Optica Applicata, Vol. LIV, No. 3, 2024
DOI: 10.37190/oa240308

Reconstructing computational spectra using deep learning's self-attention method

Hao Wu[1], Hui Wu[1], Xinyu Su[1], Jingjun Wu[2,*], Shuangli Liu[1,*]

[1]School of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, China

[2]School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

*Corresponding author: liushuangli2301@gmail.com (S.L.), jingjunwu163@163.com (J.W.)

Miniaturized computational spectrometers have become a new research hotspot due to their portability and miniaturization. However, there are several issues, like low precision and poor stability. Because the problem of spectrum reconstruction accuracy is very evident, we suggested a novel approach to raise the reconstruction accuracy. A library of optical filtering functions was acquired using the time-domain finite-difference (FDTD) method. A cross-correlation algorithm was then used to choose 100 sparse filter functions, which were then built as an encoding matrix and then, based on the encoding matrix, a self-attention mechanism algorithm to improve the accuracy. The reconstructed spectrum's mean square error (MSE) is 0.0019, and its similarity coefficient (R^2) is 0.9780. This self-attention mechanism spectral reconstruction technique will open up new possibilities for high-accuracy reconstruction for various computational spectrometer types.

Keywords: spectral reconstruction, self-attention, encoding matrix, cross-correlation.

1. Introduction

The spectrometer is widely acknowledged as a vital tool in industry and scientific research. Miniaturized spectrometers are critical for developing applications, including consumer electronics, hyperspectral imaging, and *in situ* sensing [1]. Depending on the intricacy of the necessary algorithms, there are two primary types of spectrometer miniaturization techniques: traditional and computational [2-7]. A new paradigm in spectrometer miniaturization, the computational micro spectrometer (CS), is based on computational spectroscopy. It is anticipated to solve the limitations of conventional microspectroscopy techniques, including their inability to achieve extreme miniaturization, lack of stability, and limited resolution [8].

The CS relies on computational techniques to approximate or "reconstruct" an incident light spectrum from precalibrated information encoded within a set of detectors [9]. The process of reconstruction is crucial. Spectral reconstruction techniques have

advanced significantly in recent years due to the growth of CS [10, 11]. In current computational reconstruction algorithms, Chang use the Gaussian kernel template denoising method and parameter minimization method using the l1 paradigm for reconstruction [12]. ZHANG presented a reconstruction approach based on dictionary learning and sparse optimization. The experimental results imply that dictionary learning can significantly increase the sparsity of general spectra, as l1-paradigm minimization performs well for both direct sparse and general spectra that need to be translated into dictionaries [13]. YANG constructed the most miniature microcomputing spectrometer in the world at the time by combining recovered spectral data fitted with a Gaussian basis function [14]. HUANG combined a sparse encoding method with a compressed sensing algorithm to ensure the accuracy of spectral reconstruction, but the spectral resolution was not high enough [15]. HUANG increased the spectral resolution by four times by employing this property of the codec model of the spectral features fed into the network to successfully reconstruct the spectrum information of 100 bands from 25 feature points [16].

Recent research has demonstrated that employing deep learning algorithms for spectrum reconstruction of computational spectroscopy systems produces good results. However, because only the linear transformation of the ultimately linked layer is used and the correlation between the spectra and the spectra itself is ignored, the spectral reconstruction accuracy is not very great [17, 18]. However, the study only considered the linear transformation between spectral information and ignored the correlation between spectra and spectra, which led to the accuracy of spectral reconstruction not being high enough. The self-attention mechanism algorithm in deep learning can handle sequential data well, especially with global dependency. Transformer is a neural network architecture based on an attention mechanism initially used for natural language processing tasks. Its main innovation is that it completely abandons the sequential nature of the sequence and instead establishes dependencies between positions in the input sequence using the self-attention mechanism. At the heart of the Transformer is the self-attention mechanism, which allows the model to treat all positions in the input sequence as objects of attention when computing each output. This allows the model to simultaneously attend to all other positions in the input sequence at each position, thus better capturing long-distance dependencies. It also allows the network to assess the importance of different features autonomously, therefore determining the extent of other features and assigning greater weight to essential features in the reconstruction process [19]. Thus, the properties based on the self-attention mechanism can be used in spectral reconstruction to improve the reconstruction accuracy.

In this work, we first simulate a library of filter functions using FDTD and perform initial screening. Then, we compute and design a coding matrix for the initial screened library of filter functions using a mutual correlation algorithm. Secondly, the designed coding matrix is used to construct a dataset for training and validating the model of the spectral reconