$. Then the NNLE optimization is extended as:

$$\theta^* = \underset{\theta}{argmin}\mathcal{L}(\mathcal{D}_{ext}; \theta; \alpha_t, \eta_t), \tag{5}$$

where $\mathcal{D}_{\text{ext}}$ is the expanded dataset, α_t, η_t is the provided current training hyperparameters. In MAHT, the purpose of the hyperparameter parameterization is to convert traditional

discrete hyperparameters into differentiable ones, thereby supporting end-to-end optimization based on gradients. Through parameterization, the hyperparameters are transformed into learnable variables, enabling them to automatically adjust and adapt to the dynamic requirements of different training stages or data distributions through first-order gradient descent. The input to the meta models is specially constructed datasets, as shown in Fig. 1 (e). Under the strategies from $\mathcal{F}_D(\omega)$, the support datasets are expanded and are used for a more complete training of the NNLE model. $f_p(\varphi_1, \varphi_2)$ operates within the internal optimization process of the NNLE model. α_t, η_t are the outputs of two separate models of $f_p(\varphi_1, \varphi_2)$, parameterized by φ_1 and φ_2, respectively:

$$\alpha_t = f_p(z_t, \varphi_1), \quad \eta_t = f_p(z_t, \varphi_2), \; z_t = \left[\nabla \mathscr{L}_t, \; t, \; \mathscr{L}_t\right], \tag{6}$$

where z_t is the context feature at the current training epoch, including the loss value $\mathscr{L}_t$, training epoch t, and loss change rate $\nabla \mathscr{L}_t$. Here, the trilevel optimization problem is derived from the bilevel optimization, i.e., one optimization serves as a constraint for another optimization [40]. As shown in Fig. 1 (c), the designed trilevel optimization is described as:

$$\begin{aligned} \varphi_{1,2}^{\;*} &= \underset{\varphi_{1,2}}{\operatorname{argmin}} \mathcal{L}(\mathcal{F}_N(\mathcal{D}_{que}^{(N)}; \theta^{k_m}; \varphi_{1,2})) \\ \omega^* &= \underset{\omega}{\operatorname{argmin}} \mathcal{L}(\mathcal{F}_N(\mathcal{D}_{que}^{(N)}; \theta^*; \omega)) \\ s.t. \;\; \theta^* &= \underset{\theta}{\operatorname{argmin}} \mathcal{L}\left(\mathcal{F}_D(\mathcal{D}_{sup}^{(N)}); \theta; f_{\varphi_1}(z_t), f_{\varphi_2}(z_t)\right), \end{aligned} \tag{7}$$

where θ^{k_m} indicates the optimal θ for every k_m epochs of NNLE training. Detailed data flow of the trilevel optimization dissected from an epoch perspective is shown in Fig. 1 (g). Internal optimization is the training process of $\mathcal{F}_N(\theta)$ under fixed meta model conditions. External optimization is the training of $\mathcal{F}_D(\omega)$, whose loss comes from the performance of the NNLE model evaluated on $\mathcal{D}_{que}^{(N)}(Q)$, and the parameters are updated after every k_m epochs of the NNLE training. Before the $f_p(\varphi_1, \varphi_2)$ being fully converged, the dynamic hyperparameters they provide may not be ideal, which limits the training efficiency of the NNLE model in the early training. Therefore, the parameter update steps of the two models are set differently from the hyperparameters provided steps in the trilevel optimization process. When θ is updated every k epochs, $f_p(\varphi_1, \varphi_2)$ receives z_t as input and provides new α_t, η_t. φ_1, φ_2 are updated after every k_m epochs of training. During early training, $k_m \gg k$, this setup can prevent the suboptimal hyperparameters provided by $f_p(\varphi_1, \varphi_2)$ from degrading the early training of $\mathcal{F}_N(\theta)$. In addition, $f_p(\varphi_1, \varphi_2)$ are pre-trained using a small-scale dataset sampled from the complete dataset. At the same time, a sequence of reasonably performing hyperparameters $P = (\alpha_t^{sa}, \eta_t^{sa})$ is collected using conventional methods to enable the model to mimic the initial strategy; this is described as:

$$\varphi_{1,2}^{\;*} = \underset{\varphi_{1,2}}{argmin} \sum_t \left\| f_p\left(z_t, \varphi_{1,2}\right) - P \right\|^2. \tag{8}$$

After the optimization is completed, the parameters of the meta models are fixed. However, they can still generate dynamic hyperparameters based on the input state. While recent works apply meta-learning to find global initializations for rapid cross-channel transfer learning [42], our MAHT framework fundamentally differs. MAHT utilizes meta-learning to dynamically co-optimize intra-task hyperparameter tuning and PCS-specific data rebalancing within the active training loop, ensuring stable convergence of ultra-compact NNLEs.

Ring Decomposition Resampling (RDR). To ensure that the MAHT framework maintains task diversity during the meta-training phase without deviating from the physical nature of the PCS signal, an energy-ring-based category reconstruction mechanism is introduced in the MAHT framework, as shown in Fig. 1 (d), based on the fact that the PCS signal follows a strict radial energy distribution. A resampling strategy based on the physical parameter space is employed to construct a multi-task support set. By resampling the received sequences across different channels, this strategy constructs a series of meta-tasks that follow different shaping factors. This mechanism ensures that all empirical distribution features in the meta-model input adhere to the Maxwell-Boltzmann distribution of the PCS. The original sample is separated into multiple energy rings based on the Euclidean length from each constellation point to the origin as:

$$C = \{R_1, R_2, \dots, R_K\}. \tag{9}$$

Using ring distributions to characterize data distributions as state inputs for a meta-model can significantly reduce the input and output dimensions of the meta-model, as well as the dimension of the hyperparameter search space. When the scaling factor is α_p, the theoretical probability of occurrence of the kth ring is:

$$P(R_k; \alpha_p) = \frac{M_k}{\sum_{j=1}^{K} M_j e^{-\alpha_p E_j}} e^{-\alpha_p E_k}, \tag{10}$$

where M_k is the number of constellation points contained in the ring, and E_k is the energy of the ring. To transform the existing data distribution into the target mission distribution, it is necessary to calculate the acceptance probability weight for each ring W_k:

$$W_k = \frac{P(R_k;\alpha_1)}{P(R_k;\alpha_0)} = \frac{\sum_{j=1}^{K} M_j e^{-\alpha_0 E_j}}{\sum_{j=1}^{K} M_j e^{-\alpha_1 E_j}} \cdot e^{(-(\alpha_1 - \alpha_0)E_k)}. \tag{11}$$

Based on W_k, each ring of the sample set can be resampled to construct meta-training support sets $\mathcal{D}_{sup}^{(N)}(\mathcal{T}_i)$ for different tasks. In contrast, $\mathcal{D}_{que}^{(N)}(\mathcal{T}_i)$ is constructed by sampling only from the three outer energy rings that are mutually exclusive with $\mathcal{D}_{sup}^{(N)}(\mathcal{T}_i)$ to characterize severe nonlinearity. The specific sample set workflow is shown in Fig. 1 (e). Before constructing $\mathcal{D}_{sup}^{(N)}(\mathcal{T}_i)$, the original received sequence must undergo I/Q separation, ETS sequence generation, and sample standardization and normalization to generate the raw sample set. Subsequently, the meta-training support set and query set are constructed. For the input to $\mathcal{F}_D(\omega)$, statistical processing is first performed on the samples from each ring of $\mathcal{D}_{sup}^{(N)}(\mathcal{T}_i)$, generating a ring distribution feature vector specific to that task:

$$\boldsymbol{H}_i = [n_1, n_2, \dots, n_K; \Delta_1, \Delta_2, \dots, \Delta_K]. \tag{12}$$

Here, n_K represents the proportion of samples in each ring, reflecting the task characteristics, and $\Delta_k = \frac{\sum_{i \in R_k} \|y_i - x_i\|^2}{|R_k|}$ represents the damage characteristics of each ring. Subsequently, an improved PCS oversampling strategy is developed from $\mathcal{F}_D(\omega)$ based on $\boldsymbol{H}_i$, which includes data expansion weights $S_i = [S_{i,1}, S_{i,2}, \ldots, S_{i,K}]$ for each ring. Under this strategy, the class imbalance in the $\mathcal{D}_{sup}^{(N)}(\mathcal{T}_i)$ is mitigated to obtain an expanded support dataset, which is used to train the NNLE model more effectively.

Limited oversampling space-synthetic minority oversampling (LOS-SMO). Synthetic minority oversampling (SMO) is a more reasonable approach for communication than random oversampling. It synthesizes new samples by performing linear interpolation in the feature space between minority class samples, thereby defining a smoother decision region for the outer-ring minority class samples and mitigating the risk of overfitting. The basic logic of SMO is that samples belonging to the same class typically exhibit local similarity in the feature space. Therefore, by generating new samples along the line connecting two adjacent minority-class samples, the distribution characteristics of that class can be simulated, rather than merely replicating existing data. Assuming the minority class is denoted by C, for each sample $\boldsymbol{s}_i$, calculate the Euclidean distance in the feature space from $\boldsymbol{s}_i$ to all other samples in C, select the k nearest neighbors, randomly and uniformly choose one of these neighbors $\boldsymbol{s}_i^*$, and perform linear interpolation along the line segment: $\boldsymbol{s}_{SMOTE} = \boldsymbol{s}_i + \beta \cdot (\boldsymbol{s}_i^* - \boldsymbol{s}_i)$, where β is a random number uniformly distributed between 0 and 1, and the new sample is randomly distributed between the two sampled data points, forming a continuous categorical space. However, SMO may also amplify the influence of noisy samples or outliers during the oversampling. Therefore, this paper imposes restrictions on the sampling space of the SMO, called LOS-SMO. The diagrams of SMO and LOS-SMO are illustrated in Fig. 1 (f). For visualization, the high-dimensional

