# Uncertainty-Aware Denoising Network for Artifact Removal in EEG Signals

Xiyuan Jin, Jing Wang, *Member, IEEE*, Lei Liu, and Youfang Lin

***Abstract*— The electroencephalogram (EEG) is extensively employed for detecting various brain electrical activities. Nonetheless, EEG recordings are susceptible to undesirable artifacts, resulting in misleading data analysis and even significantly impacting the interpretation of results. While previous efforts to mitigate or reduce the impact of artifacts have achieved commendable performance, several challenges in this domain still persist: 1) due to black-box skepticism, deep-learning-based automatic EEG artifact removal methods have been impeded from being applied in clinical environments. How to support reliable denoised EEG signals with high accuracy is important; and 2) effectively exploring valuable local and global information from contaminated contexts remains challenging. On the one hand, feature extraction and aggregation in prior works are often performed blindly and assumed to be accurate, which is not always the case. On the other hand, global contextual information is gradually modeled by local fixed single-scaled convolutional filters layer by layer, which is neither efficient nor effective. To address the above challenges, we propose an Uncertainty-aware Denoising Network (UDNet) with multi-scaled pooling attention for efficient context capturing. Specifically, we predict the aleatoric and epistemic uncertainty existing during the denoising process to assist in finding and reducing the uncertain feature representation. We further propose a simple yet effective architecture to capture local and global contexts at multiple scales. Our proposed method can serve as an effective metric for identifying low-confidence epochs that warrant deferral to human experts for further inspection and assessment. Experimental results on two public datasets show that the proposed model outperforms state-of-the-art baselines.**

***Index Terms*— Artifact removal, deep neural network, uncertainty estimation.**

## I. Introduction

ELECTROENCEPHALOGRAM (EEG) is a widely utilized method for detecting brain electrical activity, with applications for diagnosing various neurological pathologies [1], conducting cognitive science research [2], monitoring drivers [3], tracking health parameters [4], and constructing Brain-Computer Interfaces [5], among others. Compared to other brain signal acquisition methods, EEG has several advantages, such as being a non-invasive and user-friendly technique, having a high temporal resolution, and being more cost-effective. However, its weak amplitudes make it often contaminated by various noises, especially for physiological artifacts from recording systems, including but not limited to ocular artifacts, myogenic artifacts, and cardiac artifacts. Such physiological artifacts can significantly disrupt neural information, potentially leading to their misinterpretation as normal phenomena in practical applications like brain-computer interfaces [6]. Furthermore, they might mimic cognitive or pathological activities, thereby introducing biases into the visual interpretation and diagnosis of clinical research studies, including Alzheimer's disease [7], sleep pattern analysis [43], seizure detection [25], among others. Therefore, developing effective algorithms that can reduce the impact of artifacts in EEG recordings while simultaneously preserving neural information to the greatest possible extent, is of utmost importance.

While detecting and removing artifacts automatically in such applications presents a significant challenge due to the overlapping nature of artifacts with background EEG rhythms and target events in both the temporal and spectral domains. Moreover, differentiating artifacts from the desired signal becomes difficult as artifacts can exhibit considerable variations based on factors such as their origin, waveform shape, and frequency characteristics.

Traditional artifact removal algorithms have shown acceptable performance in various EEG-based applications. However, these algorithms are subject to certain limitations when applied to specific contexts [8], [10], [19]. In recent years, deep learning (DL) has emerged as a highly effective approach for automatic feature extraction and representation learning [20]. Consequently, significant research efforts have been dedicated to developing DL-based techniques for EEG artifact denoising [32], [33], [34], [35], [36], [37]. Compared to traditional models, DL-based approaches offer two major advantages. First, they exhibit universality, as their uniform architecture enables them to handle a wide range of artifact removal tasks without the need for the manual design of prior assumptions specific to a particular type of artifact. Second, DL models possess higher capacity, which results in substantial performance improvements. Despite these advantages, several challenges persist in the domain of EEG artifact removal when utilizing DL-based methods.

**C1: How to ensure the reliability of the denoising results.** The widespread use of DL-based EEG artifact removal methods in clinical settings has yet to take off due to the

Manuscript received 12 June 2023; revised 6 October 2023; accepted 4 November 2023. Date of publication 8 November 2023; date of current version 14 November 2023. This work was supported by the Fundamental Research Funds for the Central Universities under Grant 2023JBMC056. *(Corresponding author: Jing Wang.)*

The authors are with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China (e-mail: xiyuanjin@bjtu.edu.cn; wj@bjtu.edu.cn; lei_liu@bjtu.edu.cn; yflin@bjtu.edu.cn).

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2023.3330963, provided by the authors.
Digital Object Identifier 10.1109/TNSRE.2023.3330963

common criticism that DL models are 'black boxes', especially when applied to artificial intelligence in healthcare and medicine [21]. This raises questions about how much to trust their results when denoising real-world signals. Fortunately, uncertainty estimation can be used to gauge model reliability and is already employed extensively in many other applications [22], [23], [24]. For instance, uncertainty estimation has been widely explored and proven beneficial in various fields such as MRI reconstruction [22], image segmentation [23], and seizure prediction [24]. In these applications, uncertainty estimation not only aids in generating output results but also provides valuable confidence values, enabling better inference by agents. Furthermore, incorporating uncertainty measurement can result in more informed decisions and potentially improve the quality of predictions [24]. Despite its integral role in many domains, uncertainty estimation has received limited attention in the field of EEG artifact removal. Besides, existing studies merely treat uncertainty estimation as a regularization term, overlooking the exploration of relationships between uncertainty regions and confidence. The question of how to leverage the knowledge embedded in confident representations to improve uncertain ones remains open. Therefore, it is worth investigating the integration of uncertainty estimation into EEG denoising, while simultaneously addressing the aforementioned challenges.

**C2: How to restore accurate waveforms under extremely worse noisy context.** Leveraging convolution to extract contextual information to recover the severely damaged segments is a common choice. In fact, local contexts in EEG segments reflect the adjacent trend information, aiding the recognition of noisy positions. Current DL-based artifact removal methods model local contextual information by applying 1-D convolution with fixed kernel size layer by layer. However, single-scale kernel size is not suitable for different noisy segments. On the other hand, the global context reflects the long-term trend and potentially valuable information, which should also be taken into consideration. Correspondingly, the attention mechanism is famous for its high flexibility in modeling global dependencies, which has been widely applied to various domains [27], [30]. However, one of the key challenges in applying the attention mechanism lies in its inefficiency when dealing with long-time series, which primarily stems from the high computation and memory complexity.

To address the aforementioned challenges, we propose a pioneering approach called the Uncertainty-aware Denoising network (UDNet). UDNet focuses primarily on attaining precise denoising outcomes while concurrently providing accurate uncertainty estimation. To begin with, we incorporate an Uncertainty Estimation Module (UEM), which is responsible for assessing the combined aleatoric and epistemic uncertainty at each sampling point. Furthermore, we integrate a Feature Enhancement Module (FEM) to enhance the quality of hidden representation of denoised signals by capturing local and global contextual information using an efficient multi-scaled pooling-attention mechanism, since details from multi-scales help differently in signal recovery. By combining the UEM and FEM, our proposed UDNet can effectively leverage uncertainty information and improve the denoising process in accuracy and reliability.

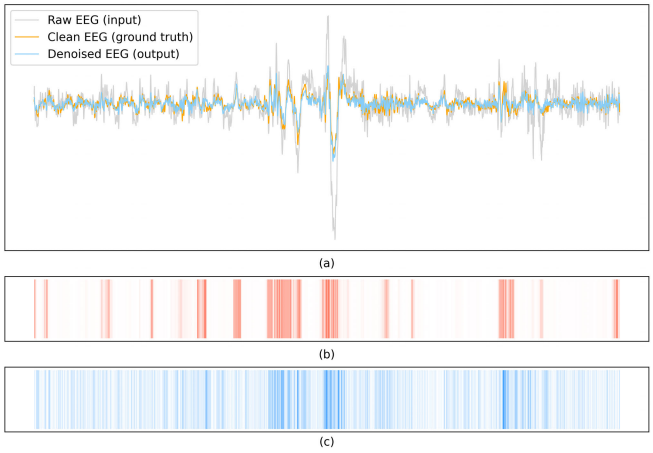


Fig. 1. The visualization of uncertainty maps and reconstruction errors (a) Denoising results. (b) Uncertainty map. (c) Error maps (i.e., the absolute value of the difference between predicted and true values $|\hat{y} - y|$). The deeper color in the uncertainty map and error map indicates a higher uncertainty over the corresponding original noisy signal or higher reconstruction error.

