IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 19 July 2024, accepted 5 August 2024, date of publication 9 August 2024, date of current version 20 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3441096

RESEARCH ARTICLE

Structural Optimization for Asymmetrical Inline Topology Filter With Transmission Zeros Using Goal-Oriented Reinforcement Learning

KIET YEW LEONG[1], **(Student Member, IEEE)**, **SOCHEATRA SOEUNG**[1], **(Senior Member, IEEE)**, **SOVUTHY CHEAB**[2], **(Senior Member, IEEE)**, **AND CHENG-KAI LU**[3], **(Senior Member, IEEE)**

[1]Department of Electrical and Electronics Engineering, Universiti Teknologi PETRONAS, Perak, Seri Iskandar 32610, Malaysia
[2]Cambodia Academy of Digital Technology, Phnom Penh 12252, Cambodia
[3]Department of Electrical Engineering, National Taiwan Normal University, Taipei 106, Taiwan

Corresponding authors: Cheng-Kai Lu (cklu@ntnu.edu.tw) and Socheatra Soeung (socheatra@gmail.com)

This work was supported in part by the National Science and Technology Council, Taiwan, under Grant NSTC 111-2222-E-003-001, Grant NSTC 112-2221-E-003-008, and Grant NSTC 113-2221-E-003-006-MY2.

ABSTRACT This paper presents a novel structural optimization approach for an asymmetrical inline topology with transmission zeros, known as Extracted Pole Unit (EPU) filter using a goal-oriented reinforcement learning method, specifically a hybrid of Soft Actor Critic (SAC) and Hindsight Experience Replay (HER). The recent popularity of reinforcement learning (RL) algorithms for optimizing cavity bandpass filters (BPFs) has led to several limitations. RL algorithms are inherently sample-inefficient, leading to prolonged model training times. A substantial number of training samples are required to achieve accurate results for a complex design using RL models. Additionally, most objective functions used for fitness calculations do not account for predistortion constants, which are crucial in synthesizing asymmetrical inline topologies with three transmission zeros as demonstrated in this work. To address these challenges, the proposed method incorporates predistortion-modified poles and transmission zeros within a feature-assisted objective function for use in the optimization process. Subsequently, a hybrid of SAC and HER is adopted as the optimization algorithm to leverage its improved sample efficiency by encouraging learning from diverse optimization scenarios and outcomes. The proposed method can optimize the self-couplings of the EPU filter in fewer optimization steps, showcasing enhanced training convergence speed and design accuracy.

INDEX TERMS Bandpass filter (BPF), extracted pole unit (EPU), neural networks (NNs), reinforcement learning, S-parameters.

I. INTRODUCTION

Structural optimization of microwave filters during the filter design phase is a stringent process. Conventionally, manual tuning is performed by an experienced engineer; who adjusts the filter's structural parameters and evaluates the filter's response based on preset goals, such as return loss, bandwidth, stopband rejection, and transmission zeros locations in S-parameters. Advanced Electronic Design Automation (EDA) software enables engineers to perform a simulation analysis of a design. Besides, with built-in mathematical optimization algorithms such as genetic algorithm, particle swarms, and conjugate gradient, the engineers can perform the local and global optimization analysis to generate a feasible set of their design structural parameters.

However, owing to the strict specifications and unique asymmetrical inline topology with transmission zeros used in this work, the manual tuning or conventional mathematical optimization process can be time-consuming and inevitably meet the requirements, which leads to more reliable solutions. This is because the objective function used does not account for the features in the target S-parameters such as

The associate editor coordinating the review of this manuscript and approving it for publication was Ali Karami Horestani.

poles and transmission zeros. The local algorithms such as conjugate gradient has higher risk of converging solutions prematurely, while the global algorithms such as genetic algorithm requires large number of parameters sweeps in the global search space when a complex design is involved. For example, the Extracted Pole Unit (EPU) structure requires precise locations of transmission zeros, TZs. Optimizing such a structure will easily induce prematurely converged and non-feasible solutions when locations of TZs are perturbed.

The novelty of this work is the use of an off-policy reinforcement learning, Soft Actor-Critic (SAC), and hybrid with Hindsight Experience Replay (HER) approach to optimize the EPU filter. Herein, we refer to this as "SAC + HER". The SAC + HER algorithm can achieve high sample efficiency and high training convergence speed in learning the non-linear relationship between the structural parameters and filter response. Using a hybrid approach with HER to store the sample transitions can reduce the effect of sparse reward problems inherent in reinforcement learning caused by the non-uniqueness input-output relationship. In the context of optimization, a feature-assisted objective function proposed by [19] is adopted but modified to consider poles in S_{11} with a series of predistortion constants α to direct poles into an acceptable region in the passband, as well as relocating the transmission zeros.

The Extracted Poles Unit (EPU) synthesized from the method used in [2] is an example of the design structure used in this experiment. The EPU filter has a pass-band frequency centered at 3.605GHz with a narrow bandwidth of 380MHz. The parameters to be optimized are the depths of the screw penetration of the resonators, which represent self-couplings. We train the SAC + HER model by running many optimization instances of the design to minimize the error between actual and preset goals. The SAC + HER model takes predistorted poles of S_{11} as inputs and generates the set of structural parameter delta values, which are later used to update the depth of the screw penetration on the EPU design.

The rest of this paper is organized as follows. Section II reviews related works. Section III describes the framework structure of goal-oriented reinforcement learning. Section IV describes the methodology and experimental setup. Section V presents the experimental results. Finally, Section VI concludes the paper.

II. RELATED WORK

The synthesis of a new class of inline filter topology with transmission zeros was introduced in [2]. The new method reconfigures eigenvalues and residues of the admittance function with predistortion constants to realize the prescribed transmission zeros owing to the dangled resonators connected to the source or load of the network. Different to the works in [3], [4], [5], [6], [7], and [8], the author in [2] adopted the predistortion constants to guide the poles in the reflection polynomials, S_{11}, to be within the values of -1 and 1 in the low-pass transformation. Depending on the location of the source and load, the poles of reflection polynomials of the dangled resonators are equal to the prescribed transmission zeros, whereas the poles of its transmission polynomials are equal to zero. The transversal topology within the source and load is predistorted with predistortion constants; then, it can be further transformed into an inline topology using the matrix rotation technique in [1]. For dangled resonators outside the source and load, their coupling matrix can be solved mathematically such that the poles of the transmission polynomials will be equal to zero. This new class of synthesis methods provides a hypothesis to investigate whether these predistortion constants can act as additional features in constructing the state vectors to train the SAC model.

Optimization is an iterative process that requires repetitive evaluations of the design model to capture feasible design parameters that satisfy the design specifications. From [29], the major challenges in modern optimization are uncertainty of optimization, multi-objective design in high-dimensional spaces, and manufacturability. To address these issues, [29] suggested advanced optimization methods using surrogate models, automated feature engineering, and machine learning techniques. Conventional common mathematical optimization methods, such as space-mapping optimization, are widely adopted in the optimization of multiplexers and switches [30]. A cognition-driven formulation of space mapping (SM) proposed in [33] is used for the equal ripple optimization of microwave filters and to estimate the yield in their design by mapping statistical variable spaces to feature parameter spaces. In [34], this cognition-driven space mapping approach that eliminates the need for explicit surrogates was further applied to EM-based filter optimization with equal-ripple responses. Unlike previous algorithms that accelerate optimization by creating a simplified model of a complex system, [31] used linear regression to calculate inverse space mapping to search for design parameters in the high-fidelity model that correspond to the desired outputs in the coarse model.

The studies in [10], [21], [28], and [32] demonstrated the possibility of using a forward model built with an artificial neural network (ANN) that aims to learn the behavior of passive or active circuits as a new modeling method. The inputs to the neural network are the circuit characteristics, whereas the outputs of the network are the electrical parameters, composed of S-parameters in magnitude dB, and phase in degrees. The application of such ANN model can be used to replace heavy simulation needed for each evaluation during optimization. However, the accuracy of the ANN model is highly dependent on the datasets and the trained ANN model is only specific to a particular design.

Subsequently, [11], [12], [13], and [14] used an inverse model to reverse the process of how a neural network should learn, taking input as electrical parameters and output as the geometrical parameters. The works in [15] considered undesired cross-couplings that potentially degrade the performance of NN by Eigenmode-based NN, taking the ideal coupling matrix of the transversal array filter as input to produce output geometrical or structural parameters. Such an inverse model is used to obtain the desired geometrical parameters of any RF design structure that has developed a new optimization approach.

Advanced data preprocessing is essential for enhancing the training of inverse models. Reference [11] introduced a method for dataset division by detecting the non-uniqueness in the input-output relationship, particularly for data with multivalued solutions. Besides, [12] employed filter decomposition to simplify high-dimensional neural network problems into sets of two-dimensional sub-neural network problems, and developed an empirical model to integrate these sub-networks. The study in [14] utilized pole-residue analysis of the electrical parameters from the transfer function as the model's inputs to handle the problem of order variations with changing geometrical parameters in large parameter spaces. Building on the non-uniqueness dataset preprocessing method in [11], the works in [16] divide datasets into partial geometrical parameters and introduce a dimensionality reduction neural network. In summary, the inverse model's accuracy can be improved by mitigating the non-uniqueness in datasets, reducing the dimensionality of the design parameters, and extracting critical features that help the neural networks to effectively map the geometrical and electrical parameters.

The inverse model is not generic for different filter design optimizations because the trained inverse model is unique to a particular design. For a more flexible solution, the surrogate model is capable in this scenario, which is heavily implemented in most EDA tools. The works in [17], [18], [19], and [20] performed design optimization using the surrogate model to partially replace heavy EM simulations to speed up the overall optimization process. In contrast to inverse model approaches, the surrogate model requires model refinement, such as active learning, to routinely improve accuracy over time. Thus, it is always used in parallel with conventional optimization algorithms.

In the optimization of microwave filters in the EM-based design process, the essential techniques discussed in [19] include sampling techniques, surrogate modeling, and optimization algorithms with AI-assisted methods. The works in [19] used a feature-assisted objective function that considers the positions of the reflection zeros, RZs, and the magnitude of S_{11} in dB. The formulation of the objective function aims to minimize the errors of the RZs positions if they are within the bandwidth around a desired center frequency in the low-pass domain, and to minimize the magnitude of the S_{11} to the desired return loss. The characteristic landscape of the feature-assisted objective function is then compared to different objective functions to evaluate the multi-modality, smoothness, and location of global optima in the valley of the landscape. From the experimentation, it was observed that the feature-assisted objective function has better smoothness and less multi-modality on the landscape, which depicts better convergence in the optimization. If the characteristic landscape has a very low multimodal profile, local minima are less likely to exist, which reduces the chances of the optimization being prematurely converged.

In addition to supervised learning, reinforcement learning has been gaining more attention ever since the success of OpenAI or DeepMind teams in solving robotic controls or automation tasks. To some extent, reinforcement learning has been used in filter optimization tasks as well in works of [22], [23], and [35]. The works in [22] used a Deep Q-Network (DQN) to predict the actions to update the filter's structural parameters, aided by another supervised neural network that is trained beforehand to predict the coupling matrix similar to [25]. This method can be extended to [27], which can automatically adjust the bandpass filters (BPFs) for multiple design goals. Because simulated filter response generation can be computationally demanding, training with a DQN can be very time-consuming. The proposed method can reduce the learning time because the second neural network replaces the simulation to predict the coupling matrix to generate S-parameters. In addition, [22] used a well-trained neural network to predict an initial set of parameters. The filter is then optimized with a DQN for better quality (e.g., achieving return loss in the specification).

Another approach to demonstrate the capability of reinforcement learning in tuning microwave filters is the works of [23] and [24] with additional modification either on the input states (observation) or reward shaping function required to train the RL model. The work in [23] uses the Locally Linear Embedding (LLE) technique to reduce the dimension of the input states represented by the S-parameters of the designs for standardizing training datasets and train the RL model with Double Deep Q-Network (DDQN). Unlike the common DQN with discrete action spaces to predict design parameter values, [24] uses a knowledge-inspired reward shaping function, together with the Deep Deterministic Policy Gradient (DDPG) algorithm to make continuous actions' prediction possible. However, DDPG uses a separate actor network to maximize the Q-learning value possesses extreme brittleness and hyperparameter sensitivity, as benchmarked in [42]. Besides taking input states as a vector, [26] uses a Convolutional Neural Networks (CNN) instead as the RL networks to learn the image representation of S-parameters curves, simplifying the state shaping and utilizing CNN's pooling layers to accomplish the feature extractions.

Different to the abovementioned approaches, an AI-assisted optimization algorithm can also be used to optimize the microwave filter. The work in [36] introduces a one-dimensional convolutional autoencoder (1D-CAE) surrogate-based electromagnetic optimization technique by incorporating particle swarm optimization (PSO) and neural networks for microwave filter design. Similarly, [37] presented the Reinforcement Learning-based Multi-Objective Differential Evolution (RLMODE) algorithm, which dynamically adjust

DE hyperparameters through reinforcement learning, can adaptively guides the solution towards feasible regions, thereby enhancing the optimization convergence.

In this paper, the proposed SAC + HER method is used to optimize the self-couplings of the EPU filter with transmission zeros. The SAC + HER method takes predistorted poles and transmission zeros as input and generates deviation values to update the penetration screw depths of the EPU filter. The synthesis method of predistortion constants is included in Section IV-C, which are adapted into a feature-assisted objective function based on [19] for use in the optimization process.

III. GOAL-ORIENTED REINFORCEMENT LEARNING (GORL) FRAMEWORK

A. SOFT ACTOR CRITIC (SAC)

SAC is an off-policy reinforcement learning algorithm that maximizes future rewards by maximizing the entropy of the policy [38]. As a result, SAC improves sample efficiency in training by exploring better with maximized entropy [41]. This is especially useful when highly sparse rewards which degrades the training performance [40] exist in highly difficult environments, such as optimizing the structural parameters of the microwave filter where non-uniqueness exists as explained in [11]. SAC does not require datasets collected beforehand; instead, the dataset mapping the input state vector and output action vector is collected and stored in a replay buffer during the learning phase. Similar to other reinforcement learning algorithms, the following main elements were implemented in the proposed method:

1) Environment: Integrating an EM simulation that updates the EPU structural parameters. The S-parameter response is simulated and evaluated to construct state observations and compute rewards to update the policy.
2) Action: Predicts the filter's structural parameters delta values.
3) State: Vector of observation is the poles value modified by predistortion constants.
4) Reward: Euclidean distance between current achieving state and the desired goal state.

The SAC utilizes three major networks: value network V, Q-network Q, and policy network π. During SAC model training, the objective function of each network must be optimized by taking their derivatives. The actions to be predicted are portrayed as a normal distribution in policy network π. The value network V is trained to minimize the squared residual error, as defined by

$$\begin{aligned} J_V(\psi) = \; & \mathbb{E}_{s_t \sim \mathcal{D}}[\frac{1}{2}(V_\psi(s_t) - \mathbb{E}_{a_t \sim \pi_\phi}[Q_\theta(s_t, a_t) \\ & - \log \pi_\phi(a_t|s_t)])^2]. \end{aligned} \tag{1}$$

where $\mathcal{D}$ is the replay buffer and $\mathbb{E}$ is the expected prediction. The formulation in (1) serves as the objective function of the value network V. It aims to minimize the discrepancy between the V-value estimated by the value network, $V_\psi(s_t)$ and the Q-value predicted by the current policy, $Q_\theta(s_t, a_t)$. Additionally, SAC introduces an entropy regularization term, $-\log \pi_\phi$ to encourage exploration to maintain stochasticity in the policy π. Maximizing entropy encourages a policy to explore different actions and learn a diverse set of behaviors, leading to better exploration and improved robustness in learning. Subsequently, two soft Q networks were trained to minimize the following error:

$$J_Q(\theta) = \mathbb{E}_{(s_t, a_t) \sim \mathcal{D}}[\frac{1}{2}(Q_\theta(s_t, a_t) - \hat{Q}(s_t, a_t))^2]. \tag{2}$$

where

$$\bar{Q}(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim \rho}[V_{\bar{\psi}}(s_{t+1})]. \tag{3}$$

Double soft Q networks are used for stability [38] by reducing the overestimation bias commonly observed in deep Q-learning algorithms. The objective function of the soft Q network (2) is based on the Bellman equation, which minimizes the mean squared error of the current prediction of the Q_θ function and the next state $\hat{Q}$ function. The next state Q function is defined in (3), which is the summation of the current reward value, $r(s_t, a_t)$ and the expected value of the next state, $V_{\bar{\psi}}(s_{t+1})$ which considers the entropy term, as in (1). This $V_{\bar{\psi}}(s_{t+1})$ is discounted by a constant γ to balance the importance of immediate versus the future rewards.

The policy network is trained to minimize the following error:

$$\begin{aligned} J_\pi(\phi) = \; & \mathbb{E}_{s_t \sim \mathcal{D}, \epsilon_t \sim \mathcal{N}}[\log \pi_\phi(f_\phi(\epsilon_t; s_t)|s_t) \\ & - Q_\theta(s_t, f_\phi(\epsilon_t; s_t))]. \end{aligned} \tag{4}$$

The SAC learns a stochastic policy that maps states to actions. In particular, it often parameterizes a Gaussian distribution in continuous action spaces, where ϵ is the epsilon term from the Gaussian distribution. The formulation in (4) is used to close the gap between the entropy of the current predicted actions against the current state, $\log \pi_\phi$ and the quality of the current state action from the soft Q network, Q_θ. Notice that the term $f_\phi(\epsilon_t; s_t)$ denotes the actions, a_t, that are calculated from the Gaussian distribution. To wrap up the three main objective functions, taking derivatives of (1), (2), and (4) gives gradient forms to be minimized, which are summarized as follows:

$$\begin{aligned} \hat{\nabla}_\psi J_V(\psi) = \; & \nabla_\psi V_\psi(s_t)(V_\psi(s_t) - Q_\theta(s_t, a_t) \\ & + \log \pi_\phi(a_t|s_t)). \end{aligned} \tag{5}$$

$$\begin{aligned} \hat{\nabla}_\theta J_Q(\theta) = \; & \nabla_\theta Q_\theta(s_t, a_t)(Q_\theta(a_t, s_t) - R(s_t, a_t) \\ & - \gamma V_{\bar{\psi}}(s_{t+1})). \end{aligned} \tag{6}$$

$$\begin{aligned} \nabla_\phi J_\pi(\phi) = \; & \nabla_\phi \log \pi_\phi(a_t|s_t) + (\nabla_{a_t} \log \pi_\phi(a_t|s_t)) \\ & - \nabla_\phi f_\phi(\epsilon_t; s_t). \end{aligned} \tag{7}$$

We used (5), (6), and (7) as the objective functions in their respective networks to be minimized in the training phase. In summary, five neural networks were utilized as two value networks, two soft Q networks, and one policy network. Both the value and soft Q networks act as critics, whereas the policy network is the actor. The actor predicts the actions, whereas the critic evaluates the performance of the action.

