

RESEARCH ARTICLE

Enhancing Attention-Based Visual Processing With Noise-Boosted Activation Functions

XIAOYUE GU¹, YUHAO REN¹, FABING DUAN¹,
AND DEREK ABBOTT², (Life Fellow, IEEE)

¹Institute of Complexity Science, Qingdao University, Qingdao 266071, China

²Centre for Biomedical Engineering (CBME), School of Electrical and Electronic Engineering, The University of Adelaide, Adelaide, SA 5005, Australia

Corresponding author: Yuhao Ren (renyuhao@qdu.edu.cn)

The work of Derek Abbott was supported by Australian Research Council under Grant DP240103404 and Grant FL240100217.

ABSTRACT Motivated by the principle of stochastic resonance, we investigate the noise-boosted activations within both channel attention mechanisms of convolutional networks and gated linear unit (GLU)-based feedforward networks (FFNs) of Vision Transformers (ViTs) under attention-based visual processing frameworks. Specifically, we replace conventional ReLU or ReLU-based GLU (ReGLU) activations with noise-boosted variants, which incorporate learnable noise scale parameters during training. Experiments on the CIFAR-10 and STL-10 image classifications, Kvasir-SEG medical image segmentation, and Cityscapes semantic segmentation show significant improvements over conventional baselines across diverse attention architectures. The learnable noise scale parameters in activations converge to non-zero values after training, demonstrating the existence of stochastic resonance in deep attention mechanisms. These results indicate that controlled noise injection can enhance information transfer efficiency of neural networks, and establish a coherent framework that connects the theoretical principle of stochastic resonance with its practical applicability in attention-based visual processing.

INDEX TERMS Attention mechanism, noise-boosted activation, vision transformer, stochastic resonance, visual processing.

I. INTRODUCTION

Visual sensors have become fundamental components in modern intelligent systems [1], [2], and are widely applied in autonomous navigation [3], surveillance systems [4], industrial automation [5], and healthcare diagnostics [6]. With the growing demand for real-time processing and privacy-preserving computation, edge computing has become an important approach where visual processing tasks are performed locally on resource-constrained devices [7], [8]. Due to their practical limitations, edge deployments require neural network architectures to achieve optimal performance under constrained computational and memory resources [9].

In this context, attention mechanisms have become essential architectural modules in modern deep learning systems, enabling more effective utilization of available resources and enhancing overall model efficiency [10], [11], [12], [13].

These mechanisms can be broadly divided into two types: self-attention mechanisms for modeling long-range dependencies in sequential data, and visual attention mechanisms for selectively emphasizing salient features in computer vision tasks [10], [11]. Building upon self-attention, transformer architecture has revolutionized natural language processing [12], and has since been successfully adapted for computer vision through Vision Transformers (ViTs) [13]. Meanwhile, channel attention mechanisms significantly improved the performance of convolutional neural networks for image classification and segmentation [14], [15].

Among these studies on attention mechanisms [10], [11], [12], [13], [14], [15], [16], [17], [18], the design of activation functions is a particularly interesting research direction, as it can enhance the expression of learned representations without increasing the complexity of the network architecture [17], [18]. This perspective is closely related to research on controlled noise injection in neural networks, which acts as an implicit regularization mechanism by simultaneously

The associate editor coordinating the review of this manuscript and approving it for publication was Hengyong Yu¹.

constraining model complexity and improving generalization [19], [20]. While noise injection can be applied at various placements of neural networks [21], [22], [23], [24], [25], [26], such as on weights or gradients [27], [28], [29], [30], [31], [32], [33], modifying activation functions, including noise-boosted activations [34], represents a particularly effective approach for improving network performance [34], [35], [36], [37], [38], [39], [40], [41].

However, integrating such noise-boosted activations [34] with attention mechanisms remains unexplored. In this work, we further reformulate the fundamental modules of channel attention and self-attention mechanisms by replacing the conventional ReLU and ReLU-based gated linear unit (ReGLU) activations with noise-boosted activation functions. Moreover, it should be noted that this noise-boosted activation model, inspired by suprathreshold stochastic resonance principles [42], establishes a unified framework that integrates commonly used activations such as ReLU, GELU and SiLU [35], [36], [37], [38], [41], [43].

Unlike explicit noise injection methods that directly add noise to gradients [27], activations [33], or weights [44] during training, our approach uses cumulative distribution functions to construct activation functions with learnable noise scale parameters. This method provides more stable training than stochastic noise sampling and does not significantly increase computational cost during inference, while still enabling the network to learn optimal noise levels that induce stochastic resonance effects. Previous work has demonstrated the effectiveness of this approach across diverse applications including function approximation, image classification, and unsupervised learning tasks [34], [39], [40]. Based on the above analysis, we enhance attention mechanisms by replacing their activation functions with noise-boosted variants. The proposed approach does not rely on any specific architectural assumptions and can be applied to a wide range of attention mechanisms without requiring structural modifications. Therefore, our method is complementary to other attention enhancement techniques and can be integrated alongside existing improvements.

This work presents two fundamental contributions: First, we incorporate noise-boosted activation functions into attention mechanisms, exploiting their information selection properties to achieve stochastic resonance in both self-attention and channel attention. Second, unlike prior noise injection approaches that use explicit stochastic sampling [27], [30], [33], our cumulative distribution function parameterization enables learnable noise scales that remain active during inference, providing stable performance gains without the computational overhead of Monte Carlo sampling. The inherent feature weighting and information bottleneck constraints of attention architectures further amplify these benefits, making them particularly suitable for noise-induced enhancements.

Attention mechanisms exhibit inherent information selection properties that make them particularly suitable for noise-boosted activation functions [10], [11]. In self-attention, the

activation functions within feedforward networks process token representations after attention weighting, where controlled noise injection can enhance the robustness of feature transformations. In channel attention mechanisms, activation functions modulate the importance weights of different feature channels [14], [15], and the introduction of learnable noise parameters enables adaptive control of these channel dependencies. The information bottleneck constraints inherent to attention mechanisms [45], where selective information transmission determines network performance, create conditions under which stochastic resonance effects can improve the representation of salient features. This theoretical foundation indicates that attention-based architectures may exhibit stronger performance improvements from noise-boosted activations than conventional feedforward networks.

We implement experiments across four scenarios to evaluate the effectiveness of noise-boosted activation functions in attention mechanisms: ViTs with GLU-based feedforward networks (FFNs) for CIFAR-10 image classification, convolutional networks with channel attention mechanisms for STL-10 image classification, dual attention networks (DATTNets) for medical image segmentation on the Kvasir-SEG dataset, and semantic segmentation on the Cityscapes dataset [46], [47], [48], [49]. Specifically, we restructure the FFNs in ViTs into GLU variants, and replace the activation functions within these GLU structures with noise-boosted activations. The experimental results demonstrate significant performance improvements across different attention mechanisms. For CIFAR-10 classification using ViTs with GLU-based FFNs, replacing ReGLU with Rayleigh linear unit (RayLU) activations achieves a validation accuracy of 82.46% compared to 81.48% obtained with the ReGLU module. For STL-10 classification using networks with channel attention, the replacement of ReLU with SiLU activations achieves 64.50% accuracy versus 60.18% for ReLU, showing a remarkable gain of approximately 4.0%. For medical image segmentation on Kvasir-SEG, the network designed with CauchyLU activations achieves an intersection over union (IoU) of 66.81% compared to 64.78% for networks with ReLU, indicating an improvement of 2.03%. For Cityscapes semantic segmentation, networks with RayLU activations achieve a mean IoU of 51.53% compared to 50.45% for networks with ReLU activations, demonstrating a 1.08% improvement on this complex dataset.

To understand these network improvements, we also observe the convergent behavior of learnable noise scale parameters toward task-specific optimal configurations across all experiments. The convergence to non-zero noise scale values and the corresponding performance improvements in neural networks can be interpreted as the stochastic resonance phenomenon within attention mechanisms. These findings confirm the effectiveness of noise-boosted activation functions in attention-based neural networks, and indicate their broad applicability across diverse visual processing tasks.

II. ATTENTION MECHANISM WITH NOISE-BOOSTED ACTIVATIONS

The designed noise-boosted activations originate from the suprathreshold stochastic resonance model introduced in [42], i.e. an ensemble of T parallel threshold units subjected to a common input signal x and mutually independent noise components η_t for $t = 1, 2, \dots, T$. Each unit implements a threshold function

$$h(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases} \quad (1)$$

The collective response of this ensemble in the presence of additive noise yields $\bar{h}(x) = \frac{1}{T} \sum_{t=1}^T h(x + \eta_t)$. For sufficiently large T , the law of large numbers ensures that $\bar{h}(x)$ converges to the expected firing probability $p(x) = \int_{-x}^{\infty} f_{\eta}(\eta) d\eta = 1 - F_{\eta}(-x)$, where $f_{\eta}(\eta)$ and $F_{\eta}(u)$ represent the probability density function and cumulative distribution function of the noise η_t , respectively [34], [40]. We regard $p(x)$ as a stochastic gate that transmits the input x with probability $p(x)$ and outputs zero with probability $1 - p(x)$. This probabilistic gating yields a unified formulation of noise-boosted activation functions

$$g(x) = xp(x) = x[1 - F_{\eta}(-x)]. \quad (2)$$

This unified model of Eq. (2) provides a general framework that can generate various noise-boosted activation functions based on different noise distributions characterized by their cumulative distribution functions $F_{\eta}(\cdot)$.

