

Self-supervised learning with noisy dataset for Rydberg microwave sensors denoising

Zongkai Liu(刘宗凯), Qiming Ren(任启明), Wenguang Yang(杨文广), Yanjie Tong(仝艳杰), Huizhen Wang(王惠珍), Yijie Zhang(张毅杰), Ruohao Zhi(智若灏), Junyao Xie(谢君尧), Mingyong Jing(景明勇), Hao Zhang(张好), Liantuan Xiao(肖连团), Suotang Jia(贾锁堂), Ke Tang(唐珂), Linjie Zhang(张临杰)
Citation: Chin. Phys. B. 35, 083301 (2026). DOI: 10.1088/1674-1056/ae6cd1
http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Excitation and ionization of OCS molecules in strong UV and NIR laser fields

Huijun Shi(师慧军), Yang Liu(刘洋), Tian Sun(孙添), Hang Lv(吕航), Haifeng Xu(徐海峰)
Chin. Phys. B. 33, 073301 (2024). DOI: 10.1088/1674-1056/ad3035

Facilitation of controllable excitation in Rydberg atomic ensembles

Han Wang(王涵), Jing Qian(钱静)
Chin. Phys. B. 32, 083302 (2023). DOI: 10.1088/1674-1056/acd0a6

High-resolution photoassociation spectroscopy of ultracold Cs_2 long-range 0_g^- state： The external well potential depth

Liu Wen-Liang(刘文良), Wu Ji-Zhou(武寄洲), Ma Jie(马杰), Xiao Lian-Tuan(肖连团), Jia Suo-Tang(贾锁堂)
Chin. Phys. B. 23, 013301 (2014). DOI: 10.1088/1674-1056/23/1/013301

SPECIAL TOPIC — Quantum frontiers with Rydberg atoms

Self-supervised learning with noisy dataset for Rydberg microwave sensors denoising

Zongkai Liu(刘宗凯)[1,2,†,‡], Qiming Ren(任启明)[1,2,†], Wenguang Yang(杨文广)[1,2], Yanjie Tong(仝艳杰)[1,2], Huizhen Wang(王惠珍)[1,2], Yijie Zhang(张毅杰)[1,2], Ruohao Zhi(智若灏)[1,2], Junyao Xie(谢君尧)[1,2], Mingyong Jing(景明勇)[1,2], Hao Zhang(张好)[1,2], Liantuan Xiao(肖连团)[1,2], Suotang Jia(贾锁堂)[1,2], Ke Tang(唐珂)[3], and Linjie Zhang(张临杰)[1,2,4,§]

[1] *State Key Laboratory of Quantum Optics Technologies and Devices, Institute of Laser Spectroscopy, Shanxi University, Taiyuan* 030006, *China*
[2] *Collaborative Innovation Center of Extreme Optics, Shanxi University, Taiyuan* 030006, *China*
[3] *Guangdong Provincial Key Laboratory of Brain-inspired Intelligent Computation, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen* 518055, *China*
[4] *Hefei National Laboratory, Hefei* 230088, *China*

(Received 25 March 2026; revised manuscript received 27 April 2026; accepted manuscript online 13 May 2026)

We report a self-supervised deep learning framework for Rydberg sensors that enables single-shot noise suppression matching the accuracy of multi-measurement averaging. The framework eliminates the need for clean reference signals (rarely available in quantum sensing) by training on two sets of noisy signals with identical statistical distributions. When evaluated on Rydberg sensing datasets, the framework outperforms wavelet transform and Kalman filtering, achieving a denoising effect equivalent to 10000-set averaging while reducing the computation time by three orders of magnitude. We further validate performance across diverse noise profiles and quantify the complexity-performance trade-off of U-Net and Transformer architectures, providing actionable guidance for optimizing deep learning-based denoising in Rydberg sensor systems.

Keywords: Rydberg atom, microwave sensor, self-supervised learning, denoising

PACS: 33.80.Rv, 33.40.+f

DOI: 10.1088/1674-1056/ae6cd1 **CSTR:** 32038.14.CPB.ae6cd1

1. Introduction

Rydberg atom microwave sensors[1–8] are promising for applied fields such as wideband communications and radar, thanks to their ultra-high microwave electric field sensitivity,[9,10] broad frequency response (direct current to THz),[11–13] and International System of Units (SI)-traceable calibration.[14–17] However, their utility in open-environment deployments is limited by noise sources that submerge the signal, such as environmental electromagnetic interference, interatomic interaction noise, and thermal noise.[3,7,18]

Traditional denoising approaches are ill-suited to this constraint. Multi-measurement averaging suppresses random noise but requires hundreds of repetitions, sacrificing temporal resolution and rendering it incompatible with time-varying signals.[19,20] Wavelet transform (based on thresholding and signal matching)[21–25] and Kalman filtering (dependent on pre-defined state-space models)[26] perform well for stationary noise but fail to stabilize under the non-stationary noise encountered in real-world Rydberg sensor operation.

Deep learning has emerged as an alternative for adaptive denoising:[27] recurrent neural networks (RNNs),[28–30] long short-term memory (LSTM) networks[31,32] and 1D convolutional neural network models[33–38] leverage temporal dependencies to separate signals from noise (e.g., in speech processing[39–43]), while Transformers use self-attention to balance noise suppression and signal detail preservation.[44–47] However, a standard deep learning structure (e.g., autoencoders[48–50]) requires paired noisy-clean signal datasets for supervised learning, where clean single-shot signals are experimentally inaccessible for Rydberg sensors. Moreover, theoretically constructed idealized clean data deviate from actual experimental conditions, which readily leads to model overfitting.[51,52]

Here, we develop a self-supervised deep learning framework for denoising Rydberg atom microwave sensor signals, as shown in Fig. 1(a). The proposed approach follows the Noise2Noise paradigm, and the Transformer architecture is adopted to better capture long-range dependencies, enabling improved preservation of weak spectral features under non-stationary noise. It is validated on Rydberg sensing signals in both the time and frequency domains and is compared with

[†]These authors contributed equally to this work.
[‡]Corresponding author. E-mail: lzk1997@sxu.edu.cn
[§]Corresponding author. E-mail: zlj@sxu.edu.cn

© 2026 Chinese Physical Society and IOP Publishing Ltd. All rights, including for text and data mining, AI training, and similar technologies, are reserved.

10000-set averaging, Kalman filtering, and wavelet filtering. It has three key advances for applied use: (1) it enables single-shot denoising with accuracy matching 10000-set averaging, preserving temporal resolution for dynamic measurements; (2) it eliminates the need for clean reference signals by training on two sets of noisy signals with identical statistical distributions, resolving the labeled data scarcity in quantum sensing; (3) it reduces computation time by three orders of magnitude compared to multi-measurement averaging. Finally, we quantify the complexity-performance trade-off of U-Net and Transformer architectures to guide model selection for specific sensor deployment needs. This work advances the practical utility of Rydberg sensors by bridging deep learning denoising with applied quantum sensing constraints.

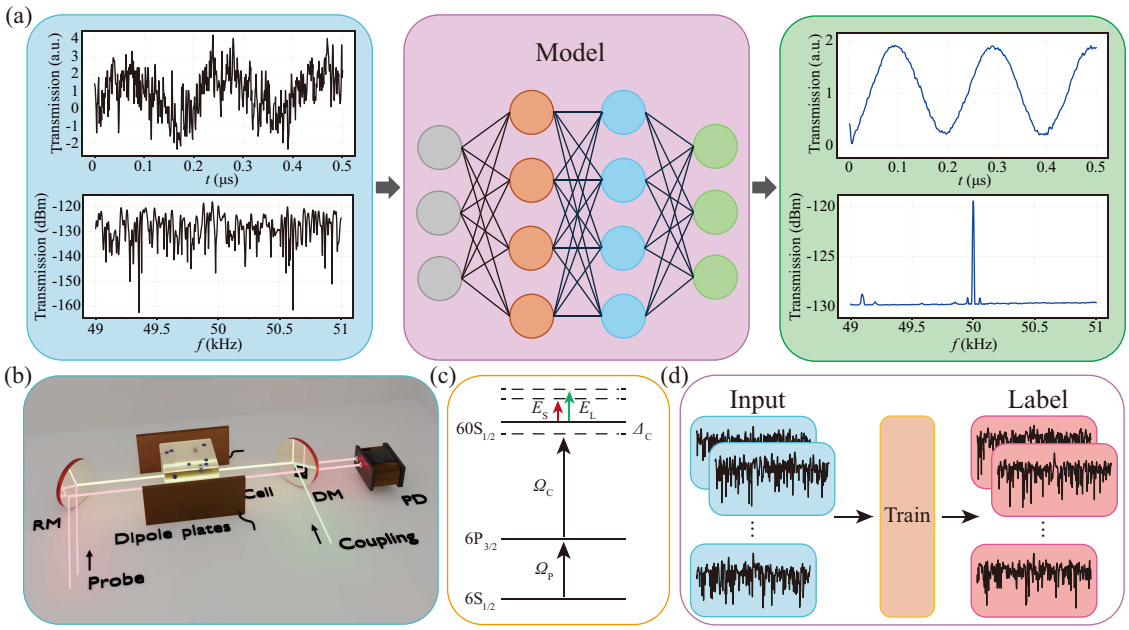


Fig. 1. The process of model testing and training. (a) A trained model is capable of generating a clean signal from an input noisy signal in both the time (upper panel) and frequency (lower panel) domains. The data are collected from experiments. (b) Experimental setup. The probe light is split into two fields: one counter-propagates with the coupling light inside a cesium vapor cell, while the other outside the cell acts as a reference field to enable balanced detection. The time-domain and frequency-domain data are collected by an oscilloscope and spectrum analyzer after balanced detection. The signal electric field E_S and local oscillator electric field E_L interact with the atoms via parallel-plate electrodes. RM-reflecting mirror, DM-dichroic mirror, PD-balanced photodetector. (c) Energy level diagram. The probe light (Rabi frequency Ω_p) and coupling light (Rabi frequency Ω_c, frequency detuning from resonance Δ_c) excite atoms from the ground state $6S_{1/2}$ to the Rydberg state $60S_{1/2}$ via the intermediate state $6P_{3/2}$. The signal and local fields with amplitudes E_S and E_L and frequencies of 63 MHz and 63.05 MHz act on the Rydberg state distribution. (d) During training, the input data and their corresponding labels are independent measurement results of the microwave signal from the same Rydberg atom, differing only in identically independently distributed noise.

2. Methods

2.1. Experimental setup

A Rydberg atom microwave detection system was constructed to acquire noisy intermediate frequency (IF) signals for validating deep learning denoising methods. The experimental setup and energy level diagram are shown in Figs. 1(b) and 1(c), respectively. Cesium atoms were confined in a room-temperature glass cell ($5 \times 5 \times 5$ cm^3). A probe field (wavelength $\lambda = 852$ nm, Rabi frequency $\Omega_p/(2\pi) = 9.06$ MHz) excited atoms from $|6S_{1/2}, F = 4\rangle$ to $|6P_{3/2}, F = 5\rangle$, while a coupling field (wavelength $\lambda = 510$ nm, Rabi frequency $\Omega_p/(2\pi) = 0.83$ MHz) further excited atoms to the $60S_{1/2}$ Rydberg state. The probe field was split into two: one as a reference outside the cell, and the other counter-propagating with the coupling field inside to form an electromagnetically induced transparency (EIT) configuration, where the coupling field modulated the probe field's atomic transmission and absorption. The probe field and the reference field were balanced-detected by a differential detector (bandwidth 1 MHz and common-mode rejection ratio 40 dB), yielding spectral signals. Both fields' frequencies were stabilized by an ultra-stable cavity (finesse $= 2 \times 10^5$) and locked to the aforementioned atomic transitions.