B. HINDSIGHT EXPERIENCE REPLAY (HER)

The primary motivation for using HER is to address the common issue of sparse rewards [39], where the SAC model may receive positive rewards only when it successfully tunes a filter. As discussed in [11], [12], and [16], the non-uniqueness problem can degrade the prediction accuracy if an RL agent has a single-input-multiple-output scenario that causes the sparse reward problem. In many real-world scenarios, learning from failures can be challenging, because the agent receives limited feedback. HER allows the agent to learn from its failures by treating unsuccessful training episodes as if they were successful but with a different goal. In this way, the agent can still gain valuable information from the experience, even if it does not achieve the original goal. This allows an RL agent to effectively learn from both successful and unsuccessful experiences, thus accelerating learning in scenarios where achieving the original goal is rare or difficult to achieve, or requires a very long training time. HER was used as the replay buffer of SAC, and the hybrid of this technique is used as the model to train the optimization of a microwave filter.

C. SAC AND HER HYBRID ALGORITHM

The framework of the proposed method is illustrated in Fig.1. The training process is summarized in Algorithm 1.

In Algorithm 1, owing to the addition of the goal state vector g, the transition is stored differently than in the original SAC algorithm in [18]. For each of the total K training epochs, there are a total T number of steps to collect the HER buffers, $\mathcal{D}$. HER stores a batch of training datasets required to train the fives networks in the SAC. In each step, actions a_t are sampled from the policy network, π_ϕ given the s_t and desired goal, g. These a_t are updated to the EPU design parameters discussed in Section III-B, and the next state, s_{t+1} is captured. Once the current epoch is terminated owing to certain criteria, such as the total number of steps are met, the optimization had reached convergence, or the optimization goals have been achieved, the last training data must be embedded into the HER. For each step, the reward signal r_t is computed by specifying the error to achieve the desired goal using (18), which is the Euclidean distance of s_t and g. The HER is required to take a set of additional buffers, in which the additional goal g' is sampled from any of the previously captured non-promising state, then a reward signal r'_t corresponding to g' is calculated. This enhances the sample efficiency by allowing the agent to learn from failures and explore alternative trajectories that lead to different outcomes. Finally, using the batch of HER, $\mathcal{D}$ to update the five networks in N gradient steps completes one cycle of SAC + HER model training.

Algorithm 1 SAC and HER Hybrid

```
Initialize parameters: $\psi, \bar{\psi}, \theta_1, \theta_2, \phi$
    $\bar{\theta_1} \leftarrow \theta_1, \bar{\theta_2} \leftarrow \theta_2, \mathcal{D} \leftarrow \emptyset$
    Sample a goal $g \in \mathcal{G}$
    for epoch $k = 1, K$, do
        Sample an initial state $s_0 \in \mathcal{S}$
        for environment step $t = 0, T$, do
            $a_t \sim \pi_\phi(a_t|s_t, g)$
            $s_{t+1} \sim \rho(s_{t+1}|a_t, s_t, g)$
        end for
        for environment step $t = 0, T$, do
            $r_t := r(s_t, a_t, g)$
            $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r_t, s_{t+1}, g)\}$
            Sample a set of additional goals $G$ from strategy
            for $g\prime \in G$ do
                $r_t\prime = r(s_t, a_t, g\prime)$
                $\mathcal{D} \leftarrow \mathcal{D} \cup \{(s_t, a_t, r_t\prime, s_{t+1}, g\prime)\}$
            end for
        end for
        for gradient step $n = 1, N$ do
            $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
            $\bar{\psi} \leftarrow \tau\psi + (1 - \tau)\bar{\psi}$
            $\theta_j \leftarrow \theta_j - \lambda_Q \hat{\nabla}_{\theta_j} J_Q(\theta_j)$ for $j \in \{1, 2\}$
            $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
        end for
    end for
Output: $\psi, \bar{\psi}, \theta_1, \theta_2, \phi$
```

IV. REINFORCEMENT LEARNING FOR EPU FILTER DESIGN

A. ENVIRONMENT SET-UP

The EM simulation was used to design and simulate the EPU filter to generate the filter response in the S-parameters. For each step in the episode, the K number of filter structural parameter delta values predicted from the SAC are updated to the design in the simulation. The simulated S-parameters were then evaluated, producing a reward value, $r_t = r(s_t, a_t)$. Before each step is terminated, the state vector, s_t, next state vector, s_{t+1}, action vector, a_t, reward value, r_t and goal state vector, g are stored in the HER, denoted by $\mathcal{D} = (s_t, a_t, r_t, s_{t+1}, g)$. For a preset number of buffer sizes, the SAC learns to generalize periodically using the buffer dataset in the HER.

B. ACTIONS

In the approach to tuning the EPU structural parameters, K numbers of the structural parameter set, $P = (p_1, p_2, \ldots, p_K)$ which can also be denoted by $p_j \in P$ where $(j = 1, 2, \ldots, K)$, are chosen for optimization. The structural parameter set, P, is preset with an initial value, where each value must be bounded with minimum and maximum values to define a feasible search space for the solution. In each step of the episode, the action vector output from the SAC policy is $a_t = \pi(s_t||g)$ where $a \in \mathcal{A}$ is the unscaled delta value of each structural parameter and $\triangle p_j = a_t b$ where b is the scale

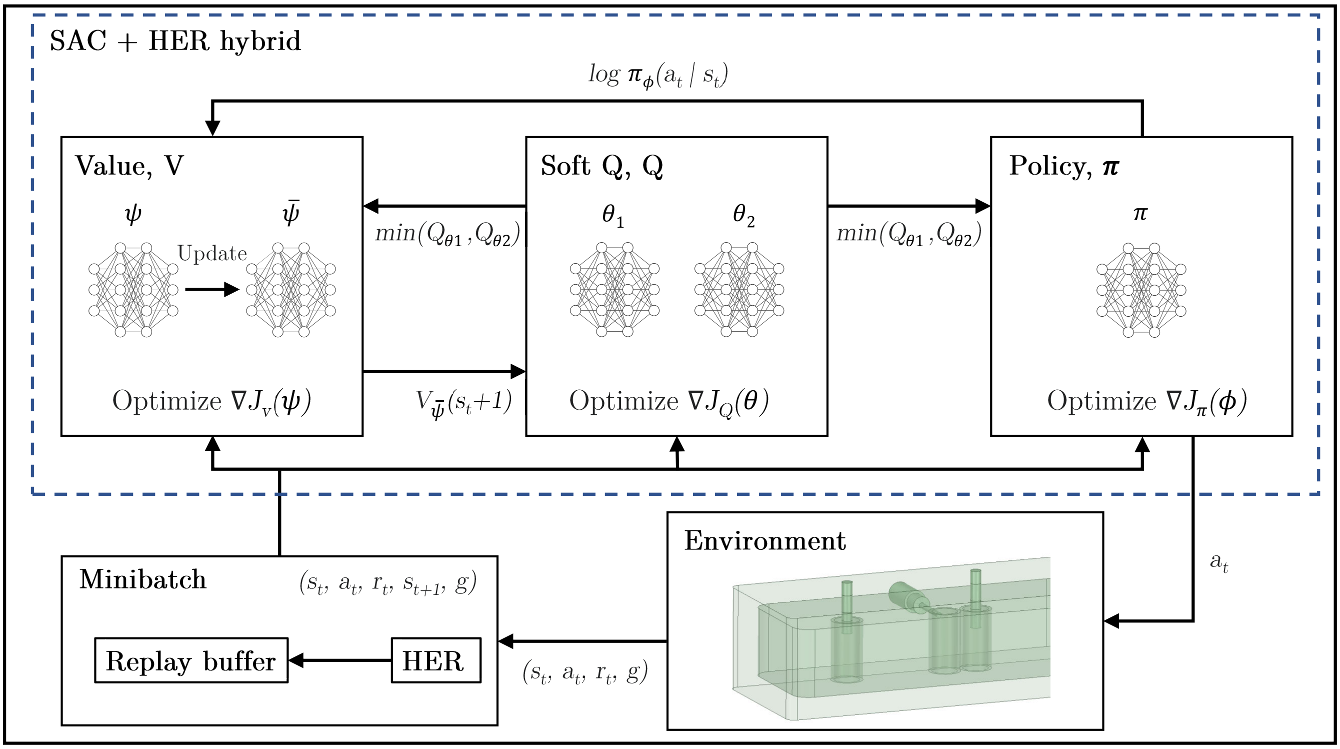

SAC + HER hybrid
$log\ \pi_\phi(a_t \mid s_t)$
Value, V
ψ
$\bar{\psi}$
Update
Optimize $\nabla J_v(\psi)$
$min(Q_{\theta 1}, Q_{\theta 2})$
$V_{\bar{\psi}}(s_t+1)$
Soft Q, Q
θ_1
θ_2
Optimize $\nabla J_Q(\theta)$
$min(Q_{\theta 1}, Q_{\theta 2})$
Policy, π
π
Optimize $\nabla J_\pi(\phi)$
Environment
a_t
Minibatch
$(s_t, a_t, r_t, s_{t+1}, g)$
Replay buffer
HER
(s_t, a_t, r_t, g)

FIGURE 1. Hybrid of SAC and HER framework.

factor that ensures the delta values are in the correct unit. To perform the update, delta values are added to the respective p_j, which is defined as

$$p_{j,t+1} = p_{j,t} + \triangle p_j. \tag{8}$$

The filter response is then simulated, which is required to extract the next state vector, s_{t+1} and compute the reward r_t.