To introduce a learnable noise scale parameter, we consider a scaled noise representation. Let $\tilde{\eta}$ denote a normalized noise variable with unit scale, define scaled $\eta = \sigma \tilde{\eta}$ where $\sigma > 0$ is the noise scale parameter. Substituting this definition into Eq. (2), the activation function becomes:

$$g(x, \sigma) = x[1 - F_{\tilde{\eta}}(-x/\sigma)], \quad (3)$$

where $F_{\tilde{\eta}}(\cdot)$ is the cumulative distribution function of the normalized noise $\tilde{\eta}$. This parameterization introduces only a single learnable parameter σ per layer, which controls the sharpness of noise-induced nonlinearity. As $\sigma \rightarrow 0$, the function $g(x, \sigma)$ approaches the ReLU activation, since $F_{\tilde{\eta}}(-x/\sigma) \rightarrow 0$ for $x > 0$ and $F_{\tilde{\eta}}(-x/\sigma) \rightarrow 1$ for $x < 0$. The parameter σ is learned during training and optimized jointly with the network weights via the stochastic gradient descent approach, allowing the network to automatically discover optimal noise levels that induce stochastic resonance effects. Thus, constructing networks with Eq. (3) does not significantly increase architectural complexity compared to standard ReLU networks [34], [40].

From the parameterized model in Eq. (3), we obtain the GELU activation [34], [35], [40]

$$g(x, \sigma) = x\Phi(x/\sigma) \quad (4)$$

under Gaussian noise, where $\Phi(x) = \frac{1}{2} + \frac{1}{2} \text{erf}(x/\sqrt{2})$ and σ denotes the noise scale parameter. Similarly, for

Rayleigh noise, the Rayleigh linear unit (RayLU) can be expressed as

$$g(x, \sigma) = \begin{cases} x, & x \geq 0, \\ xe^{-\frac{x^2}{2\sigma^2}}, & x < 0, \end{cases} \quad (5)$$

and for exponentially distributed noise, the exponential linear unit (ExLU) is given by

$$g(x, \sigma) = \begin{cases} x, & x \geq 0, \\ xe^{\frac{x}{\sigma}}, & x < 0. \end{cases} \quad (6)$$

For logistic-distributed noise, we obtain the SiLU (Swish) activation [35], [37]

$$g(x, \sigma) = \frac{x}{1 + e^{-\frac{x}{\sigma}}}. \quad (7)$$

If the noise is Cauchy distributed, the associated Cauchy linear unit (CauchyLU) takes the form

$$g(x, \sigma) = x \left[\frac{1}{\pi} \arctan\left(\frac{x}{\sigma}\right) + \frac{1}{2} \right]. \quad (8)$$

Finally, for Laplace-distributed noise, the Laplace linear unit (LaplaceLU) can be calculated as

$$g(x, \sigma) = \begin{cases} x \left(1 - \frac{1}{2}e^{-\frac{x}{\sigma}}\right), & x \geq 0, \\ \frac{1}{2}xe^{\frac{x}{\sigma}}, & x < 0. \end{cases} \quad (9)$$

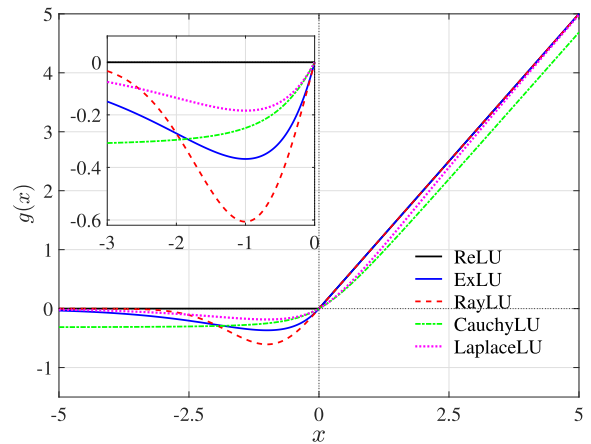


FIGURE 1. Plots of noise-boosted activation functions in Eqs. (4)–(9) ($\sigma = 1$) versus ReLU. The main plot depicts the input-output relationships, while the inset shows the negative input region. Compared to ReLU, the proposed activations exhibit smooth but non-monotonic characteristics.

Without loss of generality, we focus on the aforementioned cases of Eqs. (4)–(9) from a family of activations indicated in Eq. (2). Then, we further investigate how learnable noise scale parameters can be integrated into attention modules, and analyze whether they yield performance gains. Figure 1 illustrates the input-output characteristics of the proposed noise-boosted activation functions in Eqs. (4)–(9). In the positive input region, these functions exhibit linear

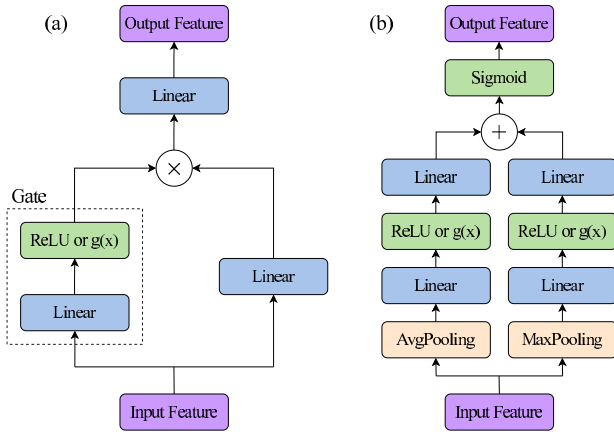


FIGURE 2. Attention modules integrated with noise-boosted activations: (a) GLU-based FFN module in ViTs and (b) channel attention module in convolutional neural networks.

growth similar to ReLU, while in the negative input region, they show different nonlinear behaviors, as shown in Fig. 1. The proposed noise-boosted activation functions are continuously differentiable across the entire real domain, while each activation function displays unique nonlinear characteristics determined by its underlying noise distribution.

As illustrated in Fig. 2, we incorporate the proposed noise-boosted activation functions into two representative attention frameworks: the GLU-based FFN module in ViTs and the channel attention mechanism in convolutional neural networks. For ViTs indicated in Fig. 2 (a), the proposed activations substitute the gating functions within the GLU structure of the feedforward module. By introducing the learnable noise scale parameter into the gating activation, the designed module can adaptively modulate information flow through the FFN, thereby enhancing the expressiveness of the learned representations. This also accords with recent work demonstrating that GLU variants employing gating activations can achieve performance improvements over conventional ReLU-based networks [18]. Following the design principle that channel attention operates globally while spatial attention works locally [15], we focus on modifying the channel attention component in visual attention mechanisms. The proposed activation functions can replace the conventional activations within channel attention pathways, where they process pooled feature representations before the computations of attention weights, as shown in Fig. 2 (b).

The effectiveness of noise-boosted activation functions is grounded in the principle of stochastic resonance, where controlled noise injection optimizes information processing in nonlinear threshold systems [42], [50]. In biophysical systems, stochastic resonance occurs when noise facilitates signal detection by enabling weak inputs to cross activation thresholds. The learnable parameter σ in our model serves an analogous function: it dynamically adjusts the effective threshold width of the activation function, similar to how

synaptic noise modulation in neural systems optimizes signal detection [50].

Following the derivation from Eqs. (1)–(3), when the ensemble size T is finite, the activation model directly corresponds to suprathreshold stochastic resonance with explicit noise injection. As $T \rightarrow \infty$, this converges to an implicit noise encoding scheme, where the noise scale parameter σ quantitatively controls the embedded noise intensity through the cumulative distribution function. The learning process of σ can be understood through information theory: backpropagation minimizes the loss function, which is equivalent to optimizing information flow for the given task [45], [51]. During training, the model automatically searches for the stochastic resonance point at each layer, where σ converges to values that maximize mutual information between layer representations and the target output, even for weak input signals such as low-contrast features in images.

This mechanism is particularly effective in attention modules. Conventional activation functions like ReLU output zero for negative inputs, which can suppress weak but important correlations that attention mechanisms seek to capture. By introducing learnable noise scales, our activation functions effectively implement a probabilistic threshold for attention weights, allowing gradients to flow more smoothly during training. The convergence of σ to task-specific optimal values enables the network to exploit stochastic resonance effects, enhancing the extraction of discriminative features through the attention mechanism. Our previous work on autoencoders [40] validated this framework, demonstrating consistent performance improvements and clear stochastic resonance phenomena characterized by optimal non-zero noise parameters and bell-shaped performance curves. We extend this approach to both self-attention and channel attention mechanisms in visual processing.