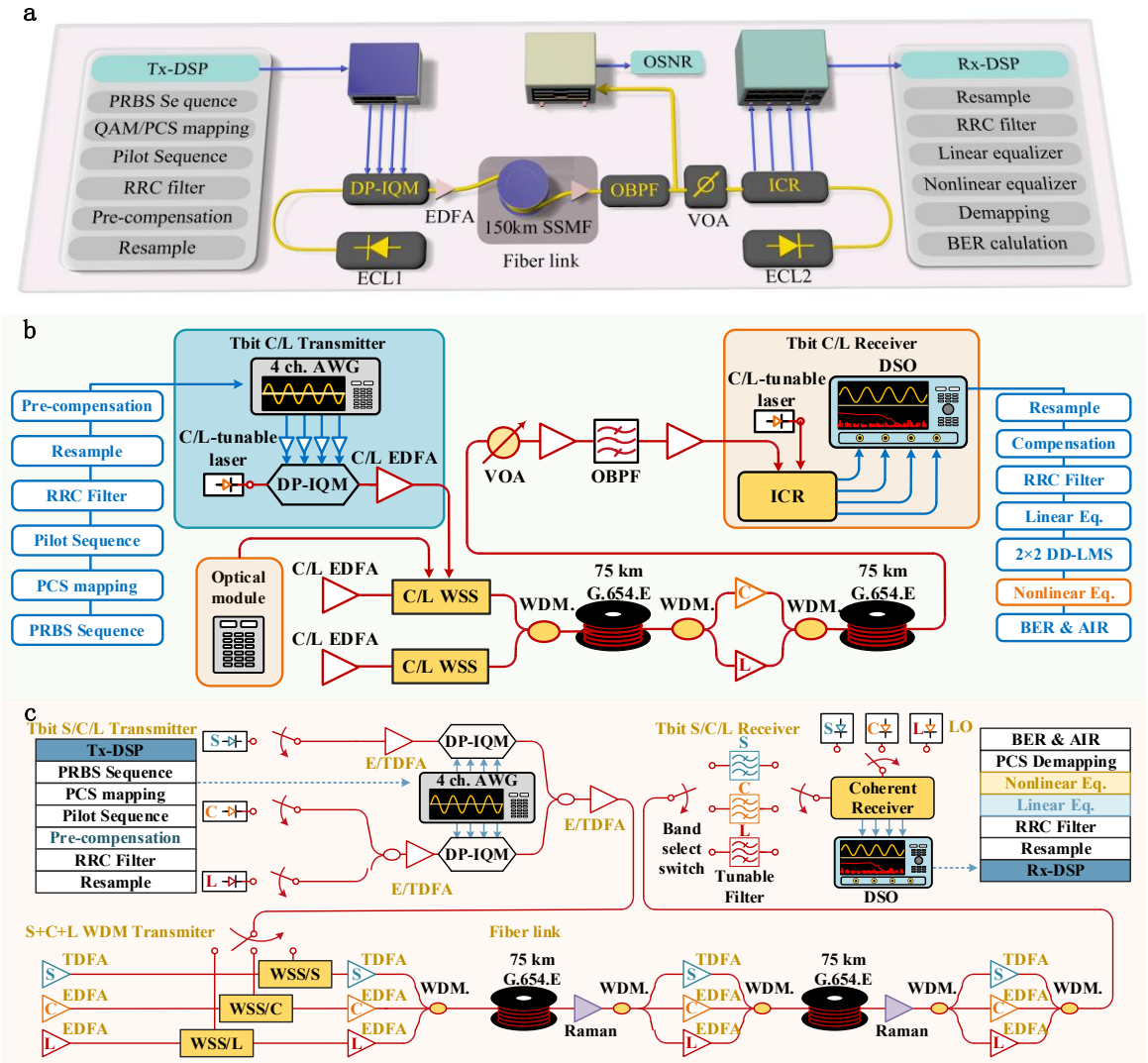


Fig. 2 Implementation scheme of the MAHT framework in coherent optical communication. a Experimental setup of the 1 Tbps DP-PCS 128 QAM BtB transmission. **b** Experimental setup of the 1 Tbps/channel 96 channels C6T+L6T-band PCS128 QAM 150 km G654.E transmission. **c** Experimental setup of the S+C+L-band PCS128 QAM coherent optical WDM transmission over 150 km G654.E.

features are reduced to two dimensions. LOS-SMO lies in the precise identification and restriction of the sampling space. Since oversampling occurs only during model training, prior information can be utilized to make decisions prior to oversampling, ensuring oversampling takes place only within the core region. After obtaining the k nearest neighbors of a sampling point, the proportion of neighbors belonging to the sampling class versus other classes is analyzed to minimize erroneous oversampling of outliers or samples near decision boundaries. Nevertheless, due to the stark disparity in the distribution of PCS signals between the inner and outer rings, pursuing absolute class balance may introduce significant additional training cost (ATC). ATC is defined here as the additional amount of training data generated by oversampling. Practically, it is necessary to balance the ATC against the performance gains.

Implementation scheme

The performance of the MAHT framework is experimentally demonstrated through a comparison, including conventional NNLE applied with manual HT, NNLE with SMO and hyperparameter grid search (GS), and NNLE with the MAHT. The experimental scenarios include:

(1) a 1 Tbps DP-PCS 128 QAM back-to-back (BtB) transmission to evaluate performance and weight efficiency in a capacity-limited coherent optical transmission scenario where higher-order PCS formats are deployed.

(2) a 1 Tbps/channel 96 channels C6T+L6T-band PCS128 QAM 150 km G654.E transmission to evaluate performance in multi-wavelength wavelength division multiplexing (WDM) transmission scenarios.

(3) a S+C+L-band PCS128 QAM (H=6.6) coherent optical transmission over 150 km G654.E to evaluate the ultimate achievable information rate (AIR) gain in WDM transmission scenarios.

The experimental setup of the three tests, which includes the DSP chains of the transmitter and receiver as well as link equipment, is shown in Fig. 2, and comprehensive details are provided in the Materials and Methods section. Offline DSP procedures, including resampling, linear equalization, nonlinear equalization with the tested NLEs, QAM de-mapping, and bit error rate (BER) calculation, are performed as part of the experimental setup. As this work discusses hyperparameter optimization under limited DSP resources, detailed optimization of the NNLE layer structure is not carried out. Instead, the performance of hyperparameter optimization