C. STATES

Reinforcement learning learns to act based on state features; it is the observation or feature to tell the SAC model what action should be performed next. The reflection parameter S_{11} is used for the feature extraction into the state vector because they are derived from the characteristic polynomials of a microwave filter. In this study, eight RZs were extracted and used in the shaping of the state vectors. The vector fitting technique from [9] used to extract the poles RZs and residues from the simulated S_{11} response. These RZs extracted from vector fitting were predistorted by the constants α owing to dangled resonators at the first resonator and last two resonators Thus, the RZs extracted by vector fitting technique and predistorted by the constants α are as follows:

$$f'_{z,i} = f_{z,i} + \alpha_i, \quad i = (1, 2, \ldots, N) \tag{9}$$

where $f_{z,i}$ denotes the RZs extracted by vector fitting and N is the filter order. Then the formulation of the state vector s can be shaped by the eight RZs, depicted as follows:

$$s_{t+1} = \{f'_{z,i}\}, \quad i = (1, 2, \ldots, N) \tag{10}$$

Note that the suffix of state, $t+1$ is denoting the next state. This is because, for every state vector, is shaped after each action is taken. This state vector captures the current state of the EPU filter's responses, which is a crucial reference for the employed model to predict the next action based on s_{t+1}.

D. OPTIMIZATION OBJECTIVE FUNCTION

To facilitate the optimization process, a meticulously crafted objective function was employed utilizing the feature-assisted objective function proposed in [19], with some modifications to consider the predistortion constant, α of the reflection polynomial from the introduction in [2]. The objective function is modified as follows:

$$\begin{aligned} F = Min \Bigg\{ & w_1 \cdot max\Bigg(\sum_{i=1}^{N-1} max\ S_{11,dB}(f'_{z,i}, f'_{z,i+1}) - D, 0\Bigg) \\ & + w_2 \cdot \sum_{i=1}^{N} max\Big(|f'_{z(i)} - CF| - BW/2, 0\Big) \\ & + w_3 \cdot \sum_{j=1}^{M} max\Big(|f_{tz(j)} - TZ(j)|, 0\Big) \Bigg\} \end{aligned} \tag{11}$$

where f'_z are the extracted RZs locations that naturally include the predistortion constant α and f_{tz} are the extracted TZs.

f_z' and f_{tz} were determined using the vector fitting method on simulated S_{11} and S_{21}. CF represents the desired center frequency, BW represents the desired bandwidth, and TZ represents the predefined transmission zero of the filter design, as in [2]. D represents the desired return loss of $S_{11,dB}$. In this study, the return loss, D is equal to -18dB value. Weights $w1$, $w2$, and $w3$ were used to control the influence of the three evaluations. The objective function is formulated with the features of the solution responses such as S_{11} poles, RZs and transmission zeros, TZs. It aims to prioritize the adjustment of RZs positions and subsequently refine the return loss within the pass-band, and then adjust the location of the TZs to agree with the predefined TZs. The predistortion constants compromise the three TZs in an inline topology. Thus, it is important to include them in the formulation. In summary, the formation of (11) guides the learning of the employed SAC model to reduce the error of misaligned TZs and ensure that all RZs are within -1 and 1.

E. EXTRACTION OF PREDISTORTION CONSTANT

The inline topology of asymmetrical network with three TZs is shown in Fig.2.

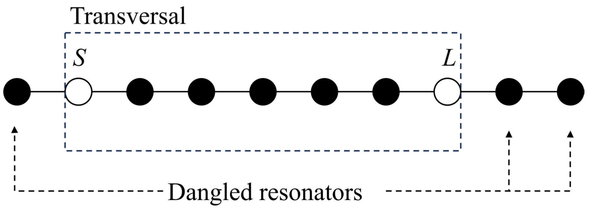


FIGURE 2. Inline topology for asymmetrical network with 3 TZs.

Considering the case in (14) is satisfied, the transversal array network to describe the inline topology with dangled resonators is shown in Fig.3.

$$r'_{11,k} \; or \; r'_{22,k} = 0, \quad k = 1, 2, \ldots, N \tag{12}$$

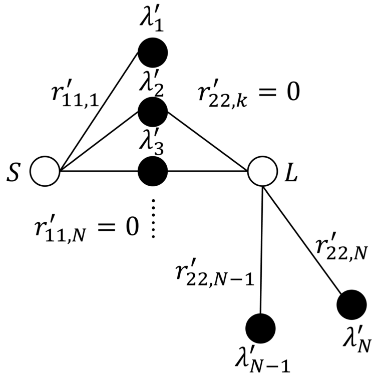


FIGURE 3. Transversal array of the asymmetrical network with 3 TZs.

In Fig.3, there are three dangled resonators: λ'_1, λ'_{N-1}, and λ'_N. In [2], the predistortion method was used to reconfigure the selected eigenvalues to the location of the prescribed TZs to realize these dangled resonators. Using the common rational polynomials from [1] yields

$$S_{11} = \frac{F_S}{E_S}, \quad S_{21} = \frac{K_{12} P_S}{E_S} \tag{13}$$

and the admittance function given by

$$Y_d = E_S + E_S^* + F_S + F_S^* \tag{14}$$

where the reflection polynomial F_S with predistortion constant, α is

$$F_S = \prod_{i=1}^{N} (s + \alpha_i + I * R_{z,i}) \tag{15}$$

It is necessary to compromise the reflection specification by altering the scaling factor K_{12} such that the unitary condition of $S_{11}S_{11}^* + S_{21}S_{21}^* = 1$ can still be satisfied. Thus, the altered scaling K_{12} is proposed as follows:

$$K_{12} = \frac{\beta P_S}{F_S} * \frac{1}{\sqrt{10^{RL/10} - 1}} \tag{16}$$

where β is the unknown that compromises the reflection specification when the predistortion constant α is placed in the reflection polynomial F_S. The calculation of α and β is proposed using optimization, where the objective function to be minimized is:

$$F_{\alpha,\beta} = Min \sum_{i=0}^{N} (|max \; roots(Y_d)| - TZ(i))^2 + (|min \; roots(Y_d)| - TZ(i))^2 \tag{17}$$

The formulation in (17) is used to ensure that the poles or eigenvalues are smaller than the three TZs. Finally, α and β are extracted when (17) is satisfied.

F. REWARDS

The employed model is goal-oriented in the context of the HER. The reward signal value r_t in this proposed method is computed by taking the Euclidean norm of the goal state g and the current state s_{t+1} after the actions have been taken: the reward function is defined by

$$r(s_{t+1}, a_t) = -(|g - s_{t+1}|^2)^{\frac{1}{2}} - F. \tag{18}$$

where F is the objective function in (11), and g is the vector consisting of the eight ideal RZs values from the synthesized EPU filter in [2]. The higher the reward function, the smaller the error between the current state s_{t+1} and the goal state, g, which also minimizes (11).

G. WORKFLOW

To combine pieces, a workflow is constructed to perform iterative optimization using SAC + HER algorithms, typically to predict actions based on the previous state vector (observation), update the design parameters, evaluate the design's S-parameter using a feature-assisted objective function, update the next state vector, and repeat the cycle until the termination criteria met. Each of this cycle is one environment step in Algorithm 1. The workflow is illustrated in Fig.4.

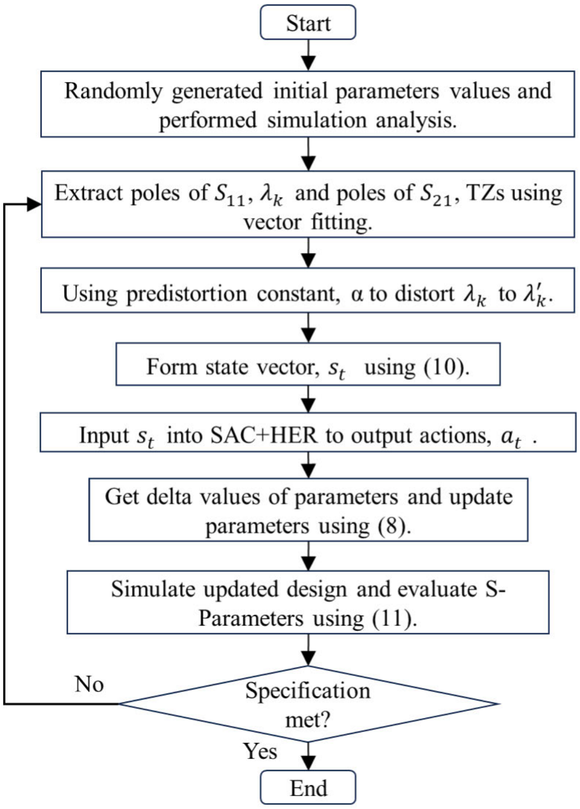

Start
Randomly generated initial parameters values and performed simulation analysis.
Extract poles of S_{11}, λ_k and poles of S_{21}, TZs using vector fitting.
Using predistortion constant, α to distort λ_k to λ'_k.
Form state vector, s_t using (10).
Input s_t into SAC+HER to output actions, a_t .
Get delta values of parameters and update parameters using (8).
Simulate updated design and evaluate S-Parameters using (11).
Specification met?
No
Yes
End

FIGURE 4. Optimization workflow.