To the best of our knowledge, our proposed approach represents the first attempt to incorporate uncertainty estimation for EEG artifact removal. As depicted in Figure 1, our model has the capability to predict the uncertainty at each sampling point during the denoising process. Consequently, it can effectively identify uncertain regions that are more likely to contain significant reconstruction errors. By leveraging this information, we are able to achieve more compelling denoising results. To summarize, this paper makes the following contributions:

- We propose a novel Uncertainty-aware Denoising Network (UDNet) that aims to enhance the quality of features and generate reliable, noise-free signals.
- We propose an Uncertainty Estimation Module (UEM) to capture the aleatoric and epistemic uncertainty of each sampling-point-wise denoising results together.
- We develop Feature Enhancement Module (FEM) to adaptively enhance the learned features, which captures both the local and global contexts through a multi-scaled pooling-attention mechanism.
- We conduct extensive experiments on two real-world datasets to evaluate the performance of the proposed model. Meanwhile, we conduct an ablation study to prove that uncertainty quantification in UDNet is beneficial for promoting denoising accuracy.

## II. Related Work

### A. Traditional Artifact Removal

Traditional methods can be divided into two main categories: those that estimate artifactual signals using a reference channel and those that decompose the EEG signal into other domains.

Specifically, regression-based methods [8] estimate noise signals by utilizing noise templates and subtracting them from the EEG data. Adaptive filtering [9] adjusts weights iteratively using optimization algorithms to reduce the amount of artifactual contamination in the primary input. However, the reliance on reference channels to enhance the accuracy of artifact removal poses limitations for certain applications [10], [11].

Fourier transform and wavelet transform (WT) [15], [16] are used to map the signal from the time domain to the spectral domain, as EEG signals and artifacts often exhibit different spectral profiles. Wavelet quantile normalization [16] attenuates artifacts of different natures, requiring no auxiliary input, parameter tuning, or human intervention. Wavelet domain optimized Savitzky–Golay (WOSG) filtering approach [17] uses the optimized SG filter in the wavelet domain for the removal of motion artifacts. Dyadic boundary points based empirical wavelet transform (DBPEWT) [18] introduced an optimal transition width-based filter bank to decompose EEG time series into sub-band (SB) signals. However, due to the overlap between artifacts and the EEG spectrum [19], complete removal of artifacts may not be achievable, leading to the potential loss of neural information. Recent approaches have proposed hybrid methods such as EEMD-ICA [28] and EEMD-CCA [29], which combine traditional techniques [13], [14] to improve performance. However, these methods still do not address the limitations imposed by prior assumptions.

### B. Deep Learning Artifact Removal

With the emergence of deep learning, profound denoising models have been developed to tackle the challenge of EEG artifact removal. One such model is the 1D-ResCNN [34], which is a one-dimensional residual convolutional neural network. It constructs a regression model capable of capturing the complex and intricate nonlinear relationship between noisy and clean EEG signals. DeepSepeparator [33], an extension of linear blind source separation methods, is designed to learn the decomposition of the clean EEG signal and artifacts within a latent space. GRUMARSC [37] focuses on identifying the most relevant artifact pattern by utilizing an attention-based adaptive feature selection mechanism to prevent erroneous reconstruction of contaminated signals. Compared to traditional methods, deep learning models offer significant advantages in terms of their universality and high capacity. However, the widespread adoption of deep learning in EEG denoising has been somewhat limited due to concerns regarding weak interpretability and safety. Therefore, there is a growing interest in developing interpretable and reliable deep learning models specifically tailored for EEG denoising.

### C. Uncertainty Estimation

In general, uncertainty in EEG artifact removal can be categorized into two types: aleatoric uncertainty (data uncertainty) and epistemic uncertainty (model uncertainty) [39]. Data uncertainty relates to the inherent noise present in the EEG signals. On the other hand, model uncertainty captures the uncertainty associated with the model parameters and can be reduced by increasing the number of training samples. Bayesian neural networks (BNNs) and their variants are commonly used to model epistemic uncertainty by introducing probability distributions over model parameters [40], [41]. However, these methods often require different training techniques for the neural network and may introduce additional model parameters, sometimes even doubling the parameter count. Gal et al. [42] proposed the Monte Carlo dropout framework (MC-Dropout), which can be directly applied to a pre-trained model. It involves applying stochastic dropouts after each hidden layer and treating the output as a random sample generated from the posterior predictive distribution. In our approach, we propose a variant of MC-dropout and focus on capturing the epistemic uncertainty of the noisy signal representation in each layer.

## III. Preliminaries

### A. Problem Statement

Let $D = \{X, Y\} = \{x_i, y_i\}_{i=1}^{N}$ be a training dataset, $y_i \in \mathbb{R}^{L\times d}$ is the cleaned EEG signal for an input noisy EEG signal $x_i \in \mathbb{R}^{L\times d}$, where $L$ denotes the length for each time series sample, $d$ denotes the number of channels of interest. Our primary objective is to learn a transformation function $f$, which is parameterized by weights $\omega$ and maps a given input $x$ to a cleaned EEG $\hat{y}$ and the associated uncertainty $\hat{\sigma} \in \mathbb{R}^{L\times d}$, denoted as $x \overset{f(\cdot)}{\rightarrow} [\hat{y}, \hat{\sigma}]$.

### B. Bayesian Inference and Uncertainty Modeling

We define our likelihood as a Gaussian with mean given by the model output: $p(y|f^{\omega}(x)) = \mathcal{N}(f^{\omega}(x), \sigma^2)$, with an observation noise scalar $\sigma$. In the inference phase, given a test sample $x^*$, the predictive probability $y^*$ is computed by:

$$p(y^*|x^*, D) = \int p(y^*|x^*, \omega)p(\omega|D)d\omega \tag{1}$$

where the posterior $p(\omega|D)$ is intractable and cannot be computed analytically. A variational posterior distribution $q_\theta(\omega)$, where $\theta$ are the variational parameters, is used to approximate the true posterior distribution by minimizing the Kullback-Leilber (KL) divergence between $p(\omega|D)$ and $q_\theta(\omega)$, resulting in the approximate predictive distribution

$$p(y^*|x^*, D) = \int p(y^*|x^*, \omega)q_\theta(\omega)d\omega \tag{2}$$

Minimizing the Kullback–Leibler divergence is equivalent to maximizing the log evidence lower bound,

$$\mathcal{L}_{\text{VI}} = \int q_\theta(\omega)p(Y|X, \omega)d\omega - \mathbf{KL}[q_\theta(\omega)||p(\omega)] \tag{3}$$

With the re-parametrization trick [45], a differentiable mini-batched Monte Carlo estimator can be obtained.

The predictive (epistemic) uncertainty can be measured by performing T inference runs and averaging predictions:

$$p(y^*|x^*, D) \approx \int p(y^*|x^*, \omega)q_\theta(\omega)d\omega \approx \frac{1}{T}\sum_{t=1}^{T} p(y^*|x^*, \hat{\omega}_t) \tag{4}$$

where $T$ corresponds to the number of sets of mask vectors from Bernoulli distribution in MC-dropout, or the number of randomly trained models in Ensemble, which potentially leads to different set of learned parameters $\omega = \{\omega_1, \ldots, \omega_t\}$.

### C. Scaled Dot-Product Attention

Scaled Dot-Product Attention [44] is a type of attention function that calculates the weights by taking the dot-product between queries and values, which offers benefits such as

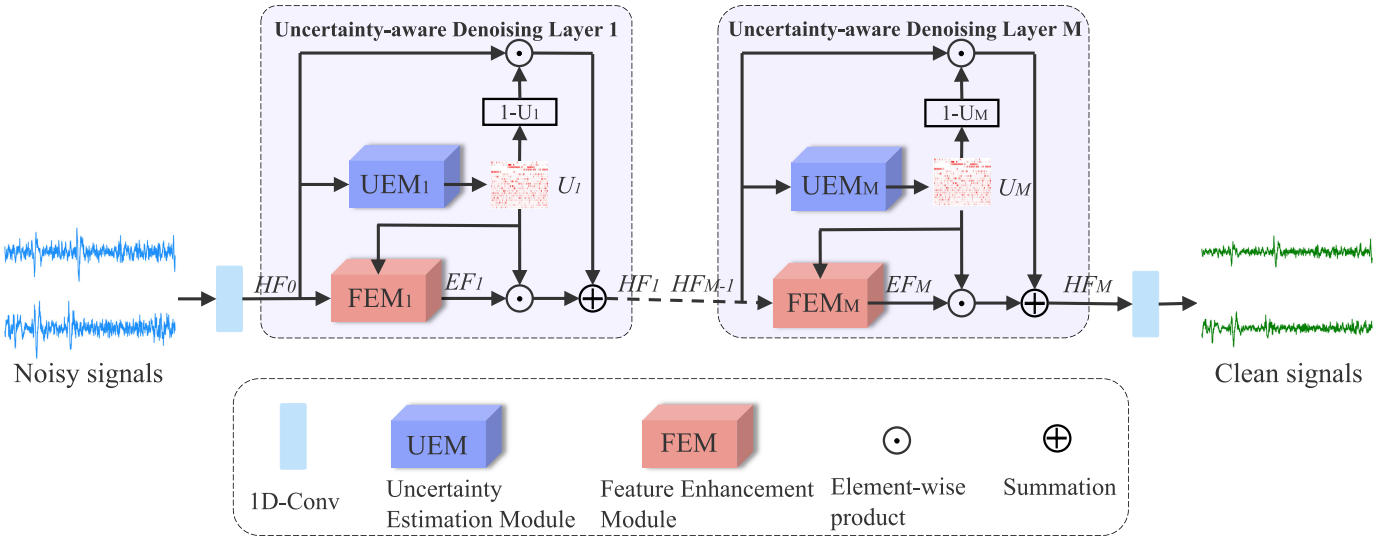


Fig. 2. Overview of the proposed UDNet. The UDNet is composed of M Uncertainty-aware Denoising Layers (UDLs), where each layer, UDL$_m$ $\{m = 1, \ldots, M\}$, incorporates the use of UEM$_m$ to estimate the corresponding uncertainty map $U_m$, along with an intermediate denoising result $\hat{J}_m$. Additionally, FEM$_m$ is employed to modulate $HF_{m-1}$, generating an enhanced feature $EF_m$ and amplifying uncertain features from $HF_{m-1}$. Furthermore, a gate unit is utilized to aggregate $HF_{m-1}$ and $EF_m$, producing a more reliable and improved representation, $HF_m$.

efficient use of space and time. Formally, it is defined as follows,

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{F}}V) \tag{5}$$

where $Q, K, V$ and $F$ are queries, keys, values, and their dimension, respectively.

## IV. Methods

Our goal is to restore clean EEG segments with better confidence (less uncertainty) from their noisy observation containing EOG or EMG artifacts. To this end, we introduce an Uncertainty-aware Denoising Network (UDNet), which focuses on improving the uncertain feature representation and leveraging the confident parts. As shown in Figure 2, UDNet consists of M Uncertainty-aware Denoising Layers (UDLs) that are interconnected to enhance the hidden features (HFs). In the $m$th UDL ($m = 1, \ldots, M$), an Uncertainty Estimation Module (UEM) is utilized to estimate the uncertainty map $U_m$ and produce an intermediate denoising result $\hat{J}_m$. Additionally, a Feature Enhancement Module (FEM) is employed to modulate the previously hidden feature $HF_{m-1}$ by leveraging the confident feature and generating the Enhanced Feature $EF_m$. To combine the enhanced feature $EF_m$ and the previously hidden feature $HF_{m-1}$, we use the uncertainty map $U_m$ as a gate. This linear combination allows UDL to update the representation in $HF_{m-1}$ with $EF_m$ and output a more confident and improved representation $HF_m$. The process can be summarized as follows:

$$HF_m = (1 - U_m) \odot HF_{m-1} + U_m \odot EF_m \tag{6}$$

where $\odot$ represents the element-wise product.

To provide more specific details, the first step in our approach is to convert the raw input $x \in \mathbb{R}^{L \times d}$ into a high dimensional representation $HF_0 \in \mathbb{R}^{L \times F}$ via linear projection. After that, $HF_0$ is updated gradually by $M$ UDLs to obtain $HF_0, HF_1, \ldots, HF_M$. Finally, $HF_M$ is converted to the original size by linear projection.

### A. Aleatoric and Epistemic Uncertainty Estimation

Bayesian deep learning provides a comprehensive framework for modeling two distinct types of uncertainty: 1) aleatoric uncertainty, which arises from the noise inherent in the observations, and 2) epistemic uncertainty, which captures uncertainty within the model itself. These two forms of uncertainty are also present in denoising models. However, conventional methods typically yield deterministic outcomes without providing any information about their associated confidence. In this paper, we introduce an Uncertainty Estimation Module (UEM) to model sampling-point-wise aleatoric uncertainty $\sigma_A^2$ and epistemic uncertainty $\sigma_E^2$ together. As shown in Figure 3, UEM contains two branches to model aleatoric and epistemic uncertainty, separately.

*1) Aleatoric Uncertainty:* Aleatoric uncertainty is explained as the inherent noise and random influences that cannot be explained explicitly. In our approach, we make the assumption that the denoising output at each sampling point, denoted as $p(J|\hat{J}, \omega)$, follows a Gaussian distribution. The mean and variance of this distribution correspond to the ground-truth signal $J$, and aleatoric uncertainty $\sigma^2$, where $\omega$ represents the network parameters. In the context of Bayesian neural networks, the functions are defined through the weights of the neural network, which serve as our sufficient statistics denoted as $\omega = (W_m)_{m=1}^M$. We perform Maximum A Posteriori (MAP) inference to obtain the optimal values for $\omega$ when given the observed data and any prior knowledge or assumptions, as follows:

$$\begin{aligned}\hat{\omega} &= \arg\max_{\omega} \log(p(J|\hat{J}, \omega)) \\ &= \arg\max_{\omega} \{-\frac{1}{2\sigma^2}||J - \hat{J}||_2^2 - \frac{1}{2}\log\sigma^2\}\end{aligned} \tag{7}$$

We treat $\sigma_A^2 = \sigma^2$. Two branches in UEM are used to predict $\sigma_A^2$ and $\hat{J}$, separately. For aleatoric uncertainty, it is conditioned on the denoising results of the previous layer and finally constrained by the following

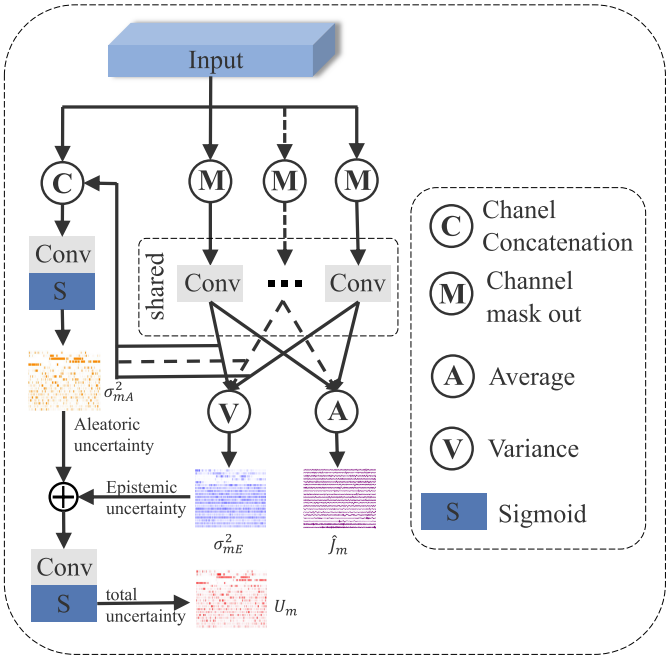


Fig. 3. The uncertainty estimation module (UEM) consists of aleatoric uncertainty and epistemic uncertainty estimation. Aleatoric uncertainty is obtained through 1-layer convolution and Sigmoid. Epistemic uncertainty is obtained through the Monte Carlo estimation of multiple shared convolutions.

minimization objective:

$$L_r^m = \frac{1}{D_m}\sum_{i=1}^{D_m}[\frac{1}{2(\sigma_{mA}^i)^2}(J_m^i - \hat{J}_m^i)^2 + \frac{1}{2}\log(\sigma_{mA}^i)^2] \quad (8)$$

where the superscript $i$ denotes the sampling point index, and $D_m$ is the number of the output points. In practice, we train the UEM to predict the log variance, $s_{mA} := \log(\sigma_{mA}^i)^2$:

$$L_r^m = \frac{1}{D_m}\sum_{i=1}^{D_m}[\frac{1}{2}\exp(-s_{mA})(J_m^i - \hat{J}_m^i)^2 + \frac{1}{2}s_{mA}] \quad (9)$$

*2) Epistemic Uncertainty:* Epistemic uncertainty encompasses the uncertainty in the model parameters, representing our lack of knowledge regarding which specific aspects of the model generated the observed data. Existing denoising models fail to capture epistemic uncertainty since they follow the deterministic network's parameters and optimize the network directly.

Bayesian neural networks have always been used to capture epistemic uncertainty, replacing the deterministic network's parameters with a prior distribution. Then performing Bayesian inference to compute the posterior distribution over these weights. While posterior distribution is not tractable for a Bayesian NN, Bayesian convolutional neural networks [40] define approximating variational distribution $q_\theta(W_m)$ for every layer $m$ to relate the approximate inference to dropout training as:

$$\begin{aligned} W_m &= M_m \cdot \mathrm{diag}([z_{m,n}]_{n=1}^{F_m}) \\ z_{m,n} &\sim \mathrm{Bernoulli}(p_m) \text{ for } m = 1,\ldots,M, n = 1,\ldots,F_m \end{aligned} \quad (10)$$

here $z_{m,n}$ are random variables following a Bernoulli distribution with probabilities $p_m$, and $M_m$ represents the variational parameters need to be optimized. The operator diag$(\cdot)$ transforms vectors into diagonal matrices.