The noise scale parameters σ_j are optimized jointly with network weights through stochastic gradient descent. At each training iteration, the parameters are updated as

$$\sigma_j^{(t)} = \sigma_j^{(t-1)} - \alpha_\sigma \frac{\partial \mathcal{L}}{\partial \sigma_j} \bigg|_{\sigma_j = \sigma_j^{(t-1)}}, \quad (10)$$

where α_σ is the learning rate and $\mathcal{L}$ denotes the loss function. For the activation $g(x, \sigma_j) = x[1 - F_\eta(-x/\sigma_j)]$, the gradient computation involves the derivative of the cumulative distribution function, which naturally couples σ_j with the weight updates through the chain rule. This coupling provides an inherent mechanism preventing collapse to trivial solutions: when $\sigma_j \rightarrow 0$, the activation saturates but the gradient magnitude increases, pushing σ_j away from zero; when $\sigma_j \rightarrow \infty$, the activation becomes approximately linear, degrading performance and driving σ_j back down. For example, for the ExLU activation, the gradient with respect to σ is [40]

$$\frac{\partial g(x, \sigma)}{\partial \sigma} = -\frac{x^2}{\sigma^2} e^{x/\sigma}, \quad (11)$$

Algorithm 1 Training With Noise-Boosted Activations in Attention Mechanisms

Input: Training dataset $\mathcal{D}$, attention architecture $\mathcal{A}$, number of training epochs N_{epochs} .

Initialization: Network weights $\mathbf{W}$ via Xavier method, noise parameters σ_j for each layer j , learning rates α_W, α_σ .

Training Loop:

```

1: for epoch = 1 to  $N_{\text{epochs}}$  do
2:   for each minibatch  $\mathcal{B} \subset \mathcal{D}$  do
3:     for each attention layer  $j$  do
4:       Compute  $g(x, \sigma_j) = x \left[ 1 - F_\eta \left( -\frac{x}{\sigma_j} \right) \right]$ 
5:     end for
6:     Compute loss  $\mathcal{L}(\mathbf{W}, \{\sigma_j\})$ 
7:     Compute gradients  $\frac{\partial \mathcal{L}}{\partial \mathbf{W}}, \frac{\partial \mathcal{L}}{\partial \sigma_j}$ 
8:      $\mathbf{W} \leftarrow \mathbf{W} - \alpha_W \frac{\partial \mathcal{L}}{\partial \mathbf{W}}$ 
9:      $\sigma_j \leftarrow \sigma_j - \alpha_\sigma \frac{\partial \mathcal{L}}{\partial \sigma_j}, \forall j$ 
10:   end for
11: end for
12: return trained  $\mathbf{W}, \{\sigma_j\}$ 

```

which approaches infinity as $\sigma \rightarrow 0$ for $x > 0$, providing the gradient-based stabilization mechanism.

This gradient-based mechanism also provides theoretical guarantees for convergence in attention architectures. From a regularization perspective, noise injection introduces an implicit penalty on loss curvature. When the input is perturbed by noise ξ with variance σ^2 , the expected loss $\mathcal{L}$ can be approximated as $\mathbb{E}_\xi[\mathcal{L}(x + \xi)] \approx \mathcal{L}(x) + \frac{\sigma^2}{2} \text{Tr}[\nabla_{xx}^2 \mathcal{L}(x)]$, where $\nabla_{xx}^2 \mathcal{L}(x)$ denotes the Hessian matrix [52]. The second term acts as adaptive regularization that smooths the loss landscape. From an information-theoretic perspective, attention mechanisms perform selective information transmission between input features X and learned representations T . Information bottleneck theory characterizes this process through mutual information $I(X; T)$ and $I(Y; T)$, where Y denotes the target output [45], [51]. Learnable noise parameters induce stochastic resonance in the information objective, reaching optimal transmission at non-zero σ values rather than $\sigma = 0$. For attention modules, this mechanism is particularly beneficial because standard ReLU activations suppress all negative values with hard thresholds, potentially discarding weak but important correlations. Noise-boosted activations implement soft probabilistic thresholds through $p(x) = 1 - F_\eta(-x/\sigma)$, allowing attention mechanisms to preserve weak feature relationships while the learnable σ automatically adjusts the threshold characteristics during training.

In practice, we initialize all noise scale parameters σ_j to a constant value for each experiment. During training, each σ_j is constrained within $[0.01, 100]$ to prevent numerical instability. We employ a higher learning rate for σ_j than that used for the network weights, as the gradients of σ_j are typically smaller, requiring larger step sizes to achieve sufficient parameter adaptation during training.

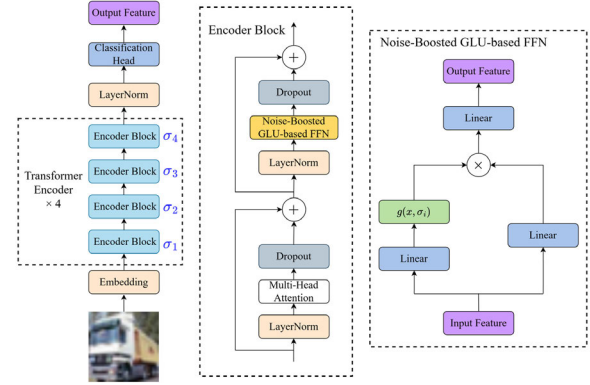


FIGURE 3. Architecture of the ViT model with GLU-based FFNs for CIFAR-10 dataset. This model consists of four transformer encoder blocks, each containing a multi-head self-attention and a GLU-based FFN. The noise-boosted FFN module (highlighted in yellow) incorporates independent learnable noise scale parameters ($\sigma_1, \sigma_2, \sigma_3, \sigma_4$) in the four encoder blocks.

The complete training procedure is outlined in Algorithm 1. The key distinction from standard training lies in the joint optimization of network weights $\mathbf{W}$ and noise scale parameters σ_j , allowing σ_j to converge to task-specific optimal values during training. Implementation details are available in our code repository [53].

III. EXPERIMENTAL RESULTS OF ATTENTION MECHANISMS

In this section, we evaluate the effectiveness of noise-boosted activation functions integrated into attention mechanisms. The evaluation is performed via classification tasks on the CIFAR-10 and STL-10 datasets and medical image segmentation on the Kvasir-SEG dataset. All experiments were implemented in PyTorch 2.5.1 on NVIDIA RTX 4090 GPU. Complete source code and implementation details are available in Ref. [53].

A. CIFAR-10 CLASSIFICATION VIA ViTs

To evaluate the performance of GLU-based architectures with different activations, experiments were implemented on the CIFAR-10 dataset using a ViT model. As illustrated in Fig. 3, the model consists of four transformer encoder blocks connected in series. Each encoder block contains a multi-head self-attention layer and a GLU-based FFN, with residual connections and layer normalization applied after each sub-layer. The GLU-based FFN uses two parallel branches combined by element-wise multiplication: one branch applies a linear layer followed by the noise-boosted activation $g(x, \sigma_j)$, while the other applies only a linear layer. Each encoder block has its own learnable noise scale parameter, denoted as $\sigma_1, \sigma_2, \sigma_3$, and σ_4 .

The model architecture comprised four transformer blocks with an embedding dimension of 192, three attention heads, a patch size of 4×4 , and a FFN expansion ratio of 4.0. Network parameters were initialized using Xavier uniform initialization. The designed model was trained using the

AdamW optimizer with a learning rate of 0.001, weight decay of 0.0001, batch size of 128, and a cosine annealing learning rate scheduler over 50 epochs. A dropout rate of 0.1 was applied to avoid overfitting. For noise-boosted activation functions, each of the four transformer blocks contained an independent noise scale parameter. The optimal values of noise scale parameters for each layer were obtained by the Bayesian optimization method [40] with a search range of [0.05, 10], and the experimental times is 10 with these optimized noise scale parameters to obtain mean performance metrics of the designed model.

We integrate noise-boosted activation functions into the GLU framework by replacing the standard activation with our designed functions. Following the naming convention in the literature where ReGLU denotes ReLU-based GLU [18], we refer to our variants as RayGLU, CauchyGLU, ExGLU, LaplaceGLU, GEGLU, and SwiGLU, with learnable noise scale parameter σ . For comparison, we also evaluate GEGLU and SwiGLU with fixed $\sigma = 1$ (denoted as “GEGLU ($\sigma = 1$ fixed)” and “SwiGLU ($\sigma = 1$ fixed)” in Table 1), corresponding to the standard configurations in the literature. Additionally, we include the ReGLU baseline.

TABLE 1. Performance comparison of ViT models with different activation functions and noise injection methods in GLU modules on the CIFAR-10 dataset. Each result represents the mean validation accuracy and loss $\pm$ standard deviation computed over 10 independent experimental runs.

Activation	Validation Accuracy (%)	Validation Loss
ReLU (baseline)	79.98 $\pm$ 0.31	0.6009 $\pm$ 0.0059
GELU (baseline)	80.32 $\pm$ 0.54	0.6063 $\pm$ 0.0111
ReGLU (baseline)	81.48 $\pm$ 0.43	0.5737 $\pm$ 0.0122
GEGLU ($\sigma = 1$ fixed)	81.72 $\pm$ 0.37	0.5827 $\pm$ 0.0133
SwiGLU ($\sigma = 1$ fixed)	81.61 $\pm$ 0.34	0.5799 $\pm$ 0.0075
RayGLU	82.46 $\pm$ 0.29	0.5605 $\pm$ 0.0071
CauchyGLU	82.01 $\pm$ 0.40	0.5673 $\pm$ 0.0099
ExGLU	81.88 $\pm$ 0.45	0.5694 $\pm$ 0.0119
GEGLU	81.83 $\pm$ 0.45	0.5778 $\pm$ 0.0115
LaplaceGLU	82.13 $\pm$ 0.33	0.5628 $\pm$ 0.0130
SwiGLU	81.87 $\pm$ 0.23	0.5638 $\pm$ 0.0059
Noisy ReLU [33]	81.64 $\pm$ 0.37	0.5666 $\pm$ 0.0099
Gradient Noise [27]	81.58 $\pm$ 0.45	0.5695 $\pm$ 0.0111
Weight Noise [44]	65.22 $\pm$ 0.39	0.5712 $\pm$ 0.0104
dropout variation	81.68 $\pm$ 0.28	0.5765 $\pm$ 0.0086

Table 1 reports the comparative performance of ViT models employing distinct activations within GLU-based architectures on the CIFAR-10 dataset. The baseline implementations comprise the conventional ReLU and GELU activations without the GLU structure, achieving accuracies of 79.98 $\pm$ 0.31% and 80.32 $\pm$ 0.54%, respectively, and the conventional ReGLU configuration yielding 81.48 $\pm$ 0.43%. For comparison, explicit noise injection methods including Noisy ReLU [33], gradient noise [27], weight noise [44], and dropout variation [30] achieve accuracies of 81.64 $\pm$ 0.37%, 81.58 $\pm$ 0.45%, 65.22 $\pm$ 0.39%, and 81.68 $\pm$ 0.28%, respectively. These results were obtained by employing Bayesian hyperparameter optimization to independently optimize the noise parameters under the same training protocol as the proposed methods. Additionally, we compare with standard GLU variants GEGLU and SwiGLU, which correspond

to our GEGLU and SwiGLU with noise scale parameter fixed at $\sigma = 1$, achieving 81.72 $\pm$ 0.37% and 81.61 $\pm$ 0.34%, respectively. As listed in Table 1, all GLU-based modules with learnable noise-boosted activation functions exhibit superior performance compared to baselines, explicit noise injection approaches, and fixed noise scale configurations. Among these, the ViT model with RayGLU modules attains the highest validation accuracy of 82.46 $\pm$ 0.29% and the lowest validation loss of 0.5605 $\pm$ 0.0071, corresponding to a 0.98% improvement over the ReGLU baseline. The ViT model with LaplaceGLU and CauchyGLU variants also demonstrate enhanced performance, achieving accuracies of 82.13 $\pm$ 0.33% and 82.01 $\pm$ 0.40%, respectively. In addition, all designed models incorporating noise-boosted activations yield smaller standard deviations across independent trials, implying improved training stability and reduced sensitivity to random initialization.

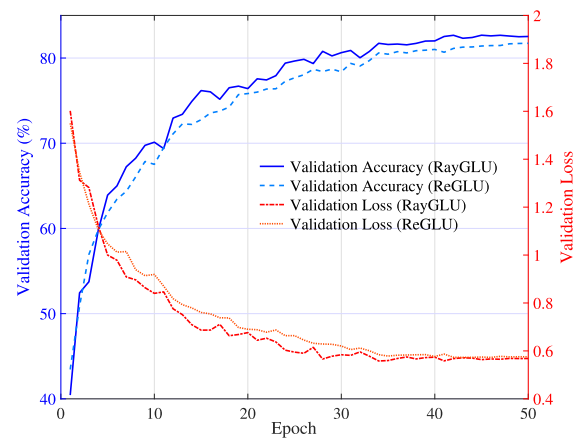


FIGURE 4. Training dynamics of GLU-based ViT models on CIFAR-10, comparing validation accuracy and loss for ReGLU versus RayGLU activations.

Figure 4 compares the training dynamics of ViT models with ReGLU and RayGLU modules on the CIFAR-10 dataset. Both architectures exhibit stable convergence throughout the training process. After 5 epochs, the ViT model with RayGLU modules begins to demonstrate superior performance, achieving higher validation accuracy and lower validation loss compared to its ReGLU counterpart. By the end of training at 50 epochs, the RayGLU-based network reaches a validation accuracy of approximately 82.5%, whereas the ReGLU-based network converges to around 81.5%. These results correspond to a representative single training run. The superiority of RayGLU over ReGLU from epoch 5 onward indicates faster convergence and more effective feature learning capabilities in the GLU architecture.

To further investigate the role of noise scale parameters in the GLU architecture, we perform a grid search analysis by plotting contour curves of the validation accuracy versus σ_1 (first transformer block) and σ_2 (second transformer block), while fixing σ_3 and σ_4 at the corresponding optimized values.

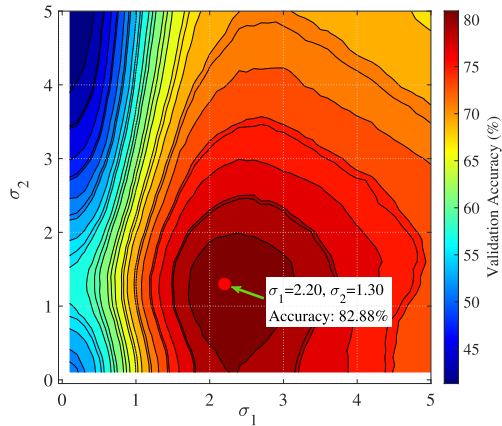


FIGURE 5. Contour plot of the validation accuracy versus noise scale parameters σ_1 and σ_2 in the RayGLU-based ViT model on the CIFAR-10 dataset. The noise scale parameters of σ_3 and σ_4 in the RayGLU-based ViT model are fixed at the corresponding optimized values $(\sigma_3, \sigma_4) = (2.70, 0.10)$. The maximum validation accuracy of 82.88% is achieved at the optimal point $(\sigma_1, \sigma_2) = (2.20, 1.30)$.

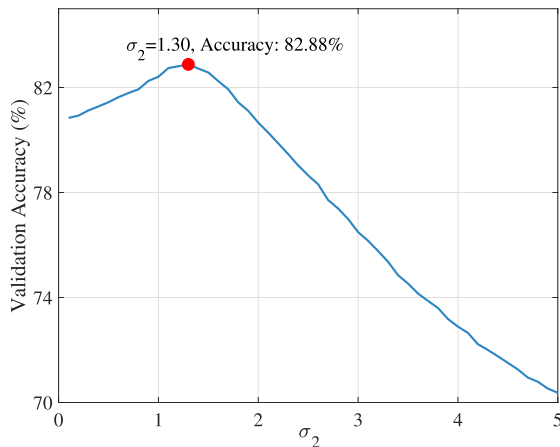


FIGURE 6. Performance versus noise scale parameter σ_2 with σ_1 fixed at 2.20. The curve shows a peak at $\sigma_2 = 1.30$ (accuracy: 82.88%). The noise scale parameters of σ_3 and σ_4 in the RayGLU-based ViT model are fixed at the corresponding optimized values $(\sigma_3, \sigma_4) = (2.70, 0.10)$. When $\sigma_2 = 0$, RayGLU reduces to ReGLU, yielding the baseline accuracy of 81.48%.

Figure 5 demonstrates that the optimal validation accuracy of 82.88% is achieved at non-zero values $\sigma_1 = 2.20$ and $\sigma_2 = 1.30$. When σ_1 or σ_2 approaches zero, the validation accuracy degrades rapidly. The variation of σ_1 has a larger impact on the model performance than σ_2 does, as indicated by the more gradual contour changes along the vertical axis. This asymmetric sensitivity pattern indicates that different transformer blocks require distinct optimal noise levels for feature learning.

To illustrate the stochastic resonance phenomenon more clearly, Fig. 6 shows the network performance as a function of σ_2 , with σ_1 fixed at 2.20. As the noise scale parameter σ_2 increases from zero, the validation accuracy first rises, reaches a peak of 82.88% at $\sigma_2 = 1.30$, and then decreases. When $\sigma_2 = 0$, the noise term vanishes and RayGLU

reduces to ReGLU, corresponding to the ReGLU baseline (81.48%) in Table 1. The performance improvement from the baseline to the optimal point shows that an appropriate amount of noise injection enhances network performance, while excessive noise degrades network performance. This peak behavior, combined with the 2D parameter space visualization in Fig. 5, confirms the occurrence of stochastic resonance in the GLU-based ViT architecture. The existence of non-zero optimal noise scale parameters validates the theoretical activation model of Eq. (2) derived from the suprathreshold stochastic resonance model [42], thereby extending classical resonance phenomena to the complex architecture of deep neural networks. Other combinations of noise scale parameters within the network also yield similar resonance characteristics, but these results are not shown here for brevity.

B. STL-10 CLASSIFICATION VIA CHANNEL ATTENTION MECHANISMS

As illustrated in Fig. 7, the experimental architecture employs a convolutional neural network with channel attention mechanisms for classification on the STL-10 dataset, which consists of 10 object categories with 5,000 training and 8,000 testing RGB images of 96×96 pixels [46]. Network parameters were initialized using Xavier uniform initialization, and all noise scale parameters were initialized to 3.0. The designed convolutional neural network is trained for 100 epochs using AdamW optimizer with learning rates at 0.001 for network weights and 0.01 for noise scale parameters. A dropout rate of 0.1 was applied to avoid overfitting. Experimental results are obtained over 10 independent runs with an exponential moving average, and performance metrics are recorded as the mean over the last 5 epochs to ensure convergence stability.

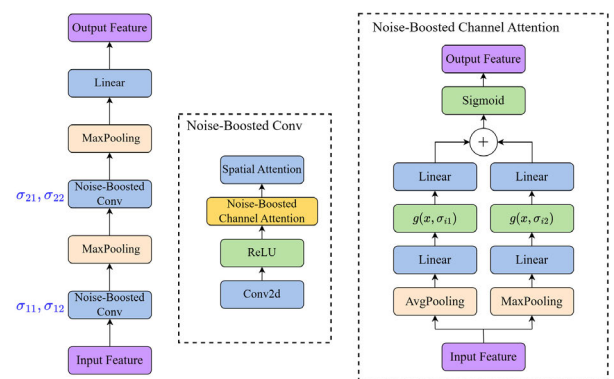


FIGURE 7. Architecture of the convolutional neural network with channel attention mechanisms for STL-10 classification. The left panel illustrates the overall network structure, consisting of two convolutional blocks followed by max pooling and a classification head. The middle panel presents the internal structure of a convolutional block, which incorporates a noise-boosted channel attention module shown on the right. Each convolutional block contains two mutually independent learnable noise scale parameters (σ_{i1} and σ_{i2} for $i = 1, 2$), corresponding to the activation functions in the two parallel branches of the noise-boosted channel attention module within the i -th block.

Table 2 presents a performance comparison between convolutional neural networks with channel attention mechanisms incorporating either conventional ReLU or noise-boosted activation functions. It is seen in Table 2 that networks with noise-boosted activations outperform the ReLU baseline across all configurations. The SiLU-based network achieves the highest classification accuracy of $64.50 \pm 1.36\%$, corresponding to a 4.32% improvement over the ReLU-based network ($60.18 \pm 1.64\%$). Networks incorporating other noise-boosted variants also demonstrate improvements of accuracy gains ranging from 1.29% to 2.77%. Furthermore, networks employing noise-boosted activations exhibit lower standard deviations in both classification accuracy and validation loss, indicating enhanced robustness and stability across independent experimental runs.

TABLE 2. Comparison of classification accuracy and validation loss for convolutional neural networks with channel attention using different activation functions on the STL-10 dataset.

Activation	Classification Accuracy (%)	Validation Loss
ReLU (baseline)	60.18 ± 1.64	1.08 ± 0.04
SiLU	64.50 ± 1.36	0.97 ± 0.04
CauchyLU	63.80 ± 1.69	0.99 ± 0.04
ExLU	61.57 ± 1.61	1.05 ± 0.04
GELU	62.95 ± 1.82	1.01 ± 0.05
LaplaceLU	62.79 ± 2.37	1.02 ± 0.06
RayLU	62.47 ± 1.18	1.02 ± 0.03

Figure 8 (a) compares the training dynamics of the noise-boosted convolutional network with ReLU and SiLU activations on the STL-10 dataset. As shown in Fig. 8 (a), after 20 epochs, the convolutional network with SiLU activations already achieves a higher classification accuracy on the validation set than the network with ReLU, and it reaches a validation accuracy of 64.5% after 100 epochs, whereas the ReLU convolutional network approaches 60.0%. A similar trend is observed in the cross-entropy loss on the validation set, where both networks exhibit comparable convergence behavior. Therefore, the superiority of the convolutional network with SiLU activations over that with ReLU is evident not only in the final performance but also throughout most of the training process. This advantage can be attributed to the adaptive learning capability endowed by the noise scale parameters.

Furthermore, we plot the learning curves of the four trainable noise scale parameters for the convolutional network with SiLU activations in Fig. 8 (b). Here, σ_{ij} denotes the noise scale parameter in the channel attention module of block i and branch j ($i, j = 1, 2$). With initial values of 3.0, it can be observed in Fig. 8 (b) that four noise scale parameters σ_{ij} continuously adapt throughout training and eventually reach stable values. However, none of them converges to zero, which indicates that the SiLU with these noise scale parameters does not degenerate into the ReLU activation function. This also indicates that the noise scale parameters

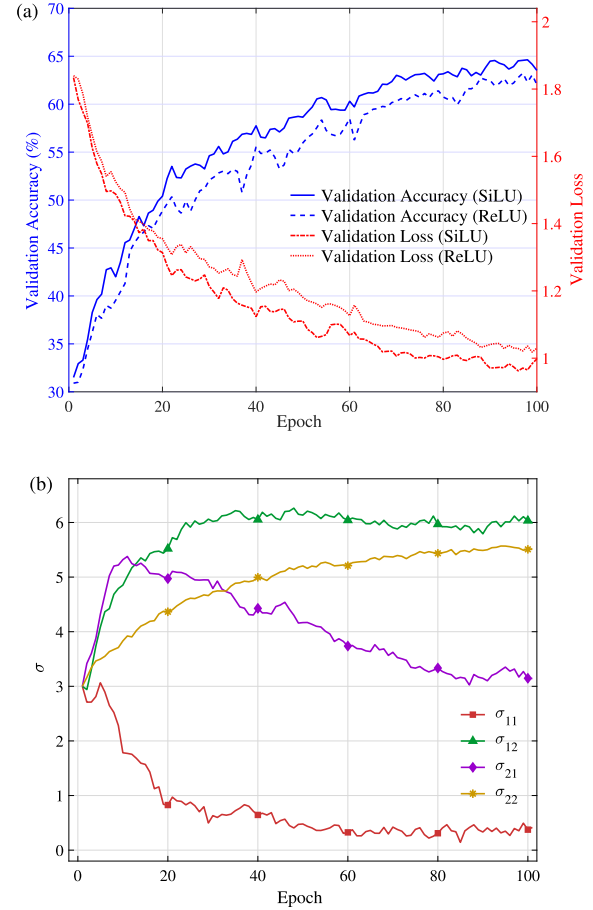


FIGURE 8. Training dynamics of convolutional networks on the STL-10 dataset. (a) Validation accuracy and loss of networks with ReLU versus SiLU activations. (b) Evolution of the learnable noise scale parameters σ_{ij} in the channel attention module of block i and branch j ($i, j = 1, 2$) for the SiLU network.

are non-negligible design variables of the neural network. The experimental results demonstrate that noise-boosted activation functions effectively enhance the performance of convolutional networks with channel attention on image classification tasks. The non-zero convergence values of learnable noise scale parameters provide evidence of the stochastic resonance effect in channel attention mechanisms, where controlled noise injection improves feature recalibration and network generalization.

To examine the learned attention patterns, we apply Gradient-weighted Class Activation Mapping (Grad-CAM) [54] on the STL-10 test set. Grad-CAM uses gradients to generate spatial importance maps, showing which image regions are most relevant for classification.

Figure 9 shows the attention maps from networks with ReLU and SiLU activations in the channel attention module. The SiLU-enhanced network produces more focused attention on target objects. In particular, the attention aligns better with object boundaries. These results indicate that noise-boosted activations improve feature extraction in attention mechanisms.

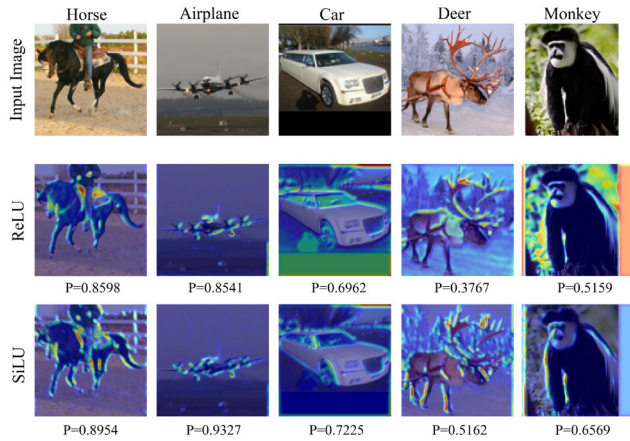


FIGURE 9. Grad-CAM [54] visualization on STL-10 test set. The attention maps are generated from the first convolutional layer for networks with ReLU and SiLU activations in the channel attention module. Ground-truth labels are shown above each image. P denotes the softmax probability for the correct class.