V. DESIGN OPTIMIZATION

A. DESIGN SPECIFICATION

An example of using the SAC + HER is demonstrated for an 8th-order Extracted Pole Unit (EPU) structure, which is designed using an inline topology with an asymmetrical network. The goal is to tune the self-couplings of the structure to obtain a good return loss in the S_{11} while relocating transmission zeros at S_{21}. In this design example, the penetration depth of the screw at each of 8 resonators need to be optimized. The structural parameters are $P = (p_1, p_2, \ldots, p_N)$, where p is the depth of screw penetration, and N is the filter order.

The 3D structure of the EPU filter is designed using Ansys Electronics Desktop (HFSS) v2019. An Application Programming Interface (API) script was constructed to enable integration between the adopted AI framework and HFSS. In Fig.5, the EPU filter's dimensions (width, w × length, l × height, h) are 30 mm × 190 mm × 18 mm. It has eight radial resonators made up of aluminium, each with diameter (d_1) of 9 mm, and the thickness of 0.5 mm. Each penetration screw is made up of aluminium with diameter (d_2) of 4 mm. The filter's casing is made up of aluminium. The source (S) and load (L) ports are made up of Teflon (inner) and Brass (outer cover). The source is connected to resonator 1 and resonator 2, while the load is connected to resonator 6 and resonator 7, both connections are done via gold pins to bridge the dangled resonators at position 1 and 7. The inter-resonators gaps are $g_{12} = 14.6$ mm, $g_{23} = 11.44$ mm, $g_{34} = 15.36$ mm, $g_{45} = 15.5$ mm, $g_{56} = 13.64$ mm, $g_{67} = 14.6$ mm, and $g_{78} = 9$ mm.

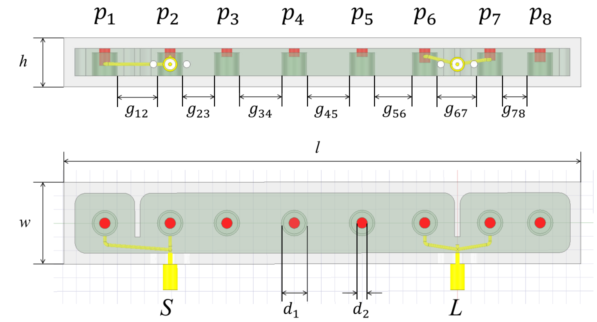


FIGURE 5. Structural parameters to be tuned in the EPU's structure are the depth of penetration of the screws, $P = (p_1, p_2, \ldots, p_N)$.

The depth of penetration is in the mm unit. The initial values for each parameter are randomly set. The minimum and maximum boundary of each parameter is defined as $0.1\ mm <= p_i <= 5\ mm$, where $i = 1, 2, \ldots, N$, to ensure that the filter design is valid and feasible. The 8 structural parameters will contribute to eight action sizes, which is the output from the policy of SAC, $a_t = \pi(s_t||g)$. The off-policy SAC is constructed to output continuous action between the normalized value of -1 and 1 for each action, a_t, then the delta value of each parameter, $\triangle p_j$ is computed by multiplying a_t with a scale factor, b. The scale factor b is calculated by $(5mm - 0.1\ mm) * 0.1$, which is the 10% difference between the maximum (5 mm) and the minimum (0.1 mm) values. The single step to update each parameter is capped at 10% of the whole value range to prevent excessively large adjustments that could destabilize the system. Then, the formulation in (8) is employed to compute the updated parameter value, thereby completing a single parameter adjustment. The hyperparameters of SAC + HER are listed in Table 1.

The framework hyperparameters are kept simple, which is adopting the default hyperparameters setup in [43]. The authors in [43] considers the continuous action values and set the entropy target to $-dim(\mathcal{A})$, where $\mathcal{A}$ is the dimension of action size. The entropy target is used to encourage exploration by promoting diversity in the model's actions. The goal selection strategy in Algorithm 1 used is future type, this means the intermediate goal, g' is achieved after the current step within the same epoch. In layman terms, the SAC algorithm considers goals that can be achieved later in the same period of learning when training data are sampled from HER for updating the policy in SAC during the training phase. By considering these additional goals during model

TABLE 1. Hyperparameters setting of SAC and HER hybrid.

General	Value
Input state size	8
Output action size	8
Network layers	256 X 256 X 256
Batch size	512
Non-linearity	ReLu
Optimizer	Adam
learning rate	1e-3
SAC	**Value**
Target smoothing coefficient (τ)	0.05
Discount (γ)	0.95
Gradient steps	1
Target update interval	1
Entropy target	-dim($\mathcal{A}$)
HER	**Value**
Replay buffer size	1e6
Number of sampled goals	4
Goal selection type	future

training, the model can learn from its failures and explore different strategies to accomplish them.

B. RESULTS OF THE OPTIMIZATION OF THE EPU FILTER

The predistortion constants α synthesized from Section IV-C are listed in Table 2. The first pole λ'_1 and last pole λ'_8 were not distorted to preserve the bandwidth of S_{11}. The middle poles from λ'_2 to λ'_7 are predistorted by the constants to compromise the dangled resonators at 1, 7, and 8. These eight predistorted poles λ' are the goal state g to be achieved in the optimization framework.

The SAC + HER model was applied in this EPU filter optimization, and the model is trained for 150 epochs. In each episode, there were a maximum of 50 steps to adjust the penetration screws. The condition to terminate the episode is either when 50 maximum steps are reached, or when preset design goals are achieved. The training of this EPU example optimization took 2 days on a High-Performance Computing (HPC) system using Intel(R) Xeon(R) CPU E5-2667 v2 @ 3.30GHz with 128GB RAM.

In Fig.6, the training performance is summarized by the cumulative rewards in each epoch. At approximately 60 epochs (3000 evaluations), the solution is converged. The trained model can now be used to optimize the EPU filter in fewer steps and produce good return loss in S_{11}, as well as relocate the three TZs.

To validate the optimization using the trained model, two examples with different set of initial values for the screw penetration depths that are randomly determined are used. In Fig.7, the SAC + HER model can effectively optimize the self-couplings of the EPU filter, achieving good agreement with the preset goals, which are the TZs location and S_{11} return loss of 18 dB. This demonstrates that the SAC + HER model successfully correlates the non-linearity between the screw penetration depths, the poles, and the TZs with the predistortion constants by using objective function in (11).

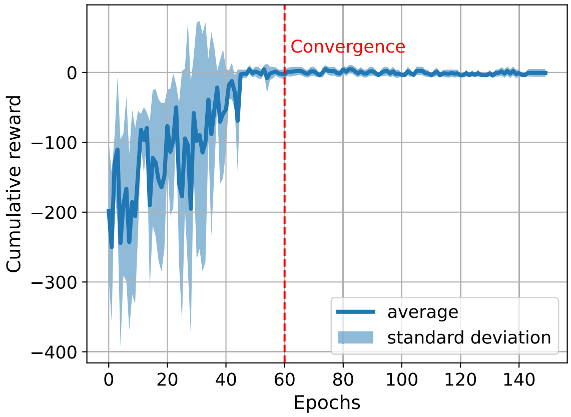
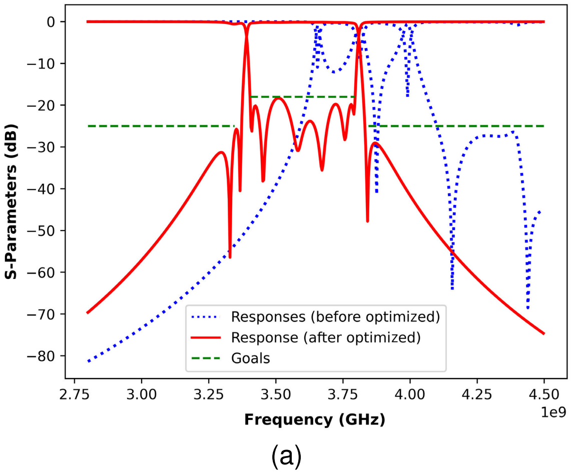


FIGURE 6. Cumulative rewards of training in 150 epochs. Training is converged at approximately epoch 60.



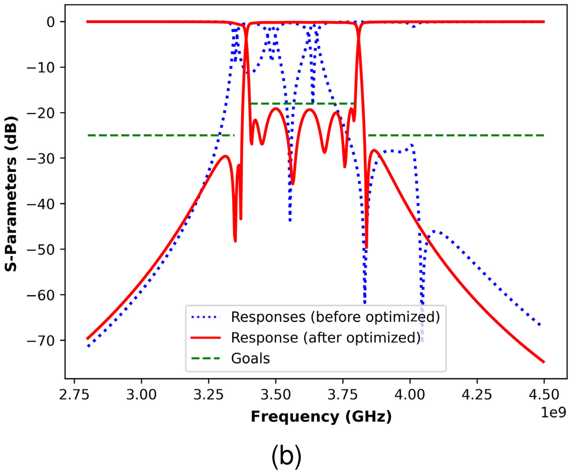


FIGURE 7. Comparison of the initial and optimized EPU filter using the SAC + HER trained model. (a) Example A and (b) Example B.