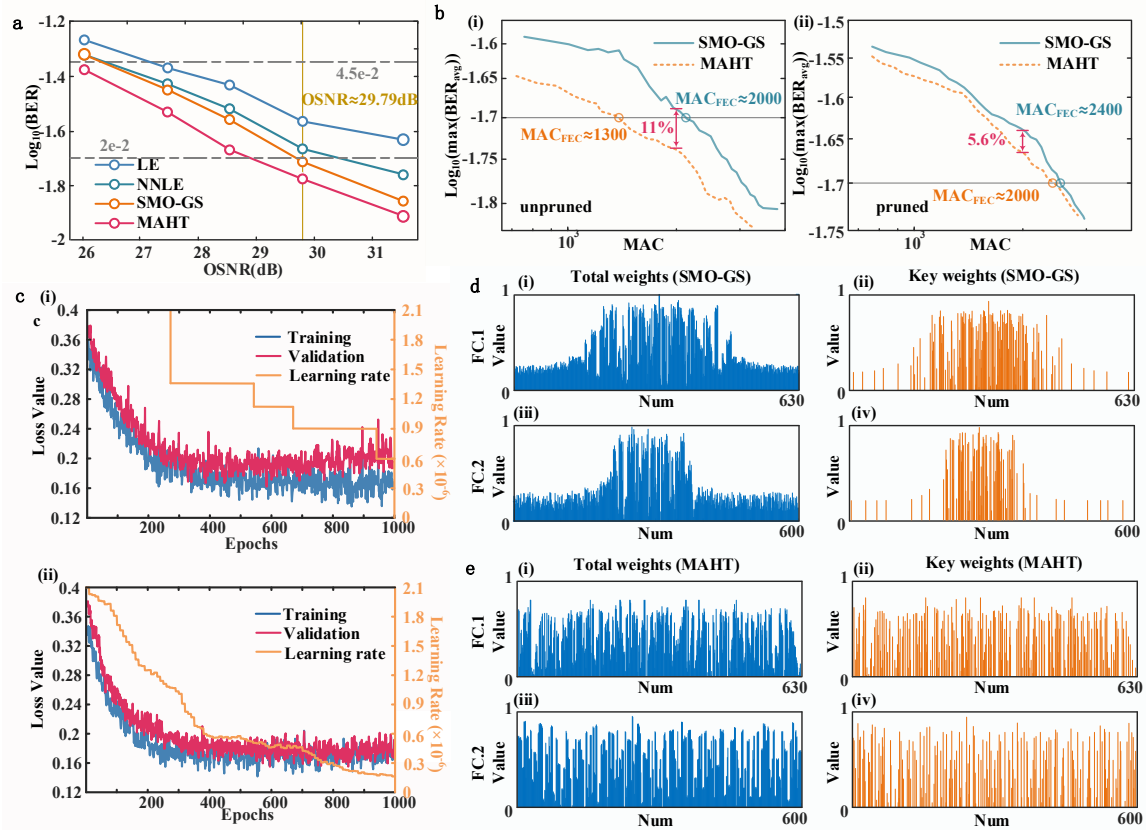


Fig. 3 Experimental results of DP-PCS 128 QAM BtB transmission. a The log10 (BER) versus OSNR curves. **b** The log10(max(BERavg)) versus MAC curves of unpruned **(i)** and pruned **(ii)** NNLE(SMO-GS) and NNLE(MAHT). **c** The loss value and adaptive learning rate versus training epochs curves of the NNLE (SMO-GS) training **(i)** and NNLE **(MAHT)** training**(ii). d** The weight distribution of NNLE (SMO-GS) before **(i)** and after **(ii)** weight cluster of the first hidden layer; **(iii) (iv)** of the second hidden layer. **e** The weight distribution of NNLE (MAHT) before **(i)** and after **(ii)** weight clustering of the first hidden layer; **(iii) (iv)** of the second hidden layer.

methods on NNLE architectures with representative sizes available in the targeted scenario is selected. Here, the NNLE architecture is configured as a 2-layer FFN with 21 input dimensions, 30, 20 hidden neurons, and 1 output dimension for the equalized symbol, with about 1200 required MACs per symbol. For the database collection, approximately 70,000 symbols of each signal frame are transmitted. The overall attenuation of the signal is increased by a wave-shaper in the transmission link to adjust the value of the received signal optical signal-to-noise ratio (OSNR). The BER of the system for different OSNR conditions is scanned until it exceeds the forward error correction (FEC) threshold.

DP-PCS 128 QAM BtB transmission

The equalization performance of each tested NNLE is quantified by measuring the pre-FEC BER while the system OSNR varies. Fig. 3 (a) shows the log10 (BER) versus OSNR curves of the 1 Tbps DP-PCS 128 QAM BtB transmission. It shows that NNLE (SMO-GS) exhibits nearly identical performance to manually tuning the NNLE at low OSNR. As SMO does not impose constraints on the oversampling space, it inevitably introduces noise into the synthesized samples when attempting to reconstruct the low-probability outer-ring distribution of the PCS signal. This effect is amplified in low OSNR scenarios. Although GS approaches the performance of MAHT in specific scenarios, its high computational cost still limits its practical application. In contrast, the MAHT not only achieves performance on par with SMO-GS in high OSNR regions but also demonstrates significantly better robustness against oversampling noise than SMO-GS due to its optimized oversampling strategy and dynamic learning rate, thereby exhibiting a higher and more stable performance ceiling than SMO-GS.

Figure 3 (c) shows the loss value versus training epochs curves of the NNLE (SMO-GS) and the NNLE (MAHT). It is evident that MAHT can adaptively provide finely tuned learning rates for each training stage, enabling more stable convergence of NNLE training. In contrast, the gradient learning rate lacks sufficient resolution and cannot adapt to training dynamics. Although both results start with the same training loss, the loss curve for SMO-GS exhibits significantly more volatility in later training, indicating gradual overfitting of the model. In contrast, the loss curve under the MAHT is smoother.

Furthermore, the effectiveness of the final weights of the two models can be evaluated by examining the performance of the trained model before and after model pruning and comparing it with the MAC curves of the scanning model. Model pruning simplifies the model topology by removing weights with low contribution. Here, global unstructured pruning is employed, where weights from all layers are pooled into a unified pooling layer for sorting, and a pruning threshold of 20% is determined based on the global distribution. Due to the dense scanning step size, to minimize testing error, the average BER from the last 10 epochs with the best performance in 5 total training runs is selected as the metric. Fig. 3 (b) shows the $\log_{10(\max(\mathrm{BER_{avg}}))}$ versus MAC curves of unpruned (i) and pruned (ii) NNLE(SMO-GS) and NNLE(MAHT). The MAC where the BER metrics of the models reach the 2×10^{-2} FEC threshold is denoted as MACFEC. In Fig. 3(b)(i), there is a performance gap between MAHT and SMO-GS before pruning across the scanned MAC values, as previously demonstrated in the OSNR scan. Specifically, SMO-GS achieves a MACFEC of approximately 2,000, while MAHT requires only about 1,300. Taking MAC=2,000 as a representative point, the estimated performance gap between the two models is approximately 11%. Comparing the impact of pruning on the two models, both models exhibit varying degrees of performance degradation after pruning. SMO-GS degrades by approximately 11.8%, with its MACFEC increasing to about 2000. Meanwhile, MAHT degrades by approximately 17%, resulting in performance comparable to that of SMO-GS. The performance gap at MAC=2000 is reduced to approximately 5.6%, and MACFEC needs to be increased to 2400 to compensate for the weights lost. Therefore, pruning damages the performance of MAHT far more than SMO-GS. This indicates that the proportion of effective weights in MAHT is significantly higher than that of SMO-GS. The results indicate that MAHT is more efficient than SMO-GS in terms of weight optimization and exhibits greater weight effectiveness.

To quantify the weight effectiveness, a weight cluster (WC) is used as an alternative method. WC reduces the diversity of model weights, identifies and extracts key weights within the model, thereby revealing the redundancy of the model. Based on the k-means algorithm, WC divides the continuously distributed weights in each hidden layer of the model into multiple discrete categories. WC aims to minimize the sum of intra-class squared errors. Upon completion, all original weights belonging to the same category are replaced with the corresponding cluster center value. Inspired by the standard network pruning method where weight effectiveness is evaluated via performance degradation after parameter removal [43], we tailor this concept for optical communications:

$$Weight\ efficiency = \frac{N_{key}}{N_{total}}\left(1 + log_{10}\left(\frac{BER_{unpruned}}{BER_{pruned}}\right)\right), \tag{13}$$

it quantifies weight efficiency using the pre- and post-pruning BER discrepancy, where a steeper BER degradation indicates lower structural redundancy and higher key weight utilization. Fig. 3 (d) and (e) show the weight distribution of the two models before and after weight clustering of the hidden layers. The results in Fig. 3 (d)(i) and (iii) show that the weight distribution of SMO-GS exhibits extreme polarization: a small number of large weights are highly concentrated, while small weights are diffusely distributed. In Fig. 3 (d)(ii) and (iv), after WC, the weight distribution is concentrated on specific key weights. According to (13), the estimated weight effectiveness of SMO-GS is only approximately 25.9%. This indicates that the model suffers from severe parameter redundancy, with most weights remaining inactive and failing to contribute effectively to performance. In contrast, Fig. 3 (e) shows that the weight distribution of MAHT remains highly stable. Both before and after WC, the weights are distributed relatively uniformly, with no obvious tendency toward zero values or isolated polarization. The estimated weight effectiveness reaches as high as 63%, meaning that the majority of weights act as key weights and

significantly contribute to performance. This weight distribution also confirms that the severe performance degradation observed in MAHT after pruning is due to the loss of key weights. The results indicate that MAHT significantly enhances the weight effectiveness. This enables MAHT to achieve more accurate and robust balancing capabilities than traditional architectures, even within a more lightweight structure.

PCS-WDM transmission

In the 96 channels C6T+L6T-band PCS128 QAM 150 km G654.E transmission, the two edge channels of the C6T band (1524.765 nm and 1571.745 nm) and the middle channel (1547.425 nm), the two edge channels of the L6T band (1574.815 nm and 1624.99 nm) and the middle channel (1599.005 nm) are selected for under test channel (UTC). The total capacity of the 96 channels in the whole band is estimated based on the average performance of the tested channels. Transmission is considered successful when the output channel BER is less than 4.5×10^{-2}. Based on this, the maximum transmission rate for the channel is measured. Transmission tests are conducted sequentially for the 6 wavelengths. The wavelength-dependent fiber loss curves over a 75-km span, along with the test wavelengths and corresponding loss values are shown in Fig. 4 (a).

Figure 4 (b) shows the constellation diagram, BER and AIR of the X-polarization signals for the 6 channels in the C6T+L6T band. As the equalization of NNLE compresses the constellation diagram and generates a non-Gaussian noise distribution, the AIR estimation based on a Gaussian channel may be slightly off. Consequently, signal amplitude normalization based on the average power of the PCS signal is applied to the results after NNLE equalization to obtain a more accurate AIR estimation. Based on the measured average performance of the 6 channels, the total capacity across all 96 channels of the C6T+L6T band is conducted. Through the combined effects of link loss

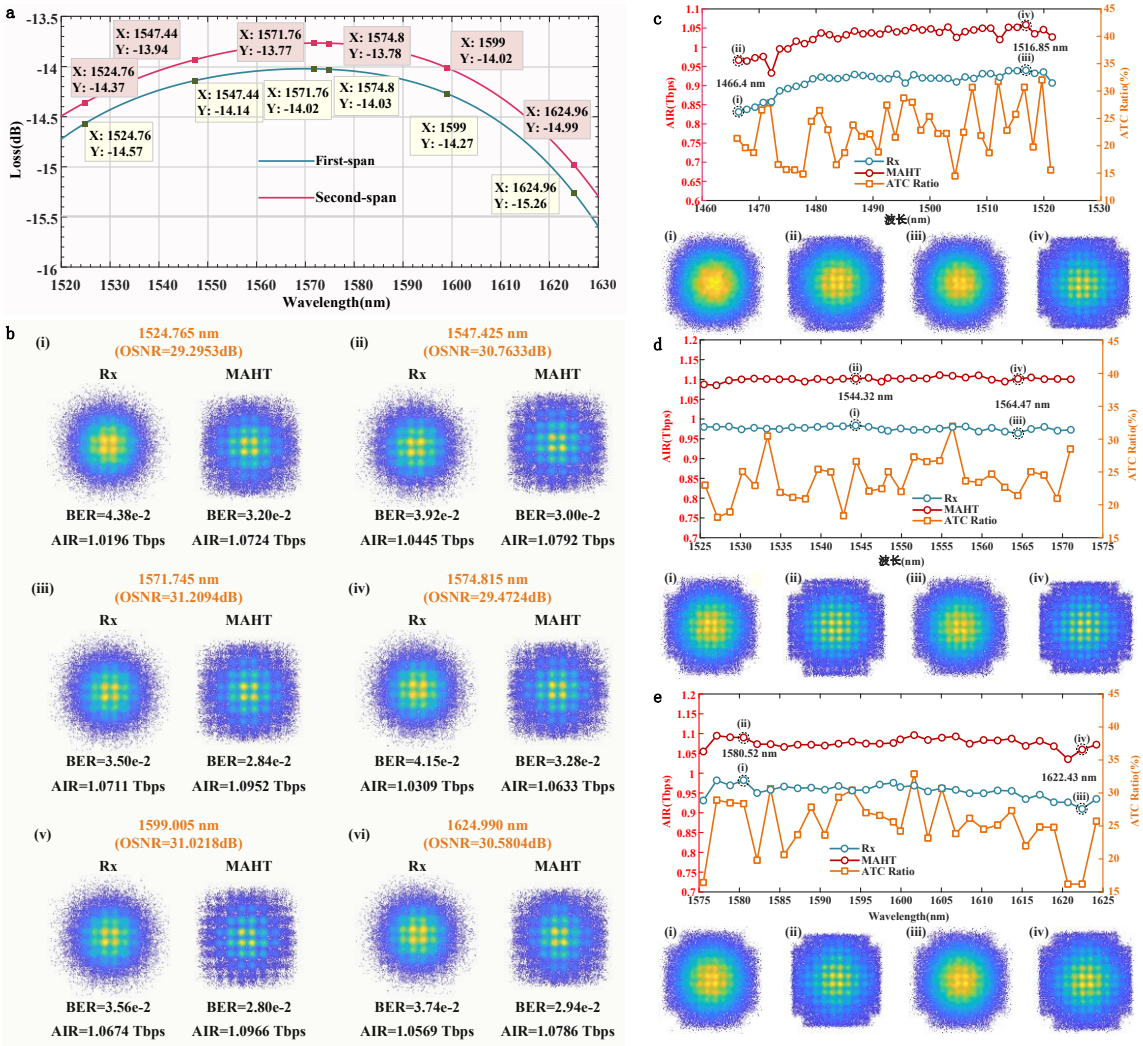


Fig. 4 Experimental results of DP-PCS 128 QAM WDM transmission. a wavelength-dependent fiber loss curves over a 75-km span, along with the test wavelengths and corresponding loss values in C6T+L6T transmission. **b** Constellation diagrams, BER and AIR of X-polarization of the tested 6 channels. **c** ATC ratio and AIR curves of MAHT across S-band channels with constellation diagrams in the inset figure. **d** ATC ratio and AIR curves of MAHT across C-band channels. **e** ATC ratio and AIR curves of MAHT across L-band channels.

compensation, a linear equalizer (LE), and the MAHT-based NNLE, the AIR for each UTC ultimately exceeded 1 Tbps. Furthermore, the MAHT provided substantial gains, delivering an AIR of approximately 52.8 Gbps on the channel with the lowest OSNR (Fig. 4 (b)(i)) and approximately 24.1 Gbps on the channel with the highest OSNR (Fig. 4 (b)(iii)). The results demonstrate that the MAHT-based NNLE effectively addresses channel non-uniformity in ultra-wideband coherent transmission systems.