Following [40], we reframe the 1-D convolution operation as a linear operation that integrates over the kernels. Specifically, let $K_k \in \mathbb{R}^{l\times F_{m-1}}$, where $k = 1,\ldots,F_m$, be the CNN's kernels with length $l$, and $F_{m-1}$ channels. The input to the layer denoted as $x \in \mathbb{R}^{L_{m-1}\times F_{m-1}}$. By convolving the input with a given stride $s$, we can interpret it as extracting segments from the input, each with dimensions $l \times F_{m-1}$. These segments are then vectorized and collected as rows in a matrix, resulting in a new representation denoted as $\bar{x} \in \mathbb{R}^{L_m \times lF_{m-1}}$, where $L_m$ represents the number of segments. The vectorized kernels are arranged as columns in the weight matrix $W_m \in \mathbb{R}^{lF_{m-1}\times F_m}$. Thus, the convolution operation can be expressed as the matrix product $\bar{x}W_m \in \mathbb{R}^{L_m\times F_m}$. To capture epistemic uncertainty, we introduce a prior distribution on convolution kernels and leverage Bernoulli variational distributions to approximately integrate each kernel-segment pair. Then we sample Bernoulli random variables $z_{m,n}$ and then multiply segment $L_m$ by the weight matrix $M_m \cdot diag([z_{m,n}]_{n=1}^{F_m})$, which is equivalent to an approximating distribution modeling each kernel-segment pair with a distinct random variable, tying the means of the random variables over the segments. Such a modeling approach randomly sets certain kernels to zero for different segments. Implementing Bayesian CNN in the inference stage is therefore equivalent to applying dropout after every convolution layer through multiple forward propagations.

Further, in order to quantify epistemic uncertainty and perform Bayesian CNN in the training stage, we find an alternative method. To be specific, we incorporate distributions over each 1D-Conv-Sigmoid layer of each UEM and employ a mask operation to approximate the inference process. Herein, we randomly mask parts of the input feature channels by setting their values to 0. These masked inputs are then passed through a shared Conv layer to reconstruct original EEG signals. This process is repeated T times, resulting in T different denoising results $\{\hat{J}_{m,t}\}_{t=1}^T$. Subsequently, we calculate the mean and variance (representing epistemic uncertainty) of $\{\hat{J}_{m,t}\}_{t=1}^T$. Specifically, we compute the predicted mean $\hat{J}_m = \frac{1}{T}\sum_{t=1}^T \hat{J}_{m,t}$ and the epistemic uncertainty $\sigma_{mE}^2 = \frac{1}{T}\sum_{t=1}^T \hat{J}_{m,t}^2 - (\frac{1}{T}\sum_{t=1}^T \hat{J}_{m,t})^2$, which quantifies the level of uncertainty in the model's prediction.

To summarize, the predicted uncertainty of each sampling point in the $m$th UEM can be approximated using the following:

$$\begin{aligned} U_m &\approx \sigma_{mE}^2 + \sigma_{mA}^2 \\ &= \frac{1}{T}\sum_{t=1}^T \hat{J}_{m,t}^2 - (\frac{1}{T}\sum_{t=1}^T \hat{J}_{m,t})^2 + \sigma_{mA}^2 \end{aligned} \quad (11)$$

With the assistance of the uncertainty map $U_m$ obtained from UEM$_m$, we can discern the level of uncertainty associated with each sampling point.

## B. Feature Enhancement Module (FEM)

The primary objective of FEM is to enhance the uncertain feature. It is observed that hard-to-denoise regions are

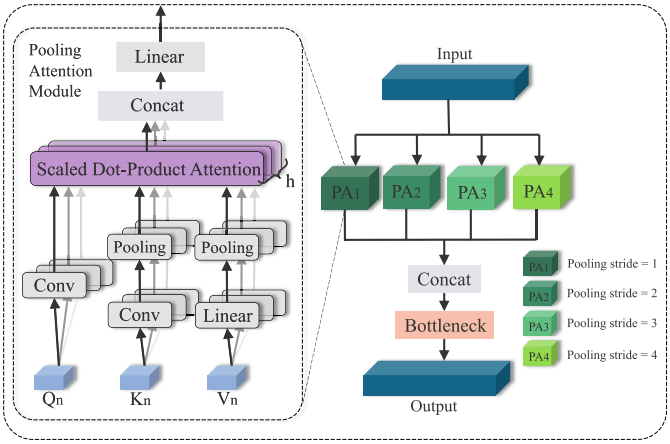


Fig. 4. The feature enhancement module (FEM) consists of multiple multi-scale pooling attention units and a bottleneck layer.

highly related to the estimated uncertainty map, which usually contains complex contexts. In the artifact removal task, it is not surprising that the original signal amplitude from different time steps is the same. The non-local similarity characteristic suggests that two distant sampling points in the clear signal, despite having the same amplitude, may experience different types of degradation. This results in complex contextual dependencies in their corresponding noisy observations. More importantly, a denoising network is required to map these two different observations to the same amplitude, which presents a challenging many-to-one mapping task. Therefore, capturing the trend around the current time step to further help identify informative contexts is the key point.

However, capturing local trends under ambiguous contexts as existing works do [33], [34], [35] is not a simple thing. The captured inaccurate contexts by 1-D convolution will continuously accumulate errors layer by layer. Alternatively, global receptive fields can help to learn effective contextual information to a certain extent since similar trends at the far end may be relatively clean. In representation learning scenarios, the attention mechanism is commonly employed to automatically extract the most pertinent information, especially from the global perspective [44]. Nevertheless, when it comes to capturing context information from long-time series data like EEG signals, employing the attention mechanism can pose challenges due to limitations in computing resources and memory.

Based on this premise, we have devised a pooling-attention architecture that endeavors to capture both local and global contextual information concurrently while mitigating the quadratic attention complexity. This design improvement enhances the efficiency of attention-like modules for EEG artifact removal applications. Additionally, we integrate multi-scale convolutions instead of linear mappings to enhance the perception of trends in the data.

*1) Trend-Aware Multi-Head Attention:* We propose the Trend-aware Multi-Head self-Attention (TMHA) mechanism to capture contextual information in signals. TMHA is built upon the self-attention mechanism, which enables the derivation of queries, keys, and values from the same sequence of symbol representations. In TMHA, we first employ Multi-Head Self Attention (MHA), which enables simultaneous attention across multiple representation subspaces.

To implement MHA, previous works apply linear projections and transform them into separate representation subspaces. The attention function (Eq. 5) is then independently performed in parallel for each subspace. Afterward, the resulting outputs from each subspace are concatenated and projected to generate the final output. By employing MHA, we can effectively capture interdependencies between different parts of the signals, facilitating the modeling of contextual information in a trend-aware manner. Formally,

$$\text{MHSelfAttention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \oplus(\text{head}_1,\ldots,\text{head}_h)W^O \tag{12}$$

$$\text{head}_j = \text{Attention}(\mathbf{Q}W_j^Q, \mathbf{K}W_j^K, \mathbf{V}W_j^V) \tag{13}$$

where $h$ denotes the number of attention heads. $W_j^Q$, $W_j^K$, $W_j^V$, and $W^O$ are projection matrices applied to $\mathbf{Q}$, $\mathbf{K}$, $\mathbf{V}$, and the final output, respectively.

The traditional multi-head self-attention mechanism, originally designed for discrete tokens like words, may not adequately capture the local trend information inherent in continuous data such as EEG signals. Applying this mechanism directly to EEG signal transformation can lead to a mismatch between the attention mechanism and the data characteristics [31]. To address this limitation and incorporate local trend information into numerical data prediction, we propose a novel approach called TrSelfAttention, which stands for Transformer-based Self-Attention. TrSelfAttention is inspired by the Convolutional Self-Attention model [31] and introduces 1D convolutions to replace the projection operations on queries and keys in Eq. 13. This modification enables the model to consider local contextual information and be more sensitive to the changing trends present in the noisy EEG signals. Mathematically, the definition of TrSelfAttention is as follows:

$$\text{TrSelfAttention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \oplus(\text{Trhead}_1,\ldots,\text{Trhead}_h)W^O \tag{14}$$

$$\text{Trhead}_j = \text{Attention}(\Phi_j^Q \star \mathbf{Q}, \Phi_j^K \star \mathbf{K}, \mathbf{V}W_j^V) \tag{15}$$

where $\star$ indicates the convolution operation and $\Phi_j^Q$, $\Phi_j^K$ are the parameters of convolution kernels.

*2) Pooling Attention:* To address the quadratic complexity issue in self-attention blocks, we introduce a pooling operation before attending to the input. This pooling operator, denoted as $P(\cdot;\Theta)$, is used to downsample the intermediate tensors $\hat{K}$ and $\hat{V}$. The parameter $\Theta := (\mathbf{k},\mathbf{s},\mathbf{p})$ specifies the pooling kernel size ($k_t$), stride ($s_t$), and padding ($p_t$). By default, we employ *non-overlapping* kernels with *shape-preserving* padding in our pooling attention operators. This results in an output tensor with a reduced signal length $\tilde{L}$, which is achieved by a factor of $\mathbf{s}$ compared to the input tensor's length $L$. The pooled tensors are denoted as $\hat{K} = P(\cdot;\Theta_K)$ and $\hat{V} = P(\cdot;\Theta_V)$. The attention computation is then performed on these shortened vectors.

$$\text{PoTrAttention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \oplus(\text{PoTrhead}_1,\ldots,\text{PoTrhead}_h)W^O \tag{16}$$

$$\text{PoTrhead}_j = \text{Attention}(\Phi_j^Q \star \mathbf{Q}, \mathcal{P}(\Phi_j^K \star \mathbf{K};\Theta_K), \mathcal{P}(\mathbf{V}W_j^V;\Theta_V)) \tag{17}$$

Naturally, the pooling operation introduces the constraint $s_K \equiv s_V$.