The experiment was conducted at room temperature in a microwave anechoic chamber lined with pyramidal foam absorbers, which provides an attenuation of < 20 dB for radio frequency (RF) fields at frequencies < 100 MHz incident normally on the material. RF electric field detection adopted an atomic heterodyne scheme: an arbitrary wave generator produced a 63 MHz signal under test (SUT) E_S and a local oscillator (LO) signal E_L with a frequency detuned from the SUT by 50 kHz. These signals were input to parallel brass electrodes (10 cm $\times$ 8 cm in size, 5 cm spacing) to generate an RF field applied to the atomic system. The atomic energy level is shifted by the SUT and local fields via the ac Stark shift, which is deduced from the probe transmission.[7,53] The shift

frequency δ is given by

$$\delta = -\alpha |E_S + E_L|^2/2, \tag{1}$$

where α is the scalar polarizability. The beating between SUT E_S and local field E_L generates a 50 kHz IF signal whose amplitude corresponds to that of the SUT. Here we set the E_L to be phase-synchronized with the E_S in our experiment, as $\phi_{LO} = \phi_{sig}$. By averaging over time, we get

$$\bar{\delta} = \bar{\delta}_0 - \frac{1}{2}\alpha E_L E_S \cos(\Delta\omega \cdot t), \tag{2}$$

where $\Delta\omega$ is the frequency difference between the measured electric field E_S and the local field E_L, and $\bar{\delta}_0 = -\frac{1}{4}\alpha\left(E_L^2 + E_S^2\right)$ is the average ac Stark shift caused by the LO and signal electric field.

The probe field is detected by a balanced differential detector, which converts the light intensity into a voltage signal P. The amplitude of P is proportional to the product of the local field strength and the measured electric field strength $E_S E_L$, which can be expressed as

$$P \propto E_L E_S \cos(2\pi\Delta\omega \cdot t). \tag{3}$$

Therefore, with the intensity of the local field E_L kept constant, the amplitude of the difference-frequency signal P can be directly used to deduce the magnitude of the signal field E_S, enabling the measurement of the signal field strength.

This signal is then read out in the time domain using an oscilloscope and in the frequency domain using a spectrum analyzer for deep learning-based denoising. The spectrum analyzer is configured with a resolution bandwidth (RBW) of 10 Hz, 5 averaging cycles for spectrum acquisition, and a frequency sweep range of 49–51 kHz with a step size of 5 Hz. In addition, different intensities of attenuation were applied to the IF signal to simulate open electromagnetic environments, where single-shot signals were submerged in noise but recoverable via multi-measurement averaging.

2.2. Dataset design and training strategy

The primary goal of training the neural network $f_\theta(\cdot)$ is to learn an optimized weight parameter ($\theta^\star$) (encompassing the weights and biases) that minimizes the discrepancy between the network's outputs and the experimental training labels (minimizes the loss function $\mathcal{L}(\cdot)$). Mathematically, the training objective is formalized as a constrained minimization problem, and the optimization process is as follows:

$$\theta^\star = \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} \mathcal{L}\left(f_\theta(x_{\text{train},i}), y_{\text{train},i}\right), \tag{4}$$

with the size of the training dataset N, $x_{\text{train}} = x + n_1$, $y_{\text{train}} = y + n_2$, where x_{train} and y_{train} denote two independently measured experimental datasets, which consist of ideal clean signals x, y and noise n_1, n_2 drawn from an independently and identically distributed noise distribution, i.e., $n_1, n_2 \in$ Noise. The distinction between self-supervised learning and supervised learning lies in their training labels. For supervised learning (e.g., autoencoders[49,54]), the training label is an ideal, noise-free reference $y_{\text{train}} = y$ (such references are rarely obtainable in sensing scenarios). In contrast, the label used in self-supervised learning (in this work) corresponds to another measurement outcome under the same conditions as the training data, $y_{\text{train}} = y + n_2$, which eliminates the need for a clean reference. The noise n_1, n_2 includes environmental electromagnetic interference, interatomic interaction noise, and thermal noise, etc.[3] To adjust the effect of the noise on the signal, we applied attenuation to the IF signal, ranging from a scenario where the IF signal is above the noise floor to one where it is submerged by noise but remains recognizable after 10000 averaging. After training converges (defined as the training loss remaining stable for 10 consecutive epochs), the optimized weights $\theta^\star$ of the deep learning layers are obtained. The prediction process leverages unseen test inputs to output, formulated as

$$\hat{y}_{\text{test}} = f_{\theta^\star}(x_{\text{test}}), \tag{5}$$

where x_{test} is independent of x_{train} to ensure unbiased generalization evaluation, and the prediction result $\hat{y}_{\text{test}}$ is compared to the 10000-set average of the original data.

Specifically, a Transformer-based end-to-end model was built, along with a dedicated dataset and training protocol, with key details below. In uniform experiments, 10000 sets of independent and identically distributed (i.i.d.) noisy spectral/time-domain data (1000 consecutive points each set) were collected and partitioned into training, labeling, and testing sets at a 4:4:2 ratio. Specifically, the training set includes measurement result data x_{train} (dimension: 4000, 1000, 1), and another independent set of measurement results y_{train} (dimension: 4000, 1000, 1), while the test set consists of x_{test} (dimension: 2000, 1000, 1). Here $x_{(\cdot)}$ and $y_{(\cdot)}$ denote the training data and their corresponding labels, respectively; the first dimension of the data represents the number of sets, the second dimension denotes the number of consecutive points, and the third dimension is a requirement of the deep learning model. The unlabeled testing data x_{test} were tested against the average of 10000 original data sets.

Next, to improve the stability of the training and avoid magnitude bias, each data point was standardized: $x' = \frac{x-\mu}{\sigma}$, where x = original data, μ = mean, σ = standard deviation, and x' = standardized data. The data are then fed into a deep learning model for training, whose architecture is shown in Appendix A.

2.3. Why the training process reduces noise in noisy datasets

To clarify the noise reduction mechanism of the proposed training process, we analyze the optimization of the objective function combined with the statistical properties of noise and the law of large numbers (LLN), as detailed below (adopting MSE loss $\mathcal{L}(a,b)=(a-b)^2$ for simplicity, extendable to other loss functions).