In the S+C+L-band PCS128 QAM coherent optical transmission over 150 km G654.E, 197 WDM UTCs is conducted with center wavelengths in the S-band (1464.96–1521.47 nm; 77 wavelengths), C-band (1523.90–1571.03 nm; 60 wavelengths), and L-band (1573.87–1624.19 nm; 60 wavelengths). Only 98 of these UTCs (spaced 100 GHz) are used for performance estimation, and the total AIR for each band is calculated by multiplying the average AIR of the UTCs in each band by the number of channels.

Figures 4 (c), (d), and (e) show the AIR and ATC results for each UTC in the S-band (38 channels), C-band (30 channels), and L-band (30 channels) with and without NNLE. The ATC ratio is achieved by the LOS-SMO strategy to fully random oversampling. It demonstrates the additional complexity reduction benefits provided by the MAHT. The specific points marked on the curves are selected as the channels with the best and worst AIR performance in the S, C, and L bands before equalization, and the inset figures show the constellation diagrams of these channels before and after equalization in the MAHT. The results show that the ATC ratio for nearly all channels in the MAHT is below 35%, and its total ATC accounts for only 23.98% of that of fully random oversampling. Furthermore, under various DSP compensations, the total AIR reached 208.61 Tbps, with the MAHT achieving an AIR gain of 23.58 Tbps.

Discussion

In this work, we proposed the MAHT framework to fundamentally dismantle the hyperparameter tuning bottlenecks of NNLEs in ultra-high-capacity optical transmission. By holistically co-optimizing PCS-induced data imbalance, training dynamics, and channel-specific hyperparameters, MAHT completely obviates prohibitive manual tuning costs while guaranteeing robust convergence and unlocking unprecedented equalization performance. Shattering existing performance barriers, our experimental demonstration of a 215.35 Tbps AIR over a 150-km G654.E fiber in an S+C+L-band system establishes a rigorous new benchmark for co-optimization in wideband WDM scenarios.

Compared to conventional methods like grid search combined with minority oversampling, MAHT demonstrates significant architectural superiority through its trilevel optimization. At the data level, a physics-prior resampling strategy effectively mitigates the spatial energy imbalance of PCS signals, while a context-aware dynamic scheduler ensures stable gradient optimization during training. Consequently, MAHT improves weight efficiency by nearly 150%, drastically reducing parameter redundancy and computational overhead to pave the way for low-latency hardware deployment. Furthermore, its multi-task meta-learning architecture extracts generalizable tuning strategies, entirely eliminating per-channel retuning costs and exhibiting seamless adaptation across diverse WDM channel conditions.

Beyond the WDM-PCS fiber communication scenarios validated here, next-generation systems will inevitably evolve toward complex higher-order modulation formats and multi-dimensional multiplexing. Given that the core mechanisms of MAHT, comprising distribution-aware resampling, dynamic scheduling, and meta-learning, operate independently of specific optical fiber physical models, this dimension-agnostic paradigm is universally applicable to advanced architectures such as multi-core or few-mode spatial division multiplexing. Moreover, it holds strong potential for time-varying channel equalization such as free-space or underwater optical communications. By fusing this meta-learning engine with specialized model topologies, such as complex-valued or Kolmogorov-Arnold networks, MAHT establishes a unified foundation for truly autonomous, self-optimizing transceiver systems in future optical networks.

Materials and methods

Experimental details of BtB transmission

The experimental setup of the 118 GBaud DP-PCS-128 QAM BtB transmission is illustrated in Fig. 2 (a). At Tx, symbol sequences of the signals are generated in a MATLAB program. A pilot sequence with a length of 2000 symbols is added for synchronization on the Rx side. Root-raised cosine (RRC) filters with a roll-off factor of 0.01 at both the Tx and Rx sides are set up to jointly implement the effect of an RC filter. The offline Tx-DSP also includes the pre-compensation for linear impairments. After a resampling process is performed to match the sampling rate, the signals are uploaded to an arbitrary waveform generator 4-channel AWG (Keysight M8199A) with a 3 dB bandwidth of 60 GHz, operating at a sample rate (SR) of 256 GSa/s. Then, the signal is directed into the DP-I/Q modulator (DP-IQM, iXblue, MPZ-LN-40) with a 3 dB bandwidth of 40 GHz. Two external cavity lasers (ECL) with a 1551 nm wavelength and a 100 kHz linewidth are used for the signal optical carrier and local oscillator (LO). An Erbium-doped fiber amplifier (EDFA) is employed after the IQM to amplify the signal, while an optical bandpass filter is used to remove the amplified spontaneous emission (ASE) noise. A variable optical attenuator (VOA) is used to adjust the received optical power (ROP) before Rx. An optical spectrum analyzer (OSA) captures the signal for OSNR calculation. Note that the OSNR scans cannot be equally spaced due to manual adjustments of the VOA. An integrated coherent optical receiver (ICR, Neophotonics MicorICR Class 40) with a 3 dB bandwidth of 43 GHz is used for coherent detection. A digital oscilloscope (DSO, Keysight UXR0704A) with a 3 dB bandwidth of 70 GHz, operating at an SR of 256 GSa/s, is employed for signal digitization. The pilot OH of the LE is around 1.5% and the 24% OH soft-decision (SD)-FEC threshold of 4.5e-2 [44] is adopted in this test. The net bit rate can be calculated

as: 118 G×2×(5.6692-0.24/1.24×log2(128))/(1+0.015)=1.003 Tbps. Offline DSP procedures, including resampling, linear equalization, nonlinear equalization with the tested NLEs, demapping, and BER calculation, are performed as part of the experimental setup. It is noted that the pre-FEC BER is used as the metric of the NLEs' performance evaluation [36].