**Computational Analysis** By pooling the key and value tensors, the computation and memory requirements of attention computation, which scales quadratically with the signal length, are dramatically reduced. Denoting the reduction factors for the signal lengths as $f_Q$, $f_K$, and $f_V$, we have $f_Q = 1$, $f_K = s_k$, $f_V = s_v$. If we consider the input tensor to the pooling operator $P(\cdot;\Theta)$ to have dimensions $D \times T$, the runtime complexity of TMHA is $O(DT/h(D + T/f_Q f_K))$ per head, and the memory complexity is $O(Th(D/h + T/f_Q f_K))$.

*3) Multi-Scale Pooling Attention:* The effectiveness of examining EEG signals in multiple scales has been demonstrated in several studies [26], [49]. In the context of EEG artifact removal, we are motivated by two factors to design a Multi-Scale Pooling Attention mechanism. (i) By working at multi-scale lower resolutions, we hope to reduce the computing requirements as well as allow maintaining satisfactory performance. (ii) Multi-scales provide a more comprehensive sense of context within the EEG signals. This contextual information at multiple lower resolutions can better guide the processing and decision-making at higher resolutions.

Herein we utilize 4 Pooling Attention units, each with different kernel sizes and pooling strides ranging from 1 to 4. These units operate on the same input feature map, enabling the extraction of information from multiple receptive field scales. By applying pooling operations at various scales, we generate feature maps for each scale. These feature maps are then combined to create a multi-scale feature map. This integration of information from different receptive field sizes enhances the representation of the input data, allowing for a more comprehensive understanding of its underlying patterns and structures.

$$X^r_{ss} = \text{PoTrAttention}_r(\mathbf{Q}, \mathbf{K}, \mathbf{V}), r \in [1, 2, 3, 4] \tag{18}$$

$$X_{ms} = Concat(X^1_{ss}, X^2_{ss}, X^3_{ss}, X^4_{ss}) \tag{19}$$

where PoTrAttention is the Trend-aware Pooling Attention with kernel sizes and pooling strides from scales 1 to 4, and $X^r_a$ is the multi-scale feature map.

To minimize the number of parameters, we incorporate a bottleneck layer, which is responsible for reducing the channels in the concatenated feature map as follows:

$$EF = Bottleneck(X_{ms}) \tag{20}$$

where $EF$ is the final multi-scale feature map and has $C_{out} = C_{in}/rate$ channels. $C_{in}$ is the channel number of $X_{ms}$ and $rate$ is the down-sampling rate of bottleneck layer.

### C. Model Training

Considering the presence of M UDLs in UDNet, we have M UEMs responsible for estimating the denoising results and uncertainty maps. To formulate the overall objective, we define the following:

$$\mathcal{L} = \theta_f \sum_i^N |\hat{y}_i - y_i|^2 + \sum_{m=1}^M \theta_m \mathcal{L}^m_r \tag{21}$$

where $N$ is the total number of samples, $\hat{y}$ is the final denoised EEG signal, $\mathcal{L}^m_r$ is the reconstruction loss described in Eq. 9, $\theta_f$ and $\theta_m$ are weight factors.

## V. Experiments

To validate the effectiveness of the proposed method, we conduct performance comparisons on semi-simulated EEG recordings using publicly available datasets, ISRUC and TUSZ. Each dataset contains EEG that has undergone visual inspection and noise reduction processing. We synthesize contaminated signals based on corresponding artifact sources.

### A. Dataset and Experiment Settings

*1) ISRUC Dataset:* ISRUC-S3 dataset [46] contains 10 healthy subjects (9 male and 1 female). Each recording in the dataset includes 6 EEG channels, 2 EOG channels, and 3 EMG channels. Furthermore, domain experts have classified these polysomnography (PSG) recordings into five sleep stages, adhering to the standards set by the American Academy of Sleep Medicine (AASM).

*2) TUSZ Dataset:* TUSZ dataset [47] stands as one of the largest annotated datasets available for EEG seizure classification. It comprises a total of 5,612 EEGs, encompassing 3,050 annotated seizures extracted from clinical recordings, and encompasses four distinct seizure types [48]. The dataset includes 19 EEG channels following the standard 10-20 system.

### B. Implementation Details

We implemented the UDNet model based on the PyTorch framework and trained by the Adam optimizer with the learning rate of $10^{-3}$. The model dimension $F$ is 64, and the number of layers $L$ is 4. We empirically set $\theta_f$ and $\theta_m$ to 1.

We perform subject-independent experiments and split each dataset into training and testing sets in the ratio of 6:1 in TUSZ and 9:1 in ISRUC. Each experiment was repeated 5 times, and the reported results represent the mean values.

We use an end-to-end way to train and rely on synthetic data as the contaminated EEG and ground truth to optimize the total objective $\mathcal{L}$. To be specific, the contaminated EEG $x$ can be generated by linearly combining the clean EEG segments $y$ with EOG or EMG artifact segments, as described by the following equation:

$$x = y + \lambda \cdot n \tag{22}$$

where the term $n$ represents either ocular or myogenic artifacts. The hyperparameter $\lambda$ is utilized to regulate the signal-to-noise ratio (SNR) in the contaminated EEG signal, as indicated:

$$\text{SNR} = 10 \times \log \frac{\text{RMS}(y)}{\text{RMS}(\lambda \cdot n)} \tag{23}$$

$$\text{RMS}(z) = \sqrt{\frac{1}{N_z} \sum_{i=1}^{N_z} z_i^2} \tag{24}$$

where $\text{RMS}(\cdot)$ denotes the root mean square, $N_z$ denotes the number of EEG points in the segment $z$, and $z_i$ denotes the $i$th sampling point. The SNR in our experiment is $-10$ dB.

### C. Performance Metrics

*1) Change in Signal to Noise Ratio:* We define the metric $\Delta$SNR as the change in the signal-to-noise ratio before and after artifact removal. The calculation of $\Delta$SNR is defined as

$$\Delta\text{SNR} = \text{SNR}_{\text{after}} - \text{SNR}_{\text{before}} \tag{25}$$

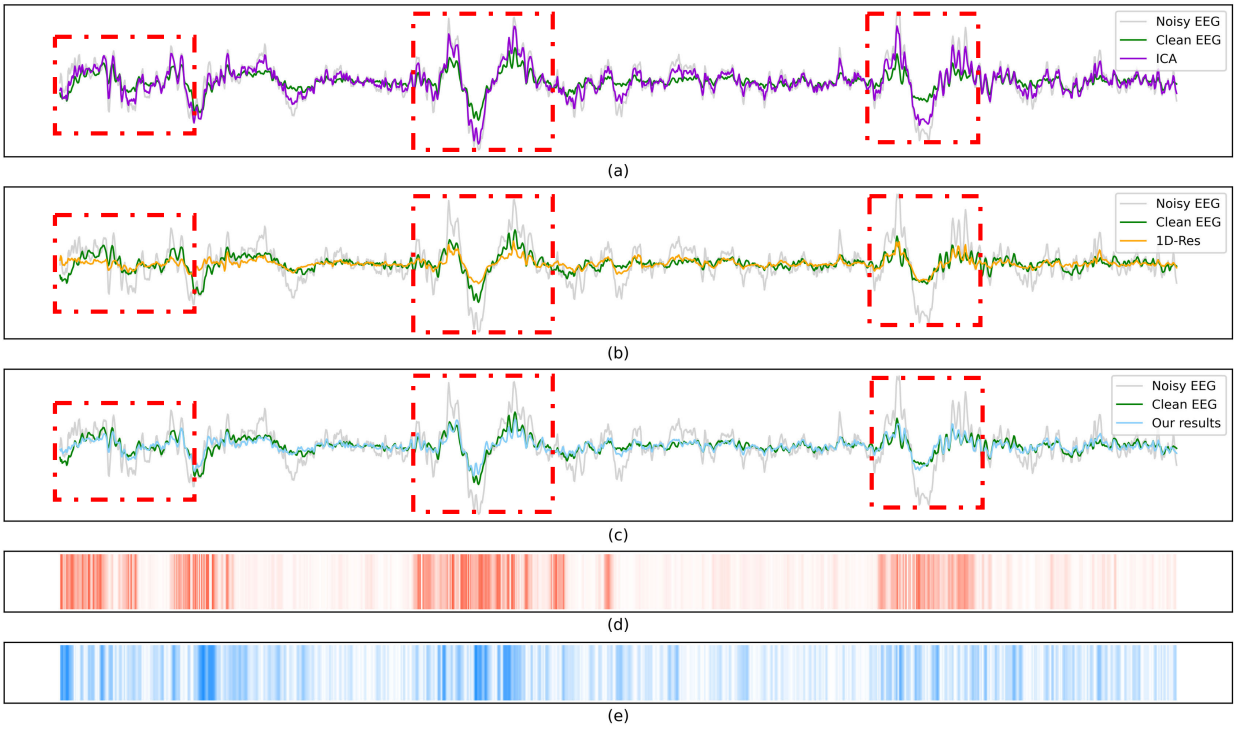


Fig. 5. Visualization of the denoised results, uncertainty maps, and reconstruction errors. The red dashed line indicates the part with severe artifacts. (a) ICA results. (b) 1D-ResCNN results. (c) UDNet results. (d) Uncertainty map. (e) Error map.

where $\text{SNR}_{\text{before}}$ and $\text{SNR}_{\text{after}}$ are the signal-to-noise ratio before and after artifact removal, respectively.