TABLE 2. Predistortion constant, α for each S_{11} poles, λ'_k.

λ'_k	-j0.9920	-j0.9178	-j0.7251	-j0.3759	j0.0861	j0.5368	j0.8498	j0.9852
α	0	0.0407	0.0228	0.0195	0.0257	0.0274	0.0447	0

TABLE 3. Results of the optimized screw penetration depths for two examples of different randomized initial values.

	Example A		Example B	
	Initial	Optimized	Initial	Optimized
	(mm)	(mm)	(mm)	(mm)
p1	0.1	3.37	2	3.32
p2	0.1	2.93	2	2.97
p3	0.1	3.72	2	3.69
p4	0.1	1.58	2	1.59
p5	0.1	1.48	2	1.53
p6	0.1	1.53	2	1.60
p7	0.1	4.32	2	4.27
p8	0.1	4.79	2	4.53
reward	0.13		0.22	

The results of optimized parameters values are presented in Table 3.

It is interesting to analyze the optimization response surface plot of the EPU filter to visualize the location of the optimal solution in a low-dimensional space. In this study, a feature reduction technique using Principal Component Analysis (PCA) is used to represent the optimization problem in a two-dimension response surface plot. The PCA technique is used to project the eight-dimensional multiple variations of design parameters with distinct values into two-dimensional sets. Then, the eigenvalues and eigenvectors of the data are computed. The first two eigenvectors with the highest eigenvalues are chosen, allowing each set of eight-dimensional design parameters to be represented by an eigenvector of two values.

A dataset X with n samples and 8 features is prepared by sampling n sets of different design parameters variations, with each analyzed in the simulation, and then evaluated using the objective function (11), where the dataset X is represented as:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,8} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,8} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,8} \end{bmatrix}$$

To perform the PCA technique to reduce the dimensions from eight features to two features, the dataset was first standardized. Then, a covariance matrix Σ, of X was computed using the standardized features.

$$\Sigma = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X}) \tag{19}$$

where $\bar{X}$ denoted the mean of the dataset X. The eigenvectors and eigenvalues of Σ which represent the principal components and the amount of variance explained by each principal component, respectively, were obtained. The dataset was projected onto the first two principal components to obtain the transformed dataset X':

$$X' = X \cdot V \tag{20}$$

where V is the matrix of eigenvectors corresponding to the two largest eigenvalues. The resulting dataset X' will have n samples with only two features, utilizing X' as the XY coordinate for each corresponding evaluated cost to construct the response surface. The 2-D response surface for the optimization of the EPU filter is shown in Fig.8.

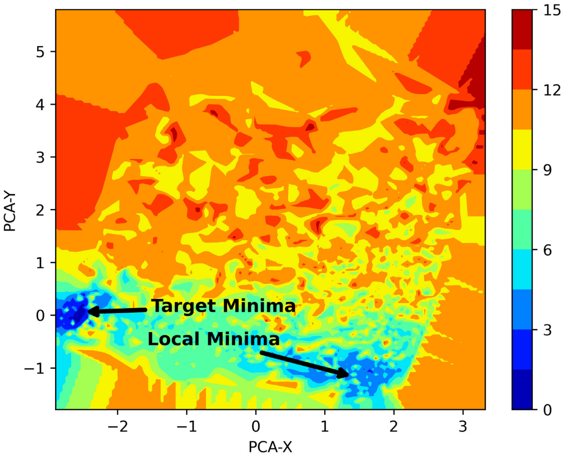


FIGURE 8. 2-D response surface of the optimization problem with red region indicates high-cost area and blue region indicates low cost area.

Note that there are two regions of possible minima: one is the global minima where the solution should converge, and the other is the local minima where the solution should be avoided. If a solution converges with the second solution, premature convergence would occur, and the goal of the design would never be met. Fig.9 depicts the design response corresponding to the solution that falls into the local minima.

In Fig.9, it is clearly observed that the second TZ, which ideally should be at 3.375GHz, switch to 3.875GHz, causing the S_{11} to not meet the desired return loss. A comparison of the optimally and locally converged design parameter solutions is summarized in Table 4.

From Table 4, the screw penetration depth at eighth resonator, $p8$ of the solution that converges in the local minima is very small compared to that in the global minima. This causes a misalignment in the location of TZ. With this SAC and HER model, the solution converges to the global minima, demonstrating a good search ability for the optimal solution.

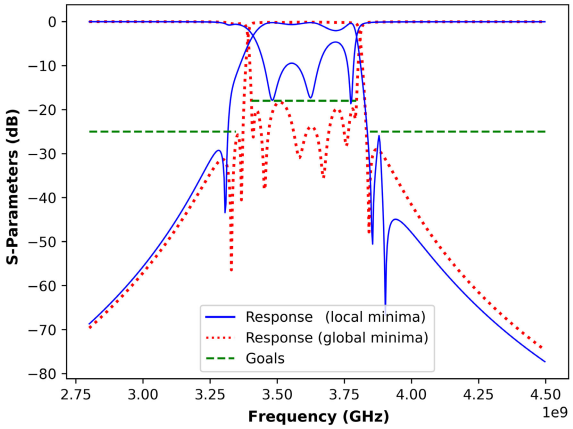


FIGURE 9. EPU S-Parameters response trapped into local minima with the second TZ should be at 3.375GHz switched to 3.875GHz.

TABLE 4. Comparison of the optimally converged design parameters and locally converged solution.

	Optimal solution (mm)	Local solution (mm)
p1	3.37	3.62
p2	2.93	2.63
p3	3.72	4.13
p4	1.58	1.53
p5	1.48	1.37
p6	1.53	1.62
p7	4.32	4.63
p8	4.79	1.47

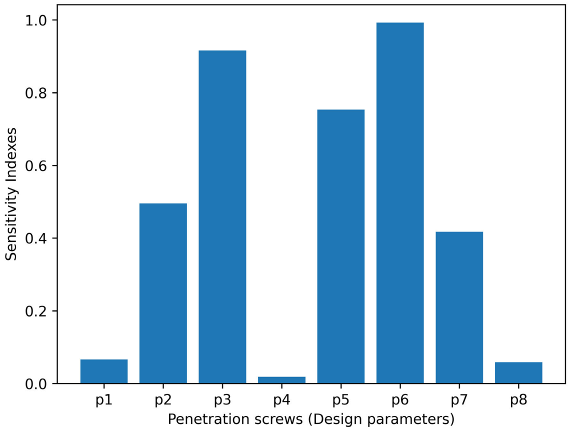


FIGURE 10. Sensitivity indexes of each penetration screws.

To analyze the sensitivity of each penetration screw to further understand the contribution to the optimization cost, a sensitivity report was constructed as shown in Fig.10. The first-order Sobol sensitivity analysis in (21) was adopted to visualize the perturbation impact of the penetration screws.

$$S_i = \frac{D_{\{i\}}}{D} \tag{21}$$

where $D_{\{i\}}$ is the contribution of a single depth of the penetration screws to the S-parameter responses. It is evident that the penetration screws 2, 3, 5, 6, and 7 have the highest impact, which is useful for understanding the priority of tuning which screws to adjust a highly detuned EPU filter.

C. COMPARISON BETWEEN RECENT WORKS AND THE PROPOSED METHOD

A comparison of reinforcement learning optimization approaches between recent studies and the proposed method is presented in Table 5. This comparison provides insights into the similarities and differences in the approaches used across various works by other authors. Unlike the approach in [22], where a forward model replaces EM simulation to speed up the training time, our proposed method integrates an API to obtain simulation results adjusted for design parameters. The network policy in SAC + HER choose actions from a continuous range of possible values, rather than from a discrete set of values as implemented in DQN and DDQN methods [38]. The continuous actions are sampled from a probability distribution over the defined minimum and maximum boundaries, making it suitable for complex optimization tasks where very small precision in parameter value is important.

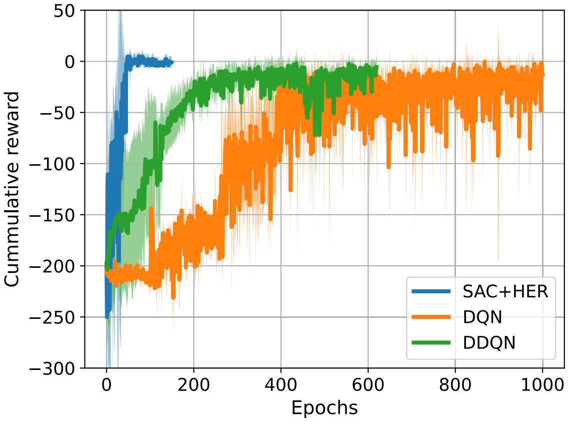


FIGURE 11. Comparison of training convergence between DQN, DDQN, and SAC+HER framework.

The proposed method focuses on using SAC + HER to handle non-uniqueness and sparse reward problems, thereby improving sample efficiency for faster training convergence. The result of training convergence is compared between DQN, DDQN and SAC + HER in the optimization of the EPU design as shown in Fig.11. It is evident that DQN requires approximately 700 epochs (35,000 evaluations) to reach convergence, which is 11 times longer than SAC + HER. Additionally, DDQN requires approximately 550 epochs (27,500 evaluations) to achieve convergence, which is 9 times longer than SAC + HER.