Experimental details of WDM transmission

The experimental platform of the C+L-band/S+C+L-band PCS 128 QAM coherent optical transmission over 150 km G654.E, including DSP chains of the transmitter and receiver and link equipment, is illustrated in Fig. 2 (b) and (c). The tunable lasers (TLs) are used for the optical carrier and local oscillator (LO) of the under-test channel (UTC) and two simulated neighboring channels. The channels are modulated by a self-developed Si-based dual-polarization IQ-modulator (DP-IQM) driven by a 4-channel 256 GSa/s arbitrary waveform generator (AWG, Keysight 8199A). They are then amplified by erbium/thulium-doped fiber amplifiers (TDFAs for S-band and EDFAs for C- and L-band), combined, and inserted into the notch in the dummy channels generated by the ASE noise using wavelength-selective switches (WSS, Finisar 16000A). In the fiber links, each span contains 75 km G654.E fiber, a pair of multiplexers and demultiplexers, a Raman amplifier, and three DFAs. At the receiver side, an optical tunable filter is used for separating the UTC from the dummy channels. An integer coherent receiver (ICR) is utilized for coherent detection, and a 256 GSa/s digital storage oscilloscope (DSO, Keysight UXR0594B) is used for digitization.

Communication DSP

In the Tx-DSP, the transmitted pseudo-random binary sequence (PRBS) is first input into the PCS encoding module to obtain a PCS 128QAM signal with the corresponding modulation entropy. This signal is then subjected to RRC matching filtering with a roll-off coefficient of 0.01 and resampling. The resampled signal undergoes static impairment pre-compensation, followed by peak clipping to improve the overall signal-to-noise ratio; the clipping parameter ranges from 0.99 to 0.999. In Rx-DSP, the received signal first passes through a static equalizer to compensate for channel delays and other fading effects. Following resampling and frequency offset estimation, an RRC filter with a roll-off coefficient of 0.01 is applied. A pilot sequence with a 1.5% overhead is incorporated into the algorithm architecture for the linear equalizer. To enhance distortion compensation, the dynamic linear equalizer is divided into two stages: the first stage is an 8×2 widely linear (WL)-multiple-input multiple-output (MIMO) dynamic equalizer based on least mean squares, which incorporates polarization demultiplexing and a phase-locked loop. At the output of the linear equalizer, a least mean squares (LMS)-based 2×2 MIMO equalizer is connected to each of the X and Y polarization signals to compensate for residual linear distortion. Finally, after passing through the tested NLE, QAM demapping and BER estimation are performed.

Acknowledgements

This research was supported by the National Key Research and Development Program of China (2023YFB2804801), National Natural Science Foundation of China (NSFC) (62331013), and Shenzhen Science and Technology Program (KJZD20230923115202006).

Author contributions

Z.C. proposed the MAHT architecture, developed the ML code, and wrote the manuscript. Z.C., Z.Z., M.C., L.D. performed the experiments. Z.Z., M.C. developed the linear compensation DSP. Z.C., Z.Z., X.X., M.C., D.L., M.T., and L.D. discussed and refined the details and feasibility of the architecture. M.C., L.D. initiated and supervised the project and provided the experimental equipment. Z.C., Z.Z., X.X., M.C., M.T., and L.D. completed the revisions of the manuscript.

Data availability

The data that support the findings of this study are available from the corresponding author on reasonable request.

Conflict of interest

The authors declare no competing interests.

References

1. van den Hout, M., S. Luís, R., Puttnam, B.J. et al, Reaching the pinnacle of high-capacity optical transmission using a standard cladding diameter coupled-core multi-core fiber. Nat. Commun. 16, 3833 (2025).
2. Le, S., Aref, V. & Buelow, H, Nonlinear signal multiplexing for communication beyond the Kerr nonlinearity limit. Nature Photon. 11, 570–576 (2017).
3. L. Xiang, et al, Nonlinearity mitigation in a fiber-wireless integrated system based on low-complexity autoencoder and BiLSTM-ANN equalizer. Opt. Express. 31, 20005-20018 (2023).
4. Lauinger V, Buchali F, Schmalen L, Blind equalization and channel estimation in coherent optical communications using variational autoencoders. IEEE J. Sel. Areas Commun. 40, 2529-39 (2022).
5. S. Sasipim, et al, Parallelization of recurrent neural network-based equalizer for coherent optical systems via knowledge distillation. J. Lightwave Technol. 42, 2275-2284 (2024).
6. H. Lu, et al, Global Receptive Field Designed Complex-Valued Convolutional Neural Network Equalizer for Optical Fiber Communications. Photonics. 11, 431 (2024).
7. J. Ziyun, X. Liu, and L. Zhang, Wide and Deep Learning-Aided Nonlinear Equalizer for Coherent Optical Communication Systems. Photonics. 11, 141 (2024).
8. S. Waddah S., S. K. O. Soman, and Octavia A. Dobre, Deep learning-assisted nonlinearity compensation in subcarrier-multiplexing coherent optical systems. J. Lightwave Technol. 42, 8162 – 8172 (2024).
9. M. Schädler, G. Böcherer, and S. Pachnicke, Soft-demapping for short reach optical communication: A comparison of deep neural networks and volterra series. J. Lightwave Technol. 39, 3095–3105 (2021).
10. Hamgini, B. Behinaein, et al, Pruning Attention in Transformers for Nonlinear Channel Compensation in Optical Systems. Opt. Fiber Commun. Conf., San Francisco, CA, USA, 2024, paper M4K.2.
11. Hao Ming et al., Ultralow Complexity Long Short-Term Memory Network for Fiber Nonlinearity Mitigation in Coherent Optical Communication Systems. J. Lightwave Techno. 40, 2427-2434 (2022).
12. F. Pedro, et al, Computational complexity optimization of neural network-based equalizers in digital signal processing: a comprehensive approach. J. Lightwave Technol. 42, 4177 – 4201 (2024).