*2) Normalized Mean Squared Error:* The normalized mean squared error (NMSE), in decibels, is defined as:

$$\text{NMSE} = 10\log\frac{\sum_i |\hat{y}_i - y_i|^2}{\sum_i |y_i|^2} \tag{26}$$

where $y_i$ is the i-th sample of the signal y.

*3) Change in Correlation:* We defined the change in correlation $\Delta$R before and after the artifact removal as:

$$\Delta R = R_{\text{after}} - R_{\text{before}} \tag{27}$$

where $R_{\text{before}}$ and $R_{\text{after}}$ are the Pearson correlation coefficients between the ground truth and the signal before and after artifact removal, respectively.

*4) Improvement in Spectral Coherence:* We defined the improvement in coherence $I_{\text{coh}}$ before and after the artifact removal as:

$$I_{\text{coh}} = (C_{\text{after}} - C_{\text{before}})/(1 - C_{\text{before}}) \tag{28}$$

where $C_{\text{before}}$ and $C_{\text{after}}$ denote the average magnitude squared coherence, calculated between the ground truth and the signal before and after artifact removal, respectively.

## D. Comparison With the State-of-the-Art Methods

To validate the effectiveness of the UDNet, we conducted experiments with the subject-independent procedure. We compare the proposed UDNet with four traditional methods (i.e., EMD-ICA, EMD-CCA [12], WT [15], and WQN [16]), two signal processing-based methods (i.e., WOSG [17] and DBPEWT [18]), and five DL-based methods (i.e., MMNN [38], Novel CNN [35], 1D-ResCNN [34], DeepSeparator [33], and GRUMARSC [37]). We included artifacts of different natures (ocular, muscular) using datasets of semi-simulated. In Table I, it was observed that UDNet achieved the highest increase $\Delta$SNR and the lowest NMSE across all datasets. Furthermore, UDNet demonstrated superior performance in terms of correlation and coherence improvement, because UDNet can effectively preserve frequency information and avoid spectral distortion. In particular, EMD-ICA and EMD-CCA showed inferior performance on seizure-related datasets since TUSZ contains more subjects and channel numbers than ISRUC. WOSG and DBPEWT exhibit stable results across different datasets. MMNN is ineffective in NMSE and $\Delta$SNR indicators, while it performs well in $\Delta$R and $I_{\text{coh}}$, demonstrating its ability to recover information in the spectral domain. Novel CNN and 1D-ResCNN had relatively stable performance against different noise types and datasets. DeepSeparator severely damages spectral domain information in EOG artifact scenarios. GRUMARSC had a significantly reduced effect on more complex epilepsy data. Overall, UDNet stood out among the tested methods, achieving the best improvement in both temporal and spectral domains.

*1) Ablation Experiment:* To assess the individual contributions of each module in our model, we designed several variant models. These variants involve modifications to specific modules while keeping the rest of the architecture unchanged. We start with a 4-layer 1-D temporal convolution architecture to construct the basic model, which serves as the foundation upon which we gradually add and stack the remaining modules to create a complete branch. First, we use 4-layer 1-D temporal convolution as the basic model to gradually stack the remaining modules to form a whole branch. Then, we integrate data and model uncertainty estimation into the basic model, separately. Finally, we integrate the multi-scaled

TABLE I
RESULT ON VALIDATION ON DIFFERENT DATASETS FOR THE PROPOSED UDNET METHOD. ΔSNR AND NMSE VALUES ARE IN dB. THE BEST PERFORMANCE FOR EACH METRIC IS HIGHLIGHTED IN BLACK

| Dataset | Methods | ΔSNR ↑ | NMSE ↓ | ΔR ↑ | $I_{coh}$ ↑ |
|---|---|---|---|---|---|
| TUSZ-EOG | EMD-ICA | 10.91 | -0.81 | 0.43 | 0.15 |
| | EMD-CCA | 12.81 | -2.69 | 0.46 | 0.17 |
| | WT | 10.16 | -0.13 | 0.35 | 0.21 |
| | WQN | 10.28 | -0.22 | 0.24 | 0.25 |
| | WOSG | 12.09 | -0.02 | 0.21 | 0.11 |
| | DBPEWT | 10.45 | -0.56 | 0.25 | 0.14 |
| | MMNN | 23.20 | -2.03 | 0.41 | 0.08 |
| | Novel CNN | 23.36 | -2.71 | 0.38 | 0.01 |
| | 1D-ResCNN | 25.47 | -4.28 | 0.60 | 0.38 |
| | DeepSeparator | 20.08 | 0.10 | 0.03 | -0.02 |
| | GRUMARSC | 22.07 | -2.04 | 0.18 | -0.03 |
| | **UDNet** | **26.99** | **-6.12** | **0.68** | **0.52** |
| TUSZ-EMG | EMD-ICA | 18.70 | -7.99 | 0.63 | 0.01 |
| | EMD-CCA | 17.99 | -7.25 | 0.60 | 0.04 |
| | WT | 11.37 | -1.27 | 0.39 | 0.09 |
| | WQN | 9.27 | 0.78 | 0.19 | 0.05 |
| | WOSG | 12.08 | -0.02 | 0.22 | 0.08 |
| | DBPEWT | 10.72 | -0.44 | 0.25 | 0.21 |
| | MMNN | 28.98 | -7.79 | 0.77 | 0.15 |
| | Novel CNN | 27.02 | -6.28 | 0.67 | -0.04 |
| | 1D-ResCNN | 30.03 | -9.00 | 0.80 | 0.27 |
| | DeepSeparator | 26.08 | -5.99 | 0.67 | 0.03 |
| | GRUMARSC | 27.33 | -7.37 | 0.70 | -0.03 |
| | **UDNet** | **31.21** | **-10.18** | **0.81** | **0.42** |
| ISRUC-EOG | EMD-ICA | 10.21 | -0.13 | 0.24 | 0.28 |
| | EMD-CCA | 10.07 | -0.02 | 0.26 | 0.28 |
| | WT | 10.61 | -0.57 | 0.20 | 0.20 |
| | WQN | 11.19 | -1.17 | 0.09 | 0.23 |
| | WOSG | 10.30 | -0.36 | 0.19 | 0.13 |
| | DBPEWT | 14.76 | -1.14 | 0.22 | 0.27 |
| | MMNN | 25.00 | -4.86 | 0.41 | 0.35 |
| | Novel CNN | 22.43 | -2.48 | 0.22 | -0.17 |
| | 1D-ResCNN | 25.14 | -5.08 | 0.41 | 0.36 |
| | DeepSeparator | 20.97 | -0.96 | -0.01 | -0.09 |
| | GRUMARSC | 23.84 | -3.85 | 0.32 | -0.05 |
| | **UDNet** | **25.49** | **-5.49** | **0.43** | **0.43** |
| ISRUC-EMG | EMD-ICA | 14.44 | -4.76 | 0.53 | -0.03 |
| | EMD-CCA | 16.34 | -6.14 | 0.57 | 0.01 |
| | WT | 10.68 | -0.59 | 0.35 | 0.07 |
| | WQN | 8.62 | 1.42 | 0.14 | 0.02 |
| | WOSG | 10.23 | -0.23 | 0.29 | 0.05 |
| | DBPEWT | 15.01 | -1.49 | 0.10 | -0.02 |
| | MMNN | 30.89 | -10.69 | 0.80 | 0.29 |
| | Novel CNN | 25.74 | -5.83 | 0.71 | -0.07 |
| | 1D-ResCNN | 30.46 | -10.32 | 0.81 | 0.30 |
| | DeepSeparator | 27.07 | -7.15 | 0.77 | 0.00 |
| | GRUMARSC | 30.05 | -10.13 | 0.52 | 0.04 |
| | **UDNet** | **31.02** | **-10.94** | **0.84** | **0.41** |

pooling-attention feature enhancement module to form the proposed model. The specific process is described as follows:

- variant a (Temporal Convolution (Base Model)): We utilize a 4-layers temporal convolution as the base model.
- variant b (+ Data Uncertainty): We add data uncertainty estimation $\sigma^2_{mA}$ **based on variant a** to form a data-uncertainty-aware temporal convolution network.
- variant c (+ Model Uncertainty): We add model uncertainty estimation $\sigma^2_{mE}$ **based on variant a** to form a model-uncertainty-aware temporal convolution network.

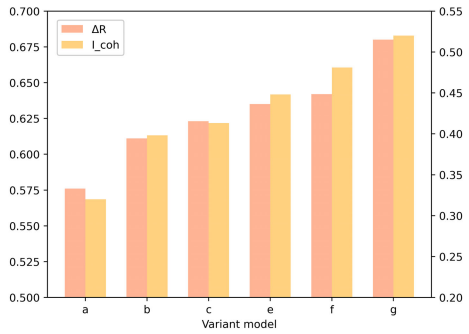


Fig. 6. A comparison of the designed variant models to assess the effectiveness of different modules in UDNet.