In Table 6, a comparison of the effectiveness of the trained DQN, DDQN, and SAC + HER models in solving

TABLE 5. Comparison of reinforcement learning optimization approaches between the recent works and the proposed method.

Reference	[22]	[23]	[24], [35]	[26]	This paper
Proposed Framework	DQN	DDQN	DQN	DQN	SAC + HER
Design Sample	Microstrip BPF	Cavity BPF	Cavity BPF	Cavity BPF	Cavity BPF with transmission zeros
EM analysis	Forward model	EM simulation	EM simulation	EM simulation	EM simulation
Initial Guess	Predicted by an inverse model	Random	Filter designer	Random	Random
Replay Buffer	Experience replay	Experience replay	Experience replay	Experience replay	Hindsight experience replay
Action Space	Discrete	Discrete	Discrete	Discrete	Continuous

TABLE 6. Comparison of the effectiveness of the trained DQN, DDQN, and SAC + HER models to solve the optimization problem.

RL Algorithm	DQN	DDQN	SAC+HER
Convergence Epoch	700	550	60
Convergence Training Steps	35,000	27,500	3,000
Training Time	22 days	15 days	2 days
Trained Model Inference Steps	92	57	6

the optimization of the EPU filter is summarized. SAC + HER shows better performance by achieving optimal convergence in 3000 steps, which is taking lesser steps than DQN and DDQN. The inference of trained models shows that the SAC + HER method requires 6 steps to successfully optimize the EPU filter, which is fifteen times faster than DQN (92 steps) and nine times faster than DDQN (57 steps).

VI. CONCLUSION

In this study, a hybrid method of goal-oriented SAC and HER is proposed to optimize the self-couplings of an asymmetrical inline topology with transmission zeros by adjusting the depth of screw penetration on each resonator. A feature-assisted objective function, incorporating predistortion constants from the synthesis process, was applied to prioritize adjustments to pole positions, return loss within the pass-band, and the location of the transmission zeros. As a result, the SAC + HER framework effectively addresses non-linearity between frequency responses and structural parameters, handling issues of non-uniqueness and sparse rewards more efficiently. The optimization results of trained SAC + HER model demonstrates better performance compared to conventional methods.

ACKNOWLEDGMENT

The authors would like to thank FILPAL for providing the EPU filter design for the simulation and optimization.

REFERENCES

[1] R. J. Cameron, C. M. Kudsia, and R. R. Mansour, *Microwave Filters for Communication Systems: Fundamentals, Design, and Applications*. Hoboken, NJ, USA: Wiley, 2007.

[2] P. W. Wong and G. S. Ng, "A new class of inline microwave filter with transmission zeros," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2023, p. 740, doi: 10.1109/ims37964.2023.10188050.

[3] S. Tamiazzo and G. Macchiarella, "Synthesis of cross-coupled filters with frequency-dependent couplings," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 3, pp. 775–782, Mar. 2017, doi: 10.1109/TMTT.2016.2633258.

[4] S. Tamiazzo, G. Macchiarella, and F. Seyfert, "A true inline coaxial-cavity filter with two symmetric zeros," *IEEE Microw. Wireless Compon. Lett.*, vol. 31, no. 6, pp. 666–669, Jun. 2021, doi: 10.1109/LMWC.2021.3064201.

[5] S. Tamiazzo, G. Macchiarella, and F. Seyfert, "Path filters: A class of true inline topologies with transmission zeros," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 1, pp. 850–863, Jan. 2022, doi: 10.1109/TMTT.2021.3126861.

[6] Y. Yang, M. Yu, and Q. Wu, "Advanced synthesis technique for unified extracted pole filters," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 12, pp. 4463–4472, Dec. 2016, doi: 10.1109/TMTT.2016.2623618.

[7] Y. Yang, M. Yu, and Q. Wu, "Advanced synthesis technique for extracted pole and NRN filters," in *IEEE MTT-S Int. Microw. Symp. Dig.*, May 2016, pp. 1–4, doi: 10.1109/MWSYM.2016.7540237.

[8] Y. Yang, M. Yu, Q. Wu, X. Yin, and J. Yang, "A fully integrated multiplexer using unified extracted pole technique," *IEEE Trans. Microw. Theory Techn.*, vol. 68, no. 8, pp. 3439–3447, Aug. 2020, doi: 10.1109/TMTT.2020.2996246.

[9] C. L. Ng, S. Soeung, S. Cheab, and K. Y. Leong, "A modified vector fitting technique to extract coupling matrix from S-parameters," *Radioengineering*, vol. 32, no. 3, pp. 325–331, Sep. 2023, doi: 10.13164/re.2023.0325.

[10] Q.-J. Zhang, K. C. Gupta, and V. K. Devabhaktuni, "Artificial neural networks for RF and microwave design-from theory to practice," *IEEE Trans. Microw. Theory Techn.*, vol. 51, no. 4, pp. 1339–1350, Apr. 2003.

[11] H. Kabir, Y. Wang, M. Yu, and Q.-J. Zhang, "Neural network inverse modeling and applications to microwave filter design," *IEEE Trans. Microw. Theory Techn.*, vol. 56, no. 4, pp. 867–879, Apr. 2008, doi: 10.1109/TMTT.2008.919078.

[12] H. Kabir, Y. Wang, M. Yu, and Q.-J. Zhang, "High-dimensional neural-network technique and applications to microwave filter modeling," *IEEE Trans. Microw. Theory Techn.*, vol. 58, no. 1, pp. 145–156, Jan. 2010, doi: 10.1109/TMTT.2009.2036412.

[13] S. Liu, F. Feng, J. Jin, S. Yan, N. Yan, Q.-J. Zhang, and K. Ma, "Inverse multivalued neural network modeling of fourth-order microwave filter," in *Proc. Int. Conf. Microw. Millim. Wave Technol. (ICMMT)*, May 2023, p. 1, doi: 10.1109/icmmt58241.2023.10277575.

[14] F. Feng, C. Zhang, J. Ma, and Q.-J. Zhang, "Parametric modeling of EM behavior of microwave components using combined neural networks and pole-residue-based transfer functions," *IEEE Trans. Microw. Theory Techn.*, vol. 64, no. 1, pp. 60–77, Jan. 2016, doi: 10.1109/TMTT.2015.2504099.

[15] M. Ohira, A. Yamashita, Z. Ma, and X. Wang, "A novel eigenmode-based neural network for fully automated microstrip bandpass filter design," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2017, pp. 1628–1631, doi: 10.1109/MWSYM.2017.8058947.

[16] Y. Wu, G. Pan, D. Lu, and M. Yu, "Artificial neural network for dimensionality reduction and its application to microwave filters inverse modeling," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 11, pp. 4683–4693, Nov. 2022.

[17] Z. Zhang, B. Liu, Y. Yu, M. Imran, Q. S. Cheng, and M. Yu, "A surrogate modeling space definition method for efficient filter yield optimization," *IEEE Microw. Wireless Technol. Lett.*, vol. 33, no. 6, pp. 631–634, Jun. 2023, doi: 10.1109/LMWT.2023.3243524.

[18] Z. Zhang, B. Liu, Y. Yu, and Q. S. Cheng, "A microwave filter yield optimization method based on off-line surrogate model-assisted evolutionary algorithm," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 6, pp. 2925–2934, Jun. 2022, doi: 10.1109/TMTT.2022.3163745.

[19] Y. Yu, Z. Zhang, Q. S. Cheng, B. Liu, Y. Wang, C. Guo, and T. T. Ye, "State-of-the-art: AI-assisted surrogate modeling and optimization for microwave filters," *IEEE Trans. Microw. Theory Techn.*, vol. 70, no. 11, pp. 4635–4651, Nov. 2022, doi: 10.1109/TMTT.2022.3208898.

[20] B. Liu, H. Yang, and M. J. Lancaster, "Global optimization of microwave filters based on a surrogate model-assisted evolutionary algorithm," *IEEE Trans. Microw. Theory Techn.*, vol. 65, no. 6, pp. 1976–1985, Jun. 2017, doi: 10.1109/TMTT.2017.2661739.

[21] S. Wu, W. Cao, M. Wu, and C. Liu, "A tuning method for microwave filter via complex neural network and improved space mapping," *Int. J. Electron. Commun. Eng.*, vol. 12, no. 3, p. 7, 2018.

[22] M. Ohira, K. Takano, and Z. Ma, "A novel deep-Q-network-based fine-tuning approach for planar bandpass filter design," *IEEE Microw. Wireless Compon. Lett.*, vol. 31, no. 6, pp. 638–641, Jun. 2021, doi: 10.1109/LMWC.2021.3062874.

[23] E. Sekhri, R. Kapoor, and M. Tamre, "Double deep Q-learning approach for tuning microwave cavity filters using locally linear embedding technique," in *Proc. Int. Conf. Mech. Syst. Mater. (MSM)*, Jul. 2020, pp. 1–6, doi: 10.1109/MSM49833.2020.9202393.

[24] Z. Wang, Y. Ou, X. Wu, and W. Feng, "Continuous reinforcement learning with knowledge-inspired reward shaping for autonomous cavity filter tuning," in *Proc. IEEE Int. Conf. Cyborg Bionic Syst. (CBS)*, Oct. 2018, pp. 53–58, doi: 10.1109/CBS.2018.8612197.