13. Jarajreh, Mutsam A., et al, Artificial neural network nonlinear equalizer for coherent optical OFDM. IEEE Photon. Technol. Lett. 27, 387-390 (2014).
14. Bonneau, K., Lederer, J., Templeton, C. et al, Peering inside the black box by learning the relevance of many-body functions in neural network potentials. Nat. Commun. 16, 9898 (2025).
15. L. Yuzhe, et al, DACNN-aided nonlinear equalizer for a probabilistic shaping coherent optical communication system. Appl. Opt. 63, 1881-1887 (2024).
16. S. Bohan, et al, 420-Gb/s/channel WDM PS-64QAM transmission over 4,000-km ULAF using ring-wise neural network equalization. Opt. Fiber Commun. Conf. San Francisco, CA, USA, 2023, paper Tu2G.6.
17. S. Bohan, et al, 147.4 Tb/s DP-64QAM MDM-WDM Transmission over 500-km FMF utilizing MIMO Equalization Based on Multi-Label Neural Network. Opt. Fiber Commun. Conf. San Francisco, CA, USA, 2024, paper Th1H.4.
18. K. Miao, et al, 645-Gbit/s/carrier PS-16QAM WDM Coherent Transmission over 6,800 km Using Modified LSTM Nonlinear Equalizer. Eur. Conf. Opt. Commun. Bordeaux, France, 2021, 1-4.
19. M. Abu-Romoh, S. Sygletos, I. D. Phillips and W. Forysiak, Neural-network-based pre-impairment method to compensate for low resolution DAC nonlinearity. Eur. Conf. Opt. Commun. Dublin, Ireland, 2019, 1–4.
20. H. Xingyuan, et al, Complex principal component analysis-based complex-valued fully connected NN equalizer for optical fibre communications. Opt. Express, 31, 42310-42326 (2023).
21. H. Junling, et al, Fast fiber nonlinearity compensation method for PDM coherent optical transmission systems based on the Fourier neural operator. Opt. Express, 32, 2245-2256 (2023).
22. Liu, X., Wang, Y., and Li, C, Nonlinear equalizer by feature engineering-based deep neural network for coherent optical communication system. Asia Communications and Photonics Conference, pp:1-3, 2020.
23. Ding, N., Qin, Y., Yang, G. et al, Parameter-efficient fine-tuning of large-scale pre-trained language models. Nat. Mach. Intell. 5, 220–235 (2023).
24. S. Luo, S. K. O. Soman, L. Lampe. et al., Learning for perturbation-based fiber nonlinearity compensation. Eur. Conf. Opt. Commun. Basel, Switzerland, 2022, paper We1C.4.
25. Zhang, S., Yaman, F., Nakamura, K. et al., Field and lab experimental demonstration of nonlinear impairment compensation using neural networks. Nat. Commun. 10, 3033 (2019).
26. Elmorsy, S.M.A., Osman, S.M. & Gamel, S.A, Enhanced spectrum sensing for 5G and LTE signals using advanced deep learning models and hyperparameter tuning. Sci. Rep. 15, 24825 (2025).
27. K. Wang et al., Complex-valued 2D-CNN equalization for OFDM signals in a photonics-aided MMW communication system at the D-Band. J. Lightwave Technol. 40, 2791–2798 (2022).
28. X. Fang, M. Bi, L. Zhang, M. Zhu and M. Hu, A Kolmogorov-Arnold Networks-based low-complexity Equalizer for High-speed IMDD system. Opt. Fiber Commun. Conf. San Francisco, CA, USA, 2025, paper M2E.4.
29. R. Fischer, P. Matalla, S. Randel and L. Schmalen, Non-linear Equalization in 112 Gb/s PONs Using Kolmogorov-Arnold Networks. Opt. Fiber Commun. Conf. San Francisco, CA, USA, 2025, paper W4H.5.
30. L. Keren, et al, FPGA implementation of multi-layer machine learning equalizer with on-chip training. Opt. Fiber Commun. Conf. San Francisco, CA, USA, 2023, paper M1F.4.
31. J. Wenqing, et al, FPGA Implementation of Power-Lite Volterra-Inspired Neural Network Equalizer in 100G PON. J. Lightwave Technol. 42, 3134-3141 (2024).
32. F. Pedro J., et al, Implementing neural network-based equalizers in a coherent optical transmission system using field-programmable gate arrays. J. Lightwave Technol. 41, 3797-3815 (2023).
33. Kim, S., Lim, M., Kim, Y. et al, Impact of Synaptic Device Variations on Pattern Recognition Accuracy in a Hardware Neural Network. Sci. Rep. 8, 2638 (2018).
34. Freire, Pedro J., et al, Deep neural network-aided soft-demapping in coherent optical systems: Regression versus classification. IEEE Trans. on Commun. 70, 7973-7988 (2022).
35. F. Diedolo et al., Nonlinear equalization for optical communications based on entropy-regularized mean square error. Eur. Conf. Opt. Commun. Basel, Switzerland, 2022, paper We2C.2.
36. N. Zekun, et al, Learnable digital signal processing: a new benchmark of linearity compensation for optical fiber communications. Light: Science & Applications, 13, 188 (2024).
37. B. Sang et al., 420-Gb/s/channel WDM PS-64QAM Transmission Over 4,000-km ULAF Using Ring-Wise Neural Network Equalization. Opt. Fiber Commun. Conf. San Francisco, CA, USA, 2023, paper Tu2G.6.
38. Z. Cao, et al., Meta-learning nonlinear equalizer for hyperparameters tuning for 206.9 Tbps PCS-WDM optical transmission. Opt. Fiber Commun. Conf. San Francisco, CA, USA, 2025, paper Th2A.19.
39. Li, R., Ye, H., Jiang, D. et al, A computational framework for neural network-based variational Monte Carlo with Forward Laplacian. Nat. Mach. Intell. 6, 209–219 (2024).
40. Singh, S., Hernández-Lobato, J.M, A meta-learning approach for selectivity prediction in asymmetric catalysis. Nat. Commun., 16, 3599 (2025).
41. V. Anna, et al, Advances and challenges in meta-learning: A technical review. IEEE Trans. Pattern Anal. Mach. Intell. 46, 476 -4779 (2024).
42. T. Zhang, Q. He, M. Luo, C. Li, and S. Liu, Task-aware meta equalizer for multi-scenario generalization in coherent DWDM systems. Opt. Express 33, 47246-47256 (2025).
43. S. Han, J. Pool, J. Tran, and W. J. Dally, Learning both weights and connections for efficient neural networks. International Conference on Neural Information Processing Systems (NIPS'15), Vol. 1. MIT Press, Cambridge, MA, USA, 1135–1143.
44. B. Fred, et al, Implementation of 64QAM at 42.66 GBaud using 1.5 samples per symbol DAC and demonstration of up to 300 km fiber transmission. Opt. Fiber Commun. Conf., San Francisco, CA, USA, 2014, paper M2A.1.