TABLE 3. Computational cost comparison of ReLU and SiLU based convolutional networks on the STL-10 classification task.

Metric	ReLU	SiLU
Total parameters	161,294	161,298
σ parameters	0	4
Model size (MB)	0.62	0.62
Training time (min)	8.29	8.42
Batch-32 inference (ms)	3.081	3.085
Throughput (images/s)	10,386	10,374
Peak GPU memory (MB)	947.05	947.07

Beyond the performance improvements, we also evaluate the computational efficiency of noise-boosted activations. Table 3 compares convolutional networks using SiLU with convolutional networks using ReLU on STL-10 classification. The noise-boosted activation introduces only four additional learnable parameters for the proposed network, with no significant increase in training time, inference latency, or memory usage.

C. MEDICAL IMAGE SEGMENTATION ON KVASIR-SEG DATASET

To validate the broad applicability of noise-boosted activation functions beyond classification tasks, we carried on an image segmentation experiment on the Kvasir-SEG polyp segmentation dataset [48]. This dataset consists of 1,000 polyp images with corresponding ground truth masks, divided into 4 : 1 training and validation sets. All images were pre-processed to 128×128 resolution using deterministic transformations.

The experiments utilized the DATNet [55] architecture with a VGG16 encoder backbone, noise-boosted channel attention modules, a context fusion bridge, and a skip-connected decoder. Each channel attention module contains a learnable noise scale parameter σ_i shared between the average pooling and max pooling branches for $i = 1, 2, \dots, 5$.

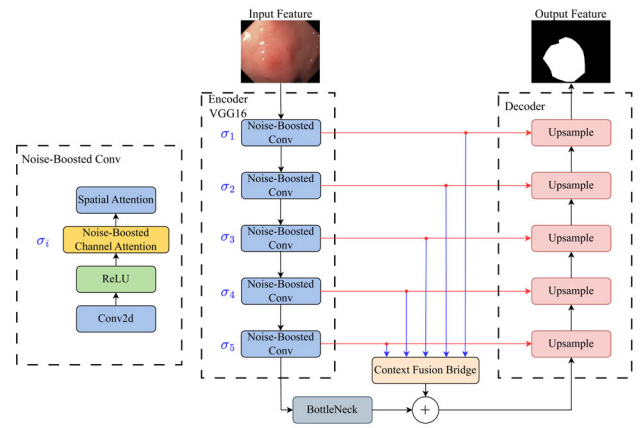


FIGURE 10. DATNet architecture with a VGG16-style encoder (blue blocks on the left) and an upsampling decoder (pink blocks on the right). Each encoder stage contains a noise-boosted channel attention module with learnable noise scale parameter σ_i shared between average and max pooling branches. The context fusion bridge between encoder and decoder integrates multi-scale features to enhance bottleneck representation.

The VGG16 encoder has 5 convolutional blocks. Activation functions were replaced only in channel attention modules, while ReLU was maintained in other network layers. Network parameters were initialized using Xavier uniform initialization, and all noise scale parameters were initialized to 5.0. Training was performed for 80 epochs with a batch size of 16, using the Adam optimizer and a combined binary cross entropy and dice loss. The learning rates were set at 0.0001 for network weights and 0.01 for noise scale parameters. A dropout rate of 0.1 was applied to avoid overfitting. Experimental results of the IoU and the validation loss were obtained by exponential moving averaging over the last 10 epochs.

TABLE 4. Results of the intersection over union (IoU) and the validation loss in channel attention for Kvasir-SEG polyp segmentation.

Activation	IoU (%)	Validation Loss
ReLU (baseline)	64.78 ± 0.97	0.2767 ± 0.0124
CauchyLU	66.81 ± 0.83	0.2597 ± 0.0142
ExLU	65.00 ± 1.01	0.2729 ± 0.0099
GELU	64.52 ± 1.65	0.2795 ± 0.0190
LaplaceLU	64.22 ± 1.93	0.2826 ± 0.0242
SiLU	66.00 ± 1.16	0.2597 ± 0.0125
RayLU	66.32 ± 1.42	0.2597 ± 0.0142

Table 4 presents the intersection over union (IoU) and the validation loss in channel attention with different activations for Kvasir-SEG polyp segmentation. It is observed that the DATNets with the six noise-boosted activation functions consistently outperformed the ReLU baseline in IoU, while generally also exhibiting a lower validation loss. The best IoU of 66.81% was achieved by the CauchyLU DATNet, and this represents a 2.03% improvement over the ReLU baseline (64.78%). Other noise-boosted DATNets demonstrated gains between 0.44% to 1.54%.

Analysis of the standard deviation indicates that the CauchyLU DATNet demonstrated slightly better stability than the ReLU baseline, achieving a standard deviation of 0.83% versus 0.97%.

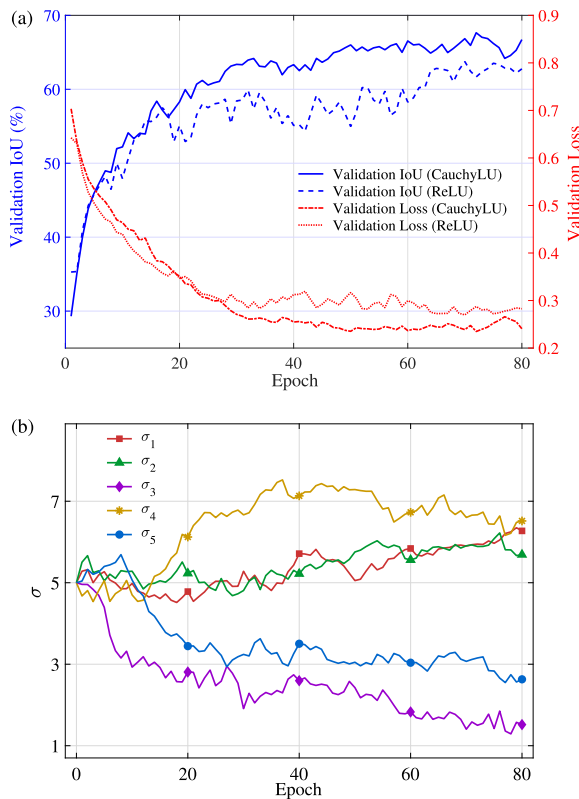


FIGURE 11. Training dynamics of DATNets on Kvasir-SEG dataset. (a) Validation accuracy and loss of networks with ReLU versus CauchyLU activations. (b) Evolution of learnable noise scale parameters across different network layers during training.

Figure 11 (a) compares the training dynamics of the DATNets with ReLU and CauchyLU activations on the Kvasir-SEG dataset. As shown in Fig. 11 (a), after 10 epochs, the network with CauchyLU activations achieves higher IoUs on the validation set than the one with ReLU, and finally converges to approximately 67.0% IoU after 80 epochs, whereas the ReLU-based network reaches an IoU around 65.0%. The validation loss of the CauchyLU network is also relatively lower. Figure 11 (b) shows the learnable curve of noise scale parameters across different network layers during training. It can be observed that the noise scale parameters also converge to distinct non-zero values, indicating that each layer adaptively learns the optimal noise characteristics suited to its specific role in the segmentation network.

Table 5 compares the performance of learnable versus fixed noise scale parameters for the CauchyLU activation. For the learnable configuration, all five noise scale parameters are independently optimized during training. For the fixed configurations, all five parameters are set to the same constant value. The learnable noise scale parameter achieves the

TABLE 5. Ablation study comparing learnable versus fixed noise scale parameters in DATNets with CauchyLU activation on the Kvasir-SEG dataset.

Configuration	IoU (%)
Learnable σ (proposed)	66.81 ± 0.90
Fixed $\sigma = 0.1$	65.69 ± 1.20
Fixed $\sigma = 0.5$	65.25 ± 2.54
Fixed $\sigma = 1.0$	65.03 ± 2.51
Fixed $\sigma = 2.0$	63.54 ± 3.38
Fixed $\sigma = 3.0$	64.92 ± 3.35
Fixed $\sigma = 5.0$	64.95 ± 2.20
Fixed $\sigma = 7.0$	66.36 ± 1.65

highest IoU of 66.81% and the lowest standard deviation of 0.90%. In contrast, networks with fixed noise scales demonstrate reduced performance across all tested values. The best fixed configuration ($\sigma = 7.0$) reaches 66.36% IoU but still underperforms the learnable approach. The learnable parameter also demonstrates superior stability, with standard deviation approximately half that of most fixed configurations.