First, we assume that the noise $n_1, n_2 \in$ Noise satisfies two general properties: (1) Zero mean: $E[n_1]=E[n_2]=0$ (no systematic bias); (2) Independence from clean signals: $E[x\cdot n_1]=E[y\cdot n_2]=0$ (noise does not carry signal-related information). The training goal is to minimize the empirical loss

$$\min_{\theta} J(\theta)=\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}\mathcal{L}\left(f_\theta(x_{\text{train},i}),y_{\text{train},i}\right). \tag{6}$$

Substituting $x_{\text{train},i}=x_i+n_{1,i}$ and $y_{\text{train},i}=y_i+n_{2,i}$, where x_{train} and y_{train} are independent measurement results corresponding to the microwave signal of the same Rydberg atom and differ only in the noise, and x and y are clean data. The objective function becomes

$$J(\theta)=\frac{1}{N}\sum_{i=1}^{N}\left[f_\theta(x_i+n_{1,i})-(y_i+n_{2,i})\right]^2. \tag{7}$$

For large N, the empirical risk converges to the expected risk by LLN

$$J(\theta)\xrightarrow{N\to\infty}E\left[\left(f_\theta(x+n_1)-(y+n_2)\right)^2\right]. \tag{8}$$

Expanding the expectation term and applying the noise assumptions

$$E\left[\cdots\right]=E\left[\left(f_\theta(x+n_1)-y\right)^2\right]+\sigma_{n2}^2, \tag{9}$$

where the cross term vanishes ($E[n_2]=0$ and independence), and $E[n_2^2]=\sigma_{n2}^2$ (constant noise variance, irrelevant to θ). Since σ_{n2}^2 is fixed, minimizing $J(\theta)$ is equivalent to minimizing the signal fitting error

$$E\left[\left(f_\theta(x+n_1)-y\right)^2\right]. \tag{10}$$

To achieve this, the model must learn a noise-agnostic mapping

$$f_\theta(x+n_1)\to y.$$

Consequently, random fluctuations of n_1 and n_2 are averaged by the LLN (for large N), and the model spontaneously suppresses noise by focusing on the clean $x-y$ relationship rather than fitting the noise. This explains why the training process reduces noise in noisy datasets.

3. Results

3.1. Denoising performance on frequency-domain signals

The deep learning model enables denoising of quantum sensing data for both time-domain and frequency-domain signals. Figure 2 presents its denoising results for IF frequency-domain signals measured by the spectrum analyzer, compared with the averaged result of 10000 repeated datasets with i.i.d. noise. From Figs. 2(a)–2(d), the amplitude of the SUT on the electrodes is gradually attenuated from 200 mV to 3 mV, which is used to simulate the scenario where the data are gradually submerged by noise. Since the IF signal arises from the

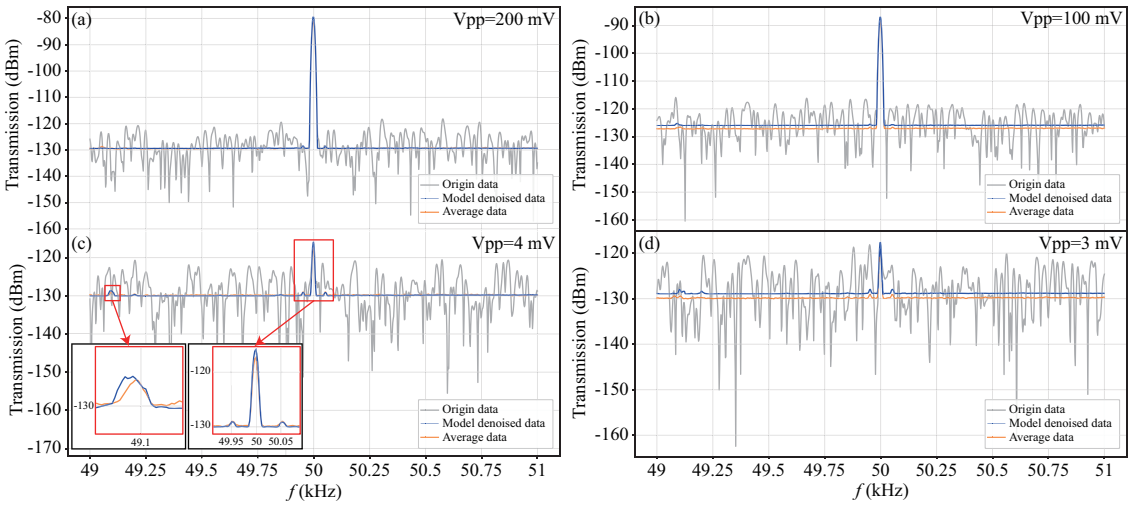


Fig. 2. Intermediate frequency (IF) frequency-domain signals under different attenuation levels. IF signal intensities are labeled in the top-right corner of subfigures (a)–(d). Gray curves: Single-shot measurements of probe field transmission signals (attenuated from above the noise floor to the noise floor level). Orange curves: Averaged results of 10000 measurements (noise-free ground truth, not used in model training), showing noise elimination with only the IF signal retained (plus IF sidebands and 49 kHz weak signals under weak attenuation in panels (c)–(d)). Blue curves: Denoising results of the Transformer-based model (trained solely on noisy signals), achieving performance comparable to multi-measurement averaging and revealing the IF signal, IF sidebands, and 49.1 kHz weak signals. The two red boxes in panel (c) are zoomed-in views of the averaged and deep learning denoising results at frequencies of 50 kHz and 49.1 kHz, respectively, to illustrate the retained submerged signal.

beating of the SUT signal and the local signal, and the frequency of the local signal is adjustable to track the SUT, the frequency of the IF signal is fixed at 50 kHz. Without loss of generality, this configuration can be generalized to SUTs with other frequencies.

For strong-amplitude signals (Figs. 2(a) and 2(b)), the model restores noisy data to a level comparable to the 10000-measurement average, while fully preserving the 50 kHz intermediate frequency (IF) signal. The model outperformed multi-measurement averaging in temporal efficiency: measurement time was reduced by three orders of magnitude (1 hour for 10000 measurements + averaging vs. 0.7 second for deep learning denoising). For spectral data with 40 dB attenuation in amplitude (Figs. 2(c) and 2(d)), the original signal is completely submerged in noise; however, the model still recovers the IF signal and retains its characteristic components. Notably, non-noise components are retained, as shown in the red box and insets in Fig. 2(c): around the 50 kHz IF signal (two symmetric small peaks) and in the 49 kHz to 49.25 kHz range (small peaks), signals emerge from noise in both the averaged and model-denoised results. These are likely characteristic signals, indicating that the model can identify not only post-averaging signals far above the noise floor but also those close to it and submerged by it. The deep learning model successfully recognizes these characteristic signals that are consistent with those of the averaged results. This consistency not only validates the model's ability to capture meaningful patterns but also confirms that the denoising process is not a spurious artifact of the deep learning framework. In other frequency regions, the denoised results are flatter than the average, showing superior noise suppression in noise-dominated ranges.