- variant d (+ Data Uncertainty + Model Uncertainty): We add two kinds of uncertainty estimation $U_m = \sigma^2_{mA} + \sigma^2_{mE}$ **based on variant a** to form an uncertainty-aware temporal convolution network.
- variant e (+ Single-scale Pooling-Attention): We replace single-scale temporal convolution with single-scale pooling-attention feature enhancement **based on variant d** equipped with an uncertainty estimation module.
- variant f (+ Multi-scale Pooling-Attention): We replace single-scale temporal convolution with multi-scale pooling-attention feature enhancement equipped **based on variant d** with the whole uncertainty estimation module.

Figure 6 demonstrates the effectiveness of the key modules in our model. We begin by evaluating the impact of the two types of uncertainty individually. The results show that both types of uncertainty lead to performance improvements. Additionally, using the fused uncertainty $U_m = \sigma^2_{mA} + \sigma^2_{mE}$, which combines both types, further enhances the performance. Moreover, trend-aware pooling attention provides a global receptive field for capturing the global trend context. Meanwhile, multi-scales are better than single scales since more potential patterns can be characterized distinctively. In summary, the ablation experiment validates the effectiveness of each module in our model.

*2) Effectiveness of UEM:* We further visualize the denoised result, uncertainty map, and reconstruction errors in Figure 5. From top to down, firstly, our denoised results are superior to traditional and deep methods. More importantly, we can observe that the uncertainty maps are intricately linked to the reconstruction errors, whereby the magnitude of the error directly correlates with the corresponding uncertainty value. Utilizing these maps, epochs that receive high uncertainty scores from our model can be appropriately deferred to clinical experts for further scrutiny and examination. Therefore, it is of great significance to provide uncertainty-aware reconstruction, which actively prevents misleading data use and decision-making.

*3) Influence of Denoising on Nonlinear Characteristics of Signals:* Figure 7 and 8 show the power spectral density (PSD) of clean EEG, contaminated EEG, and denoising EEG treated by MMNN, Novel CNN, 1D-ResCNN, DeepSeparator, GRUMARSC, and our proposed UDNet, which includes EOG noise and EMG noise. As depicted in the two figures, the PSD of the EEG signal noticeably decreases in the specific frequency range of the noise after applying the noise

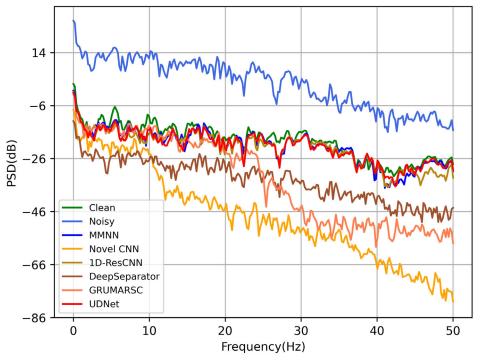


Fig. 7. PSD results of removing EOG noise from EEG signal.

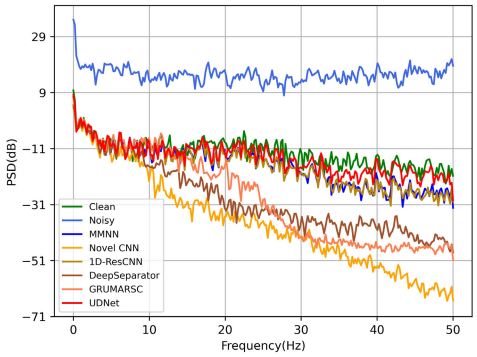


Fig. 8. PSD results of removing EMG noise from EEG signal.

reduction techniques. Importantly, our UDNet generates the closest PSD to the original clean EEG signal.

*4) Downstream Task Performance of Different Artifact Removal Methods:* To compare the quality of generating noise reduction results, we compare UDNet with 6 representative artifact removal methods on two downstream tasks (i.e., seizure classification and sleep staging) with corr-DCRNN [48], and MSTGCN [50] as the task-related classification models. We pre-train them on clean datasets and test on the contaminated datasets. Figure 9 shows the total F1 score and each class F1 score results of the TUSZ and ISRUC datasets with different artifact sources, respectively. We also provide clean EEG signal classification results for comparison.

UDNet improves the classification performance of almost all types in TUSZ and ISRUC datasets. Specifically, in the TUSZ dataset, the pre-trained corr-DCRNN performs extremely worse except for CF whose training samples are much more than the summation of all other types. Importantly, our method significantly improves the classification accuracy of minority classes, i.e., AB, CT, and GN. Almost all the remaining methods had similar results to the noisy EEG signal. For the ISRUC dataset, the results were similar to TUSZ. MSTGCN cannot accurately judge noisy EEG signals and drops severely in all five stages. Our UDNet improves the classification performance, especially for REM and Wake stages. Overall, the denoised signals generated by traditional methods only have limited improvement in downstream tasks. DL-based methods like the 1D-Res, DeepSeparator, and GRUMARSC are relatively better-performing and stable models. Our UDNet can produce robust denoising results on both datasets.

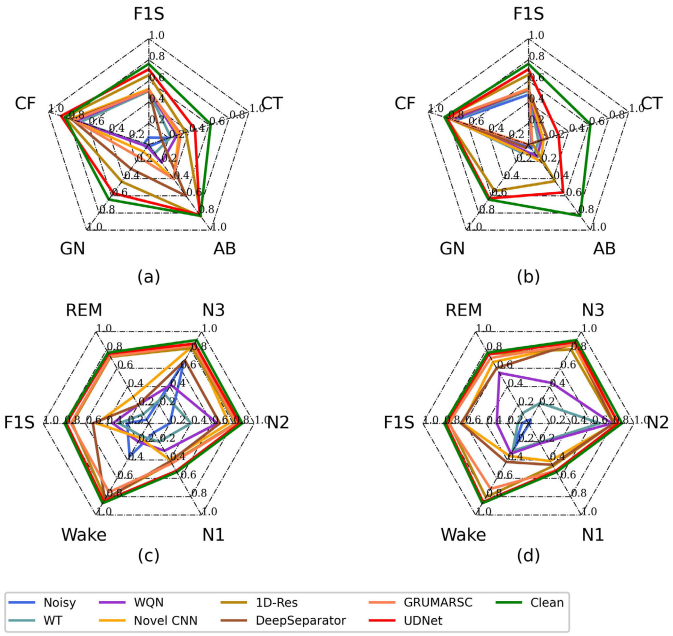


Fig. 9. Accuracy performance patterns of different seizure types and sleep stages were obtained by corr-DCRNN and MSTGCN. The total and each class's F1 scores were evaluated. Seizure types conclude: combined focal (CF), generalized non-specific (GN), absence (AB), and combined tonic (CT). Sleep stages conclude: Wake, N1, N2, N3, REM. (a)TUSZ-EOG (b)TUSZ-EMG (c)ISRUC-EOG (d)ISRUC-EMG.

## VI. Discussion

In this paper, we propose an innovative denoising network called UDNet, which aims to produce reliable and accurate denoised results. Our method has a more stable and accurate denoising performance compared with traditional signal processing methods. Additionally, by utilizing dimensionality-invariant convolution operations, our method can process noisy signals of any length. Furthermore, our approach seldom requires retraining the network when processing data closely resembles the training set once our model is fully trained. We incorporate uncertainty quantification to improve model interpretability compared to deep learning methods. While providing denoised output, the corresponding credibility can be provided at the same time, which can improve the clinical application prospects of deep learning. Extensive experimentation on synthetic datasets demonstrates the remarkable quantitative and qualitative enhancements achieved by UDNet, surpassing the current state-of-the-art techniques. Despite this, the solution in this article has a large demand for computing resources. When computing resources are limited, more Monte Carlo samplings cannot be performed, thus limiting the ability to quantify uncertainty.

## References

[1] A. Bhattacharyya et al., "A multi-channel approach for cortical stimulation artefact suppression in depth EEG signals using time-frequency and spatial filtering," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 7, pp. 1915–1926, Jul. 2019.

[2] D. Zhang, L. Yao, K. Chen, and J. Monaghan, "A convolutional recurrent attention model for subject-independent EEG signal analysis," *IEEE Signal Process. Lett.*, vol. 26, no. 5, pp. 715–719, May 2019.

[3] S. Ding, Z. Yuan, P. An, G. Xue, W. Sun, and J. Zhao, "Cascaded convolutional neural network with attention mechanism for mobile EEG-based driver drowsiness detection system," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2019, pp. 1457–1464.

[4] P. Nejedly et al., "Exploiting graphoelements and convolutional neural networks with long short term memory for classification of the human electroencephalogram," *Sci. Rep.*, vol. 9, no. 1, p. 11383, Aug. 2019.

[5] G. Zhang, V. Davoodnia, A. Sepas-Moghaddam, Y. Zhang, and A. Etemad, "Classification of hand movements from EEG using a deep attention-based LSTM network," *IEEE Sensors J.*, vol. 20, no. 6, pp. 3113–3122, Mar. 2020.

[6] M. M. N. Mannan, M. A. Kamran, S. Kang, and M. Y. Jeong, "Effect of EOG signal filtering on the removal of ocular artifacts and EEG-based brain-computer interface: A comprehensive study," *Complexity*, vol. 2018, pp. 1–18, Jul. 2018.

[7] D. Labate, F. La Foresta, N. Mammone, and F. C. Morabito, "Effects of artifacts rejection on EEG complexity in Alzheimer's disease," in *Proc. Adv. Neural Netw., Comput. Theor. Issues*, 2015, pp. 129–136.

[8] M. A. Klados, C. Papadelis, C. Braun, and P. D. Bamidis, "REG-ICA: A hybrid methodology combining blind source separation and regression techniques for the rejection of ocular artifacts," *Biomed. Signal Process. Control*, vol. 6, no. 3, pp. 291–300, Jul. 2011.

[9] C. Marque, C. Bisch, R. Dantas, S. Elayoubi, V. Brosse, and C. Pérot, "Adaptive filtering for ECG rejection from surface EMG recordings," *J. Electromyogr. Kinesiol.*, vol. 15, no. 3, pp. 310–315, Jun. 2005.

[10] B. Somers, T. Francart, and A. Bertrand, "A generic EEG artifact removal algorithm based on the multi-channel Wiener filter," *J. Neural Eng.*, vol. 15, no. 3, Feb. 2018, Art. no. 036007.

[11] H. Shahabi, S. Moghimi, and H. Zamiri-Jafarian, "EEG eye blink artifact removal by EOG modeling and Kalman filter," in *Proc. 5th Int. Conf. Biomed. Eng. Informat.*, Oct. 2012, pp. 496–500.

[12] K. T. Sweeney, S. F. McLoone, and T. E. Ward, "The use of ensemble empirical mode decomposition with canonical correlation analysis as a novel artifact removal technique," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 97–105, Jan. 2013.

[13] B. Somers and A. Bertrand, "Removal of eye blink artifacts in wireless EEG sensor networks using reduced-bandwidth canonical correlation analysis," *J. Neural Eng.*, vol. 13, no. 6, Oct. 2016, Art. no. 066008.

[14] W. De Clercq, A. Vergult, B. Vanrumste, W. Van Paesschen, and S. Van Huffel, "Canonical correlation analysis applied to remove muscle artifacts from the electroencephalogram," *IEEE Trans. Biomed. Eng.*, vol. 53, no. 12, pp. 2583–2587, Nov. 2006.

[15] M. Unser and A. Aldroubi, "A review of wavelets in biomedical applications," *Proc. IEEE*, vol. 84, no. 4, pp. 626–638, Apr. 1996.

[16] M. Dora and D. Holcman, "Adaptive single-channel EEG artifact removal with applications to clinical monitoring," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 286–295, 2022.

[17] P. Gajbhiye, N. Mingchinda, W. Chen, S. C. Mukhopadhyay, T. Wilaiprasitporn, and R. K. Tripathy, "Wavelet domain optimized Savitzky–Golay filter for the removal of motion artifacts from EEG recordings," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[18] S. Dash et al., "Dyadic boundary points based empirical wavelet transform for the elimination of eye movement and eye blink-based ocular artifacts from EEG signals," *Biomed. Signal Process. Control*, vol. 85, Aug. 2023, Art. no. 104996.

[19] P. J. Allen, O. Josephs, and R. Turner, "A method for removing imaging artifact from continuous EEG recorded during functional MRI," *NeuroImage*, vol. 12, no. 2, pp. 230–239, Aug. 2000.

[20] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, Aug. 2013.

[21] J. Amann, A. Blasimme, E. Vayena, D. Frey, and V. I. Madai, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Med. Informat. Decis. Making*, vol. 20, no. 1, pp. 1–9, Dec. 2020.

[22] Z. Zhang, A. Romero, M. J. Muckley, P. Vincent, L. Yang, and M. Drozdzal, "Reducing uncertainty in undersampled MRI reconstruction with active acquisition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2049–2053.

[23] E. Zheng, Q. Yu, R. Li, P. Shi, and A. Haake, "A continual learning framework for uncertainty-aware interactive image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 7, pp. 6030–6038.

[24] C. Li, Z. Deng, R. Song, X. Liu, R. Qian, and X. Chen, "EEG-based seizure prediction via model uncertainty learning," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 180–191, 2022.

[25] Y. Liu, Y. Lin, Z. Jia, Y. Ma, and J. Wang, "Representation based on ordinal patterns for seizure detection in EEG signals," *Comput. Biol. Med.*, vol. 126, Nov. 2020, Art. no. 104033.

[26] C. Zhao, J. Li, and Y. Guo, "SleepContextNet: A temporal context network for automatic sleep staging based single-channel EEG," *Comput. Methods Programs Biomed.*, vol. 220, Jun. 2022, Art. no. 106806.

[27] H. Fan et al., "Multiscale vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6804–6815.

[28] K. Zeng, D. Chen, G. Ouyang, L. Wang, X. Liu, and X. Li, "An EEMD-ICA approach to enhancing artifact rejection for noisy multivariate neural data," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 6, pp. 630–638, Jun. 2016.

[29] X. Chen, Q. Chen, Y. Zhang, and Z. J. Wang, "A novel EEMD-CCA approach to removing muscle artifacts for pervasive EEG," *IEEE Sensors J.*, vol. 19, no. 19, pp. 8420–8431, Oct. 2019.

[30] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention based spatial–temporal graph convolutional networks for traffic flow forecasting," in *Proc. 33rd AAAI Conf. Artif. Intell.*, vol. 33, Jan. 2019, pp. 922–929.

[31] S. Li et al., "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[32] B. Yang, K. Duan, C. Fan, C. Hu, and J. Wang, "Automatic ocular artifacts removal in EEG using deep learning," *Biomed. Signal Process. Control*, vol. 43, pp. 148–158, May 2018.

[33] J. Yu, C. Li, K. Lou, C. Wei, and Q. Liu, "Embedding decomposition for artifacts removal in EEG signals," *J. Neural Eng.*, vol. 19, no. 2, Apr. 2022, Art. no. 026052.

[34] W. Sun, Y. Su, X. Wu, and X. Wu, "A novel end-to-end 1D-ResCNN model to remove artifact from EEG signals," *Neurocomputing*, vol. 404, pp. 108–121, Sep. 2020.

[35] H. Zhang, C. Wei, M. Zhao, Q. Liu, and H. Wu, "A novel convolutional neural network model to remove muscle artifacts from EEG," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 1265–1269.

[36] P. Sawangjai et al., "EEGANet: Removal of ocular artifacts from the EEG signal using generative adversarial networks," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 10, pp. 4913–4924, Oct. 2022.

[37] W. Zhang, W. Yang, X. Jiang, X. Qin, J. Yang, and J. Du, "Two-stage intelligent multi-type artifact removal for single-channel EEG settings: A GRU autoencoder based approach," *IEEE Trans. Biomed. Eng.*, vol. 69, no. 10, pp. 3142–3154, Oct. 2022.

[38] Z. Zhang, X. Yu, X. Rong, and M. Iwata, "A novel multimodule neural network for EEG denoising," *IEEE Access*, vol. 10, pp. 49528–49541, 2022.

[39] A. Kendall and Y. Gal, "What uncertainties do we need in Bayesian deep learning for computer vision?" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[40] Y. Gal and Z. Ghahramani, "Bayesian convolutional neural networks with Bernoulli approximate variational inference," 2015, *arXiv:1506.02158*.

[41] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2015, pp. 1613–1622.

[42] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *Proc. Int. Conf. Mach. Learn. (PMLR)*, 2016, pp. 1050–1059.

[43] L. Fiorillo, P. Favaro, and F. D. Faraci, "DeepSleepNet-lite: A simplified automatic sleep stage scoring model with uncertainty estimates," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 2076–2085, 2021.

[44] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[45] D. P. Kingma, T. Salimans, and M. Welling, "Variational dropout and the local reparameterization trick," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–9.

[46] S. Khalighi, T. Sousa, J. M. Santos, and U. Nunes, "ISRUC-sleep: A comprehensive public dataset for sleep researchers," *Comput. Methods Programs Biomed.*, vol. 124, pp. 180–192, Feb. 2016.

[47] I. Obeid and J. Picone, "The temple university hospital EEG data corpus," *Frontiers Neurosci.*, vol. 10, p. 196, May 2016.

[48] S. Tang et al., "Self-supervised graph neural networks for improved electroencephalographic seizure analysis," 2021, *arXiv:2104.08336*.

[49] Z. Jia, Y. Lin, J. Wang, X. Wang, P. Xie, and Y. Zhang, "SalientSleepNet: Multimodal salient wave detection network for sleep staging," 2021, *arXiv:2105.13864*.

[50] Z. Jia et al., "Multi-view spatial–temporal graph convolutional networks with domain generalization for sleep stage classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 1977–1986, 2021.