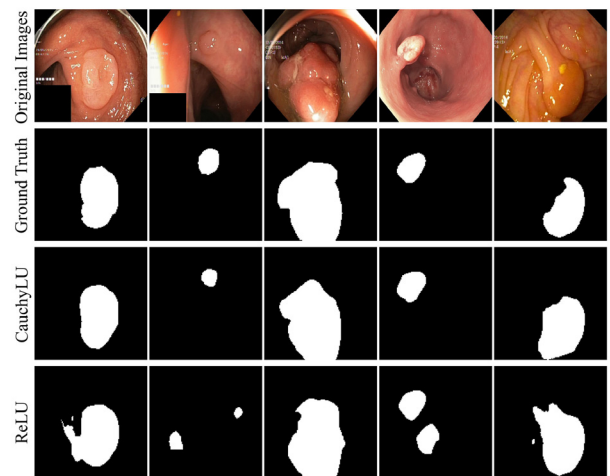


FIGURE 12. Representative segmentation samples of DATNets with ReLU and CauchyLU activations on the Kvasir-SEG dataset. The original endoscopic images and their corresponding ground truth binary masks are also shown.

Fig. 12 presents representative segmentation samples of DATNets with ReLU and CauchyLU activations on the Kvasir-SEG dataset. As we can see, DATNets with CauchyLU activation provide more accurate delineation of the polyp regions compared to those with ReLU, exhibiting improved boundary precision. The visual comparison further demonstrates that the noise-boosted activation function yields segmentation masks that align more closely with the ground truth annotations.

The Kvasir-SEG experiments demonstrate that integrating noise-boosted activation functions into channel attention modules leads to significant performance improvements in medical image segmentation. Among the evaluated activations, CauchyLU achieved the highest IoU of 66.81%,

exhibiting stable performance comparable to the ReLU baseline. The learnable noise parameters effectively adapt during training, with each layer converging to distinct optimal values. These results not only validate the effectiveness of noise-boosted activations beyond classification tasks but also provide evidence of their practical applicability in medical imaging contexts.

D. SEMANTIC SEGMENTATION ON CITYSCAPES DATASET

To evaluate the effectiveness of noise-boosted activation functions on more complex and diverse datasets, we conducted semantic segmentation experiments on the Cityscapes dataset [49]. Cityscapes is a challenging urban street scene dataset containing 2,975 training images and 500 validation images with pixel-level annotations for 19 semantic classes. All images were resized to 512×512 resolution.

The network architecture follows the DATTNet design from Section III-C, utilizing channel attention modules with noise-boosted activation functions. To accommodate the computational requirements of the dataset, the VGG16 encoder was replaced with a simplified encoder consisting of four stages of standard convolutional blocks, each containing two 3×3 convolutional layers followed by batch normalization and ReLU activation. Noise-boosted activation functions were applied only within the channel attention modules, while maintaining ReLU in other network components. Network parameters were initialized using Xavier uniform initialization, and noise scale parameters were set to 5.0. Training was performed for 80 epochs with a batch size of 4, using the Adam optimizer with learning rates of 5×10^{-4} for network weights and 5×10^{-3} for noise scale parameters. Each result represents the mean and standard deviation over 10 independent experimental runs.

TABLE 6. Semantic segmentation results on the Cityscapes validation set for networks with noise-boosted activation functions in channel attention modules.

Activation	mIoU (%)	Validation Loss
ReLU (baseline)	50.45 ± 1.29	0.2406 ± 0.0154
RayLU	51.53 ± 1.05	0.2261 ± 0.0135
CauchyLU	50.69 ± 0.98	0.2296 ± 0.0126
GELU	50.53 ± 1.28	0.2417 ± 0.0158
ExLU	50.81 ± 1.12	0.2373 ± 0.0142
SiLU	51.14 ± 0.97	0.2210 ± 0.0119
LaplaceLU	50.75 ± 1.38	0.2301 ± 0.0165

Table 6 presents the mean intersection over union (mIoU) and validation loss for different activation functions on the Cityscapes validation set. Networks with noise-boosted activation functions achieve mIoU values ranging from 50.53% to 51.53%. The RayLU activation achieves the highest mIoU, corresponding to a 1.08% improvement over the ReLU baseline (50.45%). Other noise-boosted activations including SiLU, ExLU, and LaplaceLU also demonstrate improvements ranging from 0.08% to 0.69%. The results demonstrate that noise-boosted activation functions maintain

their effectiveness when applied to complex semantic segmentation datasets with challenging urban scene understanding requirements. The consistent performance improvements across small medical imaging datasets and complex autonomous driving datasets validate the broad applicability and robustness of the proposed approach.

IV. CONCLUSION

This paper integrates noise-boosted activation functions into attention mechanisms for visual processing tasks. The unified theoretical framework in Eq. (3) generates smooth activation functions from various noise distributions with learnable scale parameters. Experimental results on CIFAR-10, STL-10, Kvasir-SEG, and Cityscapes datasets demonstrate consistent performance improvements over ReLU baselines across image classification, medical image segmentation, and large-scale semantic segmentation tasks. The architecture-agnostic nature of this approach indicates broad applicability to different attention mechanisms and deep learning models.

The learnable noise scale parameters converge to non-zero values during training, providing experimental evidence for stochastic resonance effects in deep attention mechanisms. This convergence behavior validates that controlled noise injection enhances network performance by improving information flow in the negative input region where the non-monotonic behavior of noise-boosted activations facilitates feature learning.

Future work may explore efficient optimization strategies for noise scale parameters, investigate the applicability of noise-boosted activations in other domains such as reinforcement learning or time-series prediction, and examine the long-term effects of noise in deeper network architectures.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES

The authors declare that no generative AI was used in the writing of this manuscript. However, Perplexity AI was used as a productivity tool for assisting the debugging of the manuscript's $\text{\LaTeX}$ code.

REFERENCES

- [1] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [2] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: A brief review," *Comput. Intell. Neurosci.*, vol. 2018, no. 1, 2018, Art. no. 7068349.
- [3] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *J. Field Robot.*, vol. 37, no. 3, pp. 362–386, Apr. 2020.
- [4] G. Sreenu and M. A. Saleem Durai, "Intelligent video surveillance: A review through deep learning techniques for crowd analysis," *J. Big Data*, vol. 6, no. 1, pp. 1–24, Dec. 2019.
- [5] Z. Li, Y. Yan, X. Wang, Y. Ge, and L. Meng, "A survey of deep learning for industrial visual anomaly detection," *Artif. Intell. Rev.*, vol. 58, no. 9, p. 279, Jun. 2025.
- [6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.