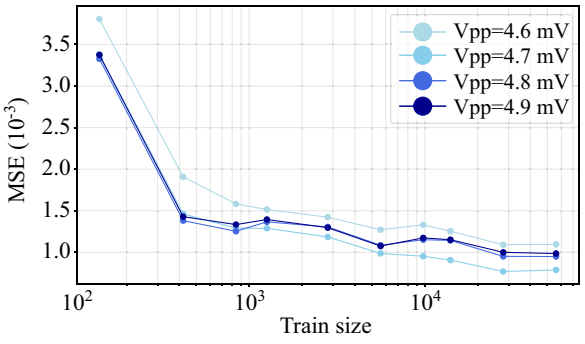


Fig. 3. MSE depends on the size of the training dataset in the frequency domain. This panel presents the MSE between the denoised results of the deep learning model and the results of 10000-set averaging, under different voltages applied to the dipole plates from 4.6 mV to 4.9 mV. As the training dataset size increases, the MSE decreases.

Denoising performance was quantified via mean squared error (MSE, against the 10000-measurement average)

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y}_i)^2, \tag{11}$$

where N is the number of data points, y_i is the denoised output, and $\bar{y}_i$ is the corresponding 10000-measurement average reference serving as the ground truth. In Fig. 3, the number of training data points is varied from 100 to 5×10^4 for peak-to-peak voltage Vpp ranging from 4.6 mV to 4.9 mV. After each training process, the MSE is calculated to evaluate how well the denoised results align with the averaging result. It decreases significantly when the number of training data points exceeds 500, indicating that a moderate increase in training data volume can drastically boost denoising efficacy. When the number of training data points exceeds 1×10^4, the improvement in MSE begins to plateau, indicating diminishing returns from further increasing the dataset size. This suggests that a training set an order of magnitude smaller is already sufficient to achieve satisfactory denoising performance.

In summary, the model achieves noise suppression performance comparable to that of multi-measurement averaging. It uniquely enables the extraction of noise-submerged signals and the identification of potential characteristic signals from single-shot data.

3.2. Quantitative validation of denoising performance: deep learning vs. traditional methods

Figure 4 validates the proposed deep learning model's superiority over traditional methods (average against 10000 sets, wavelet transform and Kalman filter) in suppressing noise in time series (recorded by the oscilloscope), using IF time-domain signals from Rydberg atom microwave heterodyne detection.

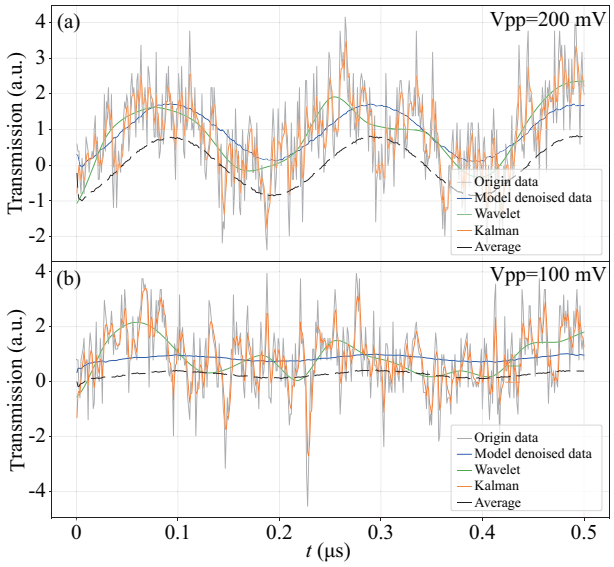


Fig. 4. Different denoising models for time-domain signals at Vpp = 200 mV (a) and 100 mV (b). The adopted denoising methods include wavelet denoising (green solid line), Kalman filtering (orange solid line), multi-measurement averaging (black dashed line), and the proposed model's denoising results (blue solid line), all applied to the single-shot measurement result of the probe field transmission signal (gray solid line). The parameters of wavelet denoising and Kalman filtering were optimized by minimizing the error between their denoising results and the multi-measurement averaging result, as detailed in the main text.

Wavelet transform denoising[24,55] (green solid line) used a Coiflet wavelet (4th type) with a threshold of 0.55. Kalman

filtering[26] (orange solid line) used the following parameters: process noise variance $Q = 0.0175$ and measurement noise variance $R = 0.06$. To make a fair comparison, the above parameters of traditional methods were optimized by grid search to minimize the MSE between the denoised data obtained by these methods and the averaged result of 10000 repeated measurements with the same IF signal and i.i.d. noise.

The 10000-measurement average (black dashed line) was visually smoother than the single-shot result (gray solid line), with noise suppressed and the IF signal prominent. The deep learning model's denoised results (blue solid line) were closer to this average than those of traditional methods, yielding smoother curves for noisy single-frequency IF signals with three orders of magnitude higher temporal efficiency.

Table 1. MSE for different denoising methods for time-domain data at Vpp = 200 mV and 100 mV.

	Deep learning	Kalman filtering	Wavelet
Vpp = 200 mV	$\mathbf{7.71 \times 10^{-4} \pm 1.5 \times 10^{-4}}$	$7.0 \times 10^{-2} \pm 1.6 \times 10^{-2}$	$7.1 \times 10^{-2} \pm 2.5 \times 10^{-2}$
Vpp = 100 mV	$\mathbf{7.9 \times 10^{-5} \pm 4.5 \times 10^{-5}}$	$1.4 \times 10^{-1} \pm 3.6 \times 10^{-2}$	$2.5 \times 10^{-1} \pm 6.4 \times 10^{-2}$

For different voltages on the dipole plates, 200 mV, Fig. 4(a), and 100 mV, Fig. 4(b), the MSE relative to the averaging result (mean±standard deviation from 2000 distinct time-domain signals in the test set) for these three methods is shown in Table 1. This two-order-of-magnitude difference in MSE confirms the deep learning model's significant denoising advantage.