[25] M. Ohira, A. Yamashita, Z. Ma, and X. Wang, "Automated microstrip bandpass filter design using feedforward and inverse models of neural network," in *Proc. Asia–Pacific Microw. Conf. (APMC)*, Nov. 2018, pp. 1292–1294, doi: 10.23919/APMC.2018.8617627.

[26] E. Sekhri, M. Tamre, and R. Kapoor, "Optimal Q-learning approach for tuning the cavity filters," in *Proc. 20th Int. Conf. Res. Educ. Mechatronics (REM)*, Wels, Austria, May 2019, pp. 1–5, doi: 10.1109/REM.2019.8744118.

[27] M. Ohira, Y. Asai, and Z. Ma, "A deep-reinforcement-learning assisted microstrip BPF design approach for multiple specifications," in *Proc. Asia–Pacific Microw. Conf. (APMC)*, Dec. 2023, p. 288, doi: 10.1109/apmc57107.2023.10439910.

[28] W. Na and Q. Zhang, "Automated parametric modeling of microwave components using combined neural network and interpolation techniques," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2013, pp. 1–3, doi: 10.1109/MWSYM.2013.6697547.

[29] J. E. Rayas-Sánchez, S. Koziel, and J. W. Bandler, "Advanced RF and microwave design optimization: A journey and a vision of future trends," *IEEE J. Microw.*, vol. 1, no. 1, pp. 481–493, Jan. 2021, doi: 10.1109/JMW.2020.3034263.

[30] M. A. Ismail and M. Yu, "Advanced design of large scale microwave devices for space applications using space mapping optimization," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2017, pp. 1515–1516, doi: 10.1109/MWSYM.2017.8058914.

[31] J. E. Rayas-Sánchez and N. Vargas-Chávez, "A linear regression inverse space mapping algorithm for EM-based design optimization of microwave circuits," in *IEEE MTT-S Int. Microw. Symp. Dig.*, Jun. 2011, pp. 1–4, doi: 10.1109/MWSYM.2011.5972954.

[32] J. E. Rayas-Sanchez, "EM-based optimization of microwave circuits using artificial neural networks: The state-of-the-art," *IEEE Trans. Microw. Theory Techn.*, vol. 52, no. 1, pp. 420–435, Jan. 2004, doi: 10.1109/TMTT.2003.820897.

[33] C. Zhang, W. Na, Q. J. Zhang, and J. W. Bandler, "Fast yield estimation and optimization of microwave filters using a cognition-driven formulation of space mapping," in *IEEE MTT-S Int. Microw. Symp. Dig.*, May 2016, pp. 1–4, doi: 10.1109/MWSYM.2016.7539995.

[34] C. Zhang, F. Feng, V.-M.-R. Gongal-Reddy, Q. J. Zhang, and J. W. Bandler, "Cognition-driven formulation of space mapping for equal-ripple optimization of microwave filters," *IEEE Trans. Microw. Theory Techn.*, vol. 63, no. 7, pp. 2154–2165, Jul. 2015, doi: 10.1109/TMTT.2015.2431675.

[35] Z. Wang, J. Yang, J. Hu, W. Feng, and Y. Ou, "Reinforcement learning approach to learning human experience in tuning cavity filters," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2015, pp. 2145–2150, doi: 10.1109/ROBIO.2015.7419091.

[36] Y. Wang, Z. Zhang, Y. Yi, and Y. Zhang, "Accurate microwave filter design based on particle swarm optimization and one-dimensional convolution autoencoders," *Int. J. RF Microw. Comput.-Aided Eng.*, vol. 32, no. 4, Apr. 2022, Art. no. e23034, doi: 10.1002/mmce.23034.

[37] X. Yu, P. Xu, F. Wang, and X. Wang, "Reinforcement learning-based differential evolution algorithm for constrained multi-objective optimization problems," *Eng. Appl. Artif. Intell.*, vol. 131, May 2024, Art. no. 107817, doi: 10.1016/j.engappai.2023.107817.

[38] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," 2018, *arXiv:1801.01290*.

[39] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, P. Abbeel, and W. Zaremba, "Hindsight experience replay," 2017, *arXiv:1707.01495*.

[40] C. Florensa, D. Held, X. Geng, and P. Abbeel, "Automatic goal generation for reinforcement learning agents," 2017, *arXiv:1705.06366*.

[41] J. Bjorck, X. Chen, C. D. Sa, C. P. Gomes, and K. Q. Weinberger, "Low-precision reinforcement learning: Running soft actor-critic in half precision," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2021, pp. 980–991.

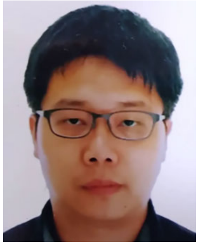
[42] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," 2016, *arXiv:1604.06778*.

[43] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, and S. Levine, "Soft actor-critic algorithms and applications," 2018, *arXiv:1812.05905*.

KIET YEW LEONG (Student Member, IEEE) received the B.Eng. degree in electrical and electronics engineering from Universiti Teknologi PETRONAS (UTP), Malaysia, in 2021, where he is currently pursuing the M.Sc. degree in electrical and electronics engineering. He is a Research and Development AI/ML Software Engineer with FILPAL Sdn., Bhd., Malaysia. His current research interests include computer-aided tuning and optimization techniques using both AI/ML and statistical approach, convex optimization, and multi-agent system based reinforcement learning optimization for microwave application.

SOCHEATRA SOEUNG (Senior Member, IEEE) received the B.Eng. degree in electrical and electronics major in computer system architecture and the M.Sc. and Ph.D. degrees in RF and microwave engineering from Universiti Teknologi PETRONAS, Malaysia. He was involved in designing, implementing, and testing RF subsystem components and RF communication link. He is currently a Lecturer and the Computation and Communication Cluster Leader of the Electrical and Electronics Engineering Department, Universiti Teknologi PETRONAS. Over the years, he has been a contributor of more than 35 technical research journals and conference papers. His research interests include RF microwave filter design and synthesis for multiband, multi-mode filter on planar and cavity structures, computer-aided tuning, and optimization techniques. He is also a MTT-S Member and an Associate Fellow of the ASEAN Academy of Engineering and Technology, and he serves as the Secretary for the IEEE ED/MTT/SSC Penang Chapter, Malaysia. He was awarded and funded as the Research Officer of RF microwave engineering under several Malaysian Ministry of Higher Education and industrial funding projects during his graduate study. He has been awarded with more than ten funding's from Malaysian Government, industries, and university research collaborations.

SOVUTHY CHEAB (Senior Member, IEEE) received the Bachelor of Engineering, Master of Science, and Ph.D. degrees in electrical and electronics engineering majoring in communication systems (telecommunication) from Universiti Teknologi PETRONAS (UTP), Malaysia, in 2011, 2012, and 2016, respectively. He has experience in teaching and conducting research in the field of telecommunications for over ten years. After completing his Ph.D. degree, he became a faculty member in the Electrical and Electronic Engineering Department, UTP, from 2016 to 2021. He was the Technical Director of FILPAL (M) Sdn., Bhd., for one year before returning to Cambodia, in 2022, to work as the Director of the Makerspace and Senior Researcher, Cambodia Academy of Digital Technology (CADT). Since 2024, he has been appointed as the Dean of the School of Digital Engineering, CADT. He has published more than 50 technical research articles. He supervised a total of six Ph.D. and seven master's students. His research interests include RF and microwave engineering and the Internet of Things. He served as an Executive Committee Member for the IEEE ED/MTT/SSC Penang Chapter, from 2016 to 2023, and was elected as the Secretary of the Chapter, from 2021 to 2022. He is an Associate Fellow of the ASEAN Academy of Engineering and Technology and the Steering Committee of ASEAN IVO, a global alliance of ICT research and development institutes and universities in the ASEAN region and Japan.



CHENG-KAI LU (Senior Member, IEEE) received the B.S. and M.S. degrees in electronics engineering from Fu Jen Catholic University, Taipei, Taiwan, in 2001 and 2003, respectively, and the Ph.D. degree in engineering from The University of Edinburgh, U.K., in 2012.

After completing his studies, he was the Director of the Research and Development Division, Chyao Shiunn Electronic Industrial Company, Shanghai, China, before joining the Science and Technology Policy Research and Information Centre, National Applied Research Laboratories, Taiwan. He is currently a Faculty Member with the Department of Electrical Engineering, National Taiwan Normal University (NTNU), Taipei. Prior to his tenure at NTNU, he was a Faculty Member with the Electrical and Electronic Engineering Department, Universiti Teknologi PETRONAS (UTP), Malaysia, from 2016 to 2021. Alongside his academic pursuits, he boasts over eight years of industrial experience. He has not only published his research works on peer-reviewed papers (including book chapters, journal papers, conferences, and reports) but also has filed multiple patents. His most notable contributions is his work in production line automation, which significantly reduced labor costs in manufacturing during his tenure as the Director of the Research and Development Division, Chyao Shiunn Electronic Industrial Company, from 2012 to 2014. Additionally, his patents were successfully licensed, resulting in partial royalties from patent licensing agreements. Notably, one of his inventions (TW Patent: PS keyboard system, 2006) was adopted by Republic of China Air Force and has been applied to light aircraft and specific long-haul flight planes. His research interests include medical imaging, embedded systems, artificial intelligence, and their applications and clinical decision support systems. From January 2017 to February 2018, he served as an Executive for the IEEE EMBS Malaysia Chapter, and IEEE ED/MTT/SSC Penang Joint Chapter, and since 2018.