- [7] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.
- [8] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.
- [9] Y. Wang, Y. Han, C. Wang, S. Song, Q. Tian, and G. Huang, "Computation-efficient deep learning for computer vision: A survey," *Cybern. Intell.*, vol. 1, no. 1, pp. 1–24, 2024.
- [10] G. Brauwers and F. Frasincar, "A general survey on attention mechanisms in deep learning," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3279–3298, Apr. 2023.
- [11] M.-H. Guo, T.-X. Xu, J. Liu, Z.-N. Liu, P.-T. Jiang, T. Mu, S.-H. Zhang, R. R. Martin, M. Cheng, and S. Hu, "Attention mechanisms in computer vision: A survey," *Comput. Vis. Media*, vol. 8, no. 3, pp. 331–368, 2022.
- [12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, Long Beach, CA, USA, 2017, pp. 5998–6008.
- [13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [14] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.
- [15] S. Woo, J. Park, J. Lee, and I. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, 2018, pp. 3–19.
- [16] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 2002.
- [17] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, "Transformers are RNNs: Fast autoregressive transformers with linear attention," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, 2020, pp. 5156–5165.
- [18] N. Shazeer, "GLU variants improve transformer," 2020, *arXiv:2002.05202*.
- [19] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization," *Neural Comput.*, vol. 7, no. 1, pp. 108–116, Jan. 1995.
- [20] A. Orvieto, A. Raj, H. Kersting, and F. Bach, "Explicit regularization in overparametrized models via noise injection," in *Proc. 26th Int. Conf. Artif. Intell. Statist.*, Valencia, Spain, 2023, pp. 7265–7287.
- [21] Y. Grandvalet, S. Canu, and S. Boucheron, "Noise injection: Theoretical prospects," *Neural Comput.*, vol. 9, no. 5, pp. 1093–1108, Jul. 1997.
- [22] K. Audhkhasi, O. Osoba, and B. Kosko, "Noise-enhanced convolutional neural networks," *Neural Netw.*, vol. 78, pp. 15–23, Jun. 2016.
- [23] B. Kosko, K. Audhkhasi, and O. Osoba, "Noise can speed backpropagation learning and deep bidirectional pretraining," *Neural Netw.*, vol. 129, pp. 359–384, Sep. 2020.
- [24] Z. Liao, Z. Wang, H. Yamahara, and H. Tabata, "Low-power-consumption physical reservoir computing model based on overdamped bistable stochastic resonance system," *Neurocomputing*, vol. 468, pp. 137–147, Jan. 2022.
- [25] Z. Liao, K. Ma, M. S. Sarker, H. Yamahara, M. Seki, and H. Tabata, "Overdamped Ising machine with stochastic resonance phenomena in large noise condition," *Nonlinear Dyn.*, vol. 112, no. 11, pp. 8967–8984, Jun. 2024.
- [26] X. Liu, H. Wang, Y. Meng, and M. Fu, "Classification of hyperspectral image by CNN based on shadow area enhancement through dynamic stochastic resonance," *IEEE Access*, vol. 7, pp. 134862–134870, 2019.
- [27] A. Neelakantan, L. Vilnis, Q. V. Le, I. Sutskever, L. Kaiser, K. Kurach, and J. Martens, "Adding gradient noise improves learning for very deep networks," in *Proc. 5th Int. Conf. Learn. Represent.*, Toulon, France, 2017. [Online]. Available: <https://openreview.net/pdf?id=rkjZ2Pcxe>
- [28] R. Kleinberg, Y. Li, and Y. Yang, "An alternative view: When does SGD escape local minima?" in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm, Sweden, 2018, pp. 2698–2707.
- [29] L. Holmstrom and P. Koistinen, "Using additive noise in back-propagation training," *IEEE Trans. Neural Netw.*, vol. 3, no. 1, pp. 24–38, Jan. 1992.
- [30] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [31] Y. Li and F. Liu, "Whiteout: Gaussian adaptive noise regularization in deep neural networks," in *Proc. 17th Int. Symp. Neural Netw.*, Cairo, Egypt, 2020, pp. 176–189.
- [32] A. Graves, "Practical variational inference for neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, Granada, Spain, 2011, pp. 2348–2356.
- [33] Ç. Gülçehre, M. Moczulski, M. Denil, and Y. Bengio, "Noisy activation functions," in *Proc. 33rd Int. Conf. Mach. Learn.*, New York City, NY, USA, 2016, pp. 3059–3068.
- [34] F. Duan, F. Chapeau-Blondeau, and D. Abbott, "Optimized injection of noise in activation functions to improve generalization of neural networks," *Chaos, Solitons Fractals*, vol. 178, Jan. 2024, Art. no. 114363.
- [35] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.
- [36] P. Ramachandran, B. Zoph, and Q. V. Le, "Searching for activation functions," 2017, *arXiv:1710.05941*.
- [37] S. Elfving, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Netw.*, vol. 107, pp. 3–11, Nov. 2018.
- [38] N. Ma, X. Zhang, M. Liu, and J. Sun, "Activate or not: Learning customized activation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Aug. 2021, pp. 8028–8038.
- [39] L. Duan, F. Duan, F. Chapeau-Blondeau, and D. Abbott, "Noise-boosted backpropagation learning of feedforward threshold neural networks for function approximation," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–12, 2021.
- [40] Y. Ren, F. Duan, F. Chapeau-Blondeau, and D. Abbott, "Self-gating stochastic-resonance-based autoencoder for unsupervised learning," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 110, no. 1, Jul. 2024, Art. no. 014107.
- [41] S. Safaei, Z. Woods, K. Rasheed, T. R. Taha, V. Safaei, J. B. Gutiérrez, and H. R. Arabnia, "SWAG: A novel neural network architecture leveraging polynomial activation functions for enhanced deep learning efficiency," *IEEE Access*, vol. 12, pp. 73363–73375, 2024.
- [42] N. G. Stocks, "Suprathreshold stochastic resonance in multilevel threshold systems," *Phys. Rev. Lett.*, vol. 84, no. 11, pp. 2310–2313, Mar. 2000.
- [43] A. Apicella, F. Donnarumma, F. Isgrò, and R. Prevete, "A survey on modern trainable activation functions," *Neural Netw.*, vol. 138, pp. 14–32, Jun. 2021.
- [44] M. Fortunato, M. G. Azar, B. Piot, J. Menick, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, and O. Pietquin, "Noisy networks for exploration," in *Proc. Int. Conf. Learn. Represent.*, Vancouver, BC, Canada, 2018, pp. 1–21. [Online]. Available: <https://openreview.net/forum?id=rywHCPkAW>
- [45] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Jerusalem, Israel, Apr. 2015, pp. 1–5.
- [46] A. Coates, A. Y. Ng, and H. Lee, "An analysis of single-layer networks in unsupervised feature learning," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, vol. 15, Fort Lauderdale, FL, USA, 2011, pp. 215–223.
- [47] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, Granada, Spain, 2011. [Online]. Available: http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf
- [48] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. D. Lange, D. Johansen, and H. D. Johansen, "Kvasir-SEG: A segmented polyp dataset," in *Proc. 26th Int. Conf. Multimedia Model.*, Daejeon, South Korea, 2020, pp. 451–462.
- [49] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 3213–3223.
- [50] M. D. McDonnell, N. G. Stocks, C. E. Pearce, and D. Abbott, *Stochastic Resonance: From Suprathreshold Stochastic Resonance To Stochastic Signal Quantization*. New York, NY, USA: Cambridge Univ. Press, 2008.
- [51] E. Song, C. An, Y. Ren, and F. Duan, "Understanding the benefits of noise in deep neural networks via the information bottleneck," *Meas. Sci. Technol.*, vol. 36, no. 5, 2025, Art. no. 056114.
- [52] C. An, F. Duan, F. Chapeau-Blondeau, and D. Abbott, "Exploring the bayes-oriented noise injection approach in neural networks," *Phys. Lett. A*, vol. 555, Sep. 2025, Art. no. 130804.

- [53] Y. Ren. *VisualAttentionSR2025*. Accessed: Feb. 12, 2026.[Online]. Available: <https://github.com/renyuhao-qdu/VisualAttentionSR2025>
- [54] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626.
- [55] M. Zhang, Y. Zhang, S. Liu, Y. Han, H. Cao, and B. Qiao, "Dual-attention transformer-based hybrid network for multi-modal medical image segmentation," *Sci. Rep.*, vol. 14, no. 1, Oct. 2024, Art. no. 25704.



XIAOYUE GU was born in China, in 2000. She received the B.S. degree in mathematics and applied mathematics from Chengdu University of Technology, in 2022. She is currently pursuing the master's degree in system science with Qingdao University. Her research interests include computer vision and machine learning.



YUHAO REN was born in China, in 1987. He received the B.Eng. degree in electronic information engineering from Qingdao University, Qingdao, China, in 2010, and the Ph.D. degree in system science from Qingdao University, in 2018. Since 2018, he has been a Lecturer with the School of Automation, Qingdao University. His research interests include signal processing and machine learning.



FABING DUAN was born in China, in 1974. He received the master's degree in engineering mechanics from China University of Mining and Technology, Beijing, in 1999, the Ph.D. degree in solid mechanics from Zhejiang University, China, in 2002. From 2002 to 2003, he was a Postdoctoral Fellow with the University of Angers, France. Since 2004, he has been with Qingdao University, China, and a Professor of system science. His research interests include signal processing and machine learning.



DEREK ABBOTT (Life Fellow, IEEE) was born in South Kensington, London, U.K. He received the B.Sc. degree (Hons.) in physics from Loughborough University, U.K. in 1982, and the Ph.D. degree in electrical and electronic engineering from The University of Adelaide, Australia, in 1997, under the supervision of Kamran Eshraghian and Bruce R. Davis. From 1978 to 1986, he was a Research Engineer with the GEC Hirst Research Center, London.

From 1986 to 1987, he was a VLSI Design Engineer with Austek Microsystems, Australia. Since 1987, he has been with The University of Adelaide, where he is currently a Full Professor with the School of Electrical and Electronic Engineering. His research programs span a number of areas of communications, stochastics, game theory, security, photonics, energy policy, biomedical engineering, and computational neuroscience. His research interests include the areas of multidisciplinary physics and electronic engineering applied to complex systems. He is a fellow of the Institute of Physics, U.K., and an Honorary Fellow of Engineers Australia. He received a number of awards, including the South Australian Tall Poppy Award for Science, in 2004, an Australian Research Council Future Fellowship, in 2012, the David Dewhurst Medal, in 2015, the Barry Inglis Medal, in 2018, the M. A. Sargent Medal for eminence in engineering, in 2019, and Australian Research Council Laureate Fellowship, in 2024. He has served as an Editor and/or a Guest Editor for a number of journals, including the *IEEE JOURNAL OF SOLID-STATE CIRCUITS*, *Journal of Optics B, Chaos, Fluctuation and Noise Letters*, *Royal Society OS*, *Proceedings of the IEEE*, and *IEEE PHOTONICS JOURNAL*. He has served on the board of *Proceedings of the IEEE*, and is on the Editorial Boards of *Scientific Reports* (Nature), *Frontiers in Physics*, *PNAS Nexus*, and *IEEE ACCESS*. He served on the IEEE Publication Services and Products Board, from 2019 to 2024, was the Editor-in-Chief (EIC) of *IEEE ACCESS* from 2019 to 2024, and has been on the Editorial Board of *IEEE ACCESS*, since 2015.

...