3.3. Relationship between model complexity and denoising performance

To investigate the influence of model complexity on denoising performance, Fig. 5 compares two deep learning architectures (i.e., Transformer and U-Net models[56–58]), with their structure diagrams shown in Appendix A, on noisy time- and frequency-domain data. The Transformer model (U-Net model) has parameter sizes of 4.50×10^5 (1.59×10^4) and training times of 2000 s per epoch (10 s per epoch), respectively; other parameters (e.g., epochs, dataset configuration, and optimizer) were kept consistent.

In both domains, the denoised results of the Transformer-based model are closer to the 10000-measurement average than those of its counterpart. For the time domain, the alternative architecture fails to capture IF signal features, resulting in outputs that contain significant high-frequency noise alongside the 50 kHz oscillation. The MSE of the Transformer model is 7.17×10^{-4} and the MSE of the U-Net model is 1.25×10^{-2}. In contrast, the Transformer model identifies and extracts the shared IF signal while suppressing noise interference. For the frequency domain, only the IF signal is prominent in the results of the alternative architecture. However, the Transformer model retains the IF signal, its sidebands, and the weak 49.1 kHz signal, showing higher similarity to the 10000-measurement average. The MSE of the Transformer model is 6.7×10^{-5} and the MSE of the U-Net model is 2.45×10^{-4}. These results demonstrate that increased computational complexity (quantified by the number of training parameters) enhances weak signal extraction, despite longer training times. This enables more robust recovery of weak signals from noise; future work may adopt deeper or higher-complexity architectures to further enhance performance.

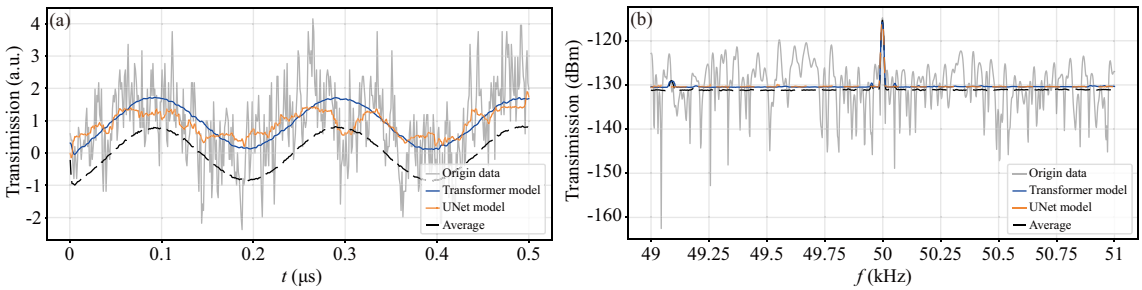


Fig. 5. Denoising results of different deep learning models under the same self-supervised architecture. (a) Denoising results of Transformer (blue solid line) and convolution-based U-Net structure (orange solid line), both applied to the single-shot measurement result of the probe field transmission signal (gray solid line); the averaged result of 10000 measurements (black dashed line) serves as the ground-truth reference. (b) Denoising results of Transformer (blue solid line) and convolution-based U-Net structure (orange dashed line) applied to the frequency-domain data of the single-shot measurement result (gray solid line). The Transformer result exhibits not only the intermediate frequency (IF) signal but also its sidebands and the weak 49.1 kHz signal, and is closer to the averaged result than the U-Net.

4. Discussion

In this work, we propose a self-supervised deep learning framework for denoising in Rydberg atom microwave sensors. This framework achieves denoising performance comparable to the averaging method in both the frequency and time domains of microwave sensing signals, while exhibiting extremely low latency. It reduces processing time by three orders of magnitude compared to the averaging method. Compared

with Kalman filtering and wavelet transform-based denoising methods, the proposed approach demonstrates superior denoising performance. Furthermore, by replacing the computational units within the deep learning framework, we identify an optimization direction for further performance enhancement, namely increasing the model's computational complexity.

The performance of the deep learning denoising framework stems inherently from the model's ability to learn the intrinsic properties of the target signals.[59] Rydberg atom microwave signals are constrained by the physical laws governing energy level transitions, exhibiting well-defined underlying structures in their amplitude variations, frequency distributions, and temporal evolution. In contrast, environmental interference and quantum fluctuations manifest as random noise components that follow specific statistical laws. By extracting features of common signals from a large dataset, the model can leverage this prior knowledge to retrieve stable shared components from single noisy measurements while suppressing random noise variations across different samples. This process is equivalent to achieving "virtual averaging" in a data-driven manner, preserving the temporal resolution of single-shot measurements without sacrificing the speed limitations associated with traditional averaging methods.

Compared to conventional techniques, the proposed deep learning-based denoising method for quantum sensing offers three distinct advantages. First, the model inference process is extremely fast, with single-data processing completed within milliseconds, fully meeting the speed requirements for dynamic process monitoring. Second, centered on data-driven learning, the method eliminates the need for established noise models, enabling stronger adaptability to non-stationary noise commonly encountered in open environments. Third, model training can be accomplished without the need for clean reference signals, making it suited for scenarios in the quantum sensing field where labeled data are scarce.

Future research will focus on the following limitations: for the data-driven paradigm, model performance is highly dependent on the quality of the training dataset, and sufficient data covering various noise scenarios are a prerequisite for ensuring denoising effectiveness. Additionally, although the Transformer architecture delivers superior denoising performance, its computational complexity poses challenges for deployment in embedded sensing systems with limited computing resources.

Appendix A: Model structure

The structures of the transformer model and U-Net model are shown in Figs. A1 and A2, respectively.

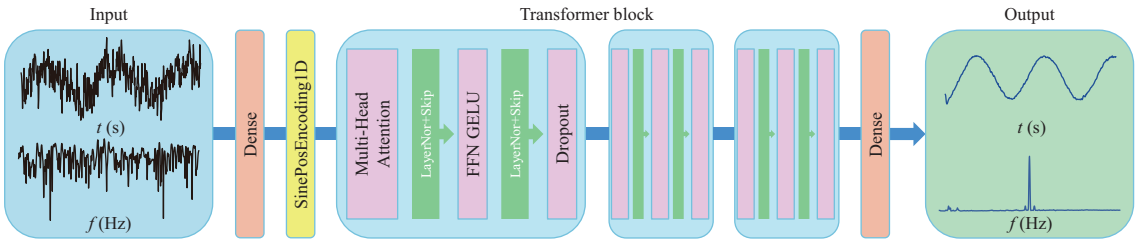


Fig. A1. Transformer-based model structure. Architecture of the proposed Transformer-based model for quantum signal denoising. The pipeline includes: (1) Input embedding and 1D sine positional encoding for the raw noisy signal; (2) Stacked Transformer blocks with multi-head attention, GELU-activated FFNs, residual connections, and dropout; (3) A final dense layer to predict the clean signal, with both time-domain waveform and frequency-domain spectrum shown as outputs.

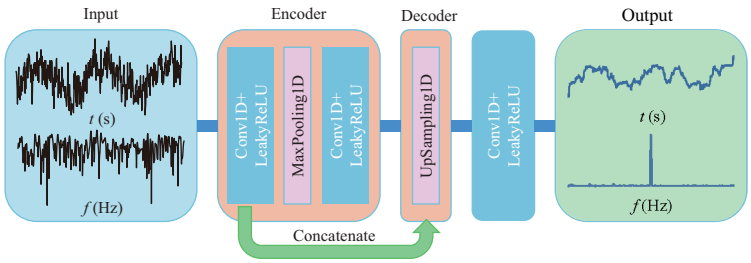


Fig. A2. U-Net structure. The proposed 1D U-Net architecture for quantum signal denoising. The pipeline includes: (1) An encoder with Conv1D + LeakyReLU blocks and MaxPooling1D for hierarchical feature extraction; (2) A decoder with UpSampling1D and Conv1D layers for signal reconstruction; (3) Skip connections between encoder and decoder layers to retain high-resolution information. The network outputs the denoised signal, with its time-domain waveform and frequency spectrum shown.

The structure of the Transformer model is as follows. First, a Dense layer projects the dimension of the data to a 64-dimensional space for feature mapping. Second, a sine encoding layer (SinePosEncoding) assigns unique vectors via sine values to preserve sequence order. Then, three stacked encoder blocks are used, each with 8-head multi-head self-attention for local-global dependencies, a 128-dimensional feed-forward neural network with a Gaussian error linear unit (GELU) acti-

vation function for signal representation, residual connections for fusion, and dropout (with a rate $p = 0.1$). Finally, a Dense layer maps 64-dimensional features back to spectral intensity for denoising. The weights of the above layers were updated via backpropagation using the Adam optimizer (with a decay coefficient for first-order momentum $\beta_1 = 0.9$, a decay coefficient for second-order momentum $\beta_2 = 0.999$, a stability parameter that prevents division by zero $\varepsilon = 10^{-7}$, and a learning rate of 1.0×10^{-5}) with MSE loss ($\mathcal{L}$) between predictions and y_{train}. The hardware for the deep learning model is a CPU with the following specifications: 12th Gen Intel(R) Core(TM) i7-12700 (2.10 GHz).

The U-Net model comprises an encoder, a decoder, and a 1D convolutional layer, where the encoder and decoder are connected via layer concatenation. The encoder consists of three components. The first convolutional layer contains 24 convolution kernels, followed by a LeakyReLU activation function. Subsequently, a max-pooling layer with a pooling size of 2 is applied to halve the temporal resolution. The second convolutional layer further extracts high-level temporal features at the downsampled temporal scale. The decoder restores temporal resolution using a 1D upsampling layer with a scale factor of 2. Via skip connections, the upsampled features are concatenated with the corresponding high-resolution features from the encoder, forming the characteristic U-shaped network structure. This design enables the network to effectively fuse low-level temporal detail information and high-level semantic features. Finally, a 1D convolutional layer (with 48 convolution kernels and a kernel size of 3) is used to fuse multi-scale features, accompanied by the LeakyReLU activation function.

Data availability statement

The datasets generated and/or analyzed during the current study are available in the Zenodo repository, accessible via the following digital object identifier (DOI): 10.5281/zenodo.18091258.

Acknowledgments

Project supported by the National Key R&D Program of China (Grant No. 2022YFA1404003), the National Natural Science Foundation of China (Grant Nos. T2495252, 12104279, 123B2062, and 12574318), the Innovation Program for Quantum Science and Technology (Grant No. 2021ZD0302100), the Fund for Shanxi '1331 Project' Key Subjects Construction, and the Sanjin Talent Program of Shanxi Province (Grant No. SJYC2025280).

References

[1] Sedlacek J A, Schwettmann A, Kübler H, Löw R, Pfau T and Shaffer J P 2012 *Nat. Phys.* **8** 819
[2] Shaffer J P and Kübler H 2018 *Proc. SPIE* **10674** 106740C
[3] Jing M, Hu Y, Ma J, Zhang H, Zhang L, Xiao L and Jia S 2020 *Nat. Phys.* **16** 911
[4] Fancher C T, Scherer D R, John M C S and Marlow B L S 2021 *IEEE Trans. Quantum Eng.* **2** 1
[5] Yuan J, Yang W, Jing M, Zhang H, Jiao Y, Li W, Zhang L, Xiao L and Jia S 2023 *Rep. Prog. Phys.* **86** 106001
[6] Liu B, Zhang L, Liu Z, Deng Z, Ding D, Shi B and Guo G 2023 *EMS* **1** 1
[7] Zhang H, Ma Y, Liao K, Yang W, Liu Z, Ding D, Yan H, Li W and Zhang L 2024 *Sci. Bull.* **69** 1515
[8] Liu B, Zhang L H, Wang Q F, *et al.* 2025 *Chin. Phys. Lett.* **42** 053201
[9] Yang B, Yan Y, Li X, Xiao L, Li X, Chen L, Deng J and Cheng H 2024 *Phys. Rev. Appl.* **21** L031003
[10] Zhou A, Lin Y, Mao R, Yang K, Ding Z, Wan W and Fu Y 2025 *IEEE Trans. Antennas Propag.* **73** 10948
[11] Anderson D A and Raithel G 2017 *Appl. Phys. Lett.* **111** 053504
[12] Adams C S, Pritchard J D and Shaffer J P 2019 *J. Phys. B: At. Mol. Opt. Phys.* **53** 012002
[13] Meyer D H, Castillo Z A, Cox K C and Kunz P D 2020 *J. Phys. B: At. Mol. Opt. Phys.* **53** 034001
[14] Holloway C L, Gordon J A, Jefferts S, Schwarzkopf A, Anderson D A, Miller S A, Thaicharoen N and Raithel G 2014 *IEEE Trans. Antennas Propag.* **62** 6169
[15] Holloway C L, Simons M T, Gordon J A, Dienstfrey A M, Anderson D A and Raithel G 2017 *J. Appl. Phys.* **121** 233106
[16] Anderson D A, Sapiro R E and Raithel G 2021 *IEEE Trans. Antennas Propag.* **69** 5931
[17] Schlossberger N, Prajapati N, Berweger S, Rotunno A P, Artusio-Glimpse A B, Simons M T, Sheikh A A, Norrgard E B, Eckel S P and Holloway C L 2024 *Nat. Rev. Phys.* **6** 606
[18] Wu B, Yao J, Wu F, An Q and Fu Y 2024 *Chin. Phys. B* **33** 024205
[19] Boudreaux-Bartels G and Parks T 1986 *IEEE Trans. Acoust. Speech Signal Process.* **34** 442
[20] Dantus M and Lozovoy V V 2004 *Chem. Rev.* **104** 1813
[21] Pan Q, Zhang L, Dai G and Zhang H 1999 *IEEE Trans. Signal Process.* **47** 3401
[22] Pasti L, Walczak B, Massart D and Reschiglian P 1999 *Chemom. Intell. Lab. Syst.* **48** 21
[23] Grossmann A and Morlet J 2006 *Decomposition of Hardy Functions into Square Integrable Wavelets of Constant Shape* (Princeton: Princeton University Press) pp. 126
[24] Huimin C, Ruimei Z and Yanli H 2012 *Phys. Procedia* **33** 1354
[25] Dautov C P and Özerdem M S 2018 *Proc. 26th Signal Process. Commun. Appl. Conf.* (SIU) pp. 1
[26] Basar T 2001 *A New Approach to Linear Filtering and Prediction Problems* (New York: Wiley) pp. 167
[27] Yu S, Ma J and Wang W 2019 *Geophysics* **84** V333
[28] Grossberg S 2013 *Scholarpedia* **8** 1888
[29] Caterini A L and Chang D E 2018 *Recurrent Neural Networks* (Cham: Springer) pp. 59
[30] Wang J, Li X, Li J, Sun Q and Wang H 2022 *Big Data Res.* **27** 100296
[31] Hochreiter S and Schmidhuber J 1997 *Neural Comput.* **9** 1735
[32] Van Houdt G, Mosquera C and Nápoles G 2020 *Artif. Intell. Rev.* **53** 5929
[33] Wang Z, Yan W and Oates T 2017 *Proc. Int. Jt. Conf. Neural Netw.* (IJCNN) pp. 1578
[34] Kiranyaz S, Ince T, Abdeljaber O, Avci O and Gabbouj M 2019 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (ICASSP) p. 8360
[35] Ismail Fawaz H, Forestier G, Weber J, Idoumghar L and Muller P A 2019 *Data Min. Knowl. Discov.* **33** 917
[36] Chen C C, Liu Z, Yang G, Wu C C and Ye Q 2020 *Electronics* **10** 59
[37] Tonekaboni S, Eytan D and Goldenberg A 2021 arXiv:2106.00750
[38] Zhou P, Liu Y, Liang J, Song Q and Li X 2025 *Proc. 31st ACM SIGKDD Conf. Knowl. Discov. Data Min.* (KDD) (New York: ACM) pp. 4120
[39] Graves A, Mohamed A R and Hinton G 2013 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (ICASSP) pp. 6645
[40] Sak H, Senior A and Beaufays F 2014 arXiv: 1402.1128
[41] Weninger F, Erdogan H, Watanabe S, Vincent E, Le Roux J, Hershey J R and Schuller B 2015 *Proc. Int. Conf. Latent Var. Anal. Signal Sep.* (Springer) pp. 91
[42] Oruh J, Viriri S and Adegun A 2022 *IEEE Access* **10** 30069

[43] Wang J, Saleem N and Gunawan T S 2024 *Cogn. Comput.* **16** 1221
[44] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 *Adv. Neural Inf. Process. Syst.* **30** 5998
[45] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B 2021 *Proc. IEEE/CVF Int. Conf. Comput. Vis.* (ICCV) pp. 9992
[46] Fan C M, Liu T J and Liu K H 2022 *Proc. IEEE Int. Symp. Circuits Syst.* (ISCAS) pp. 2333
[47] Wang H, Lin J, Li Y, Dong X, Tong X and Lu S 2024 *IEEE Trans. Geosci. Remote Sens.* **62** 1
[48] Vincent P, Larochelle H, Bengio Y and Manzagol P A 2008 *Proc. 25th Int. Conf. Mach. Learn.* (ICML) (New York: ACM) pp. 1096
[49] Lu X, Tsao Y, Matsuda S and Hori C 2013 *Proc. Interspeech* pp. 436
[50] Bank D, Koenigstein N and Giryes R 2023 *Machine Learning for Data Science Handbook* pp. 353
[51] Lever J, Krzywinski M and Altman N 2016 *Nat. Methods* **13** 703
[52] Montesinos López O A, Montesinos López A and Crossa J 2022 *Multivariate Statistical Machine Learning Methods for Genomic Prediction* (Cham: Springer) pp. 109
[53] Liu B, Zhang L H, Liu Z K, Zhang Z Y, Zhu Z H, Gao W, Guo G C, Ding D S and Shi B S 2022 *Phys. Rev. Appl.* **18** 014045
[54] Zezario R E, Hussain T, Lu X, Wang H M and Tsao Y 2020 *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* (ICASSP) pp. 6669
[55] Fang H T and Huang D S 2004 *Opt. Commun.* **233** 67
[56] Ronneberger O, Fischer P and Brox T 2015 *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.* (MICCAI) (Springer) pp. 234
[57] Reymann M P, Würfl T, Ritt P, Stimpel B, Cachovan M, Vija A H and Maier A 2019 *Proc. IEEE Nucl. Sci. Symp. Med. Imaging Conf.* (NSS/MIC) pp. 1
[58] Ibtehaz N and Rahman M S 2020 *Neural Netw.* **121** 74
[59] Izadi S, Sutton D and Hamarneh G 2023 *Artif. Intell. Rev.* **56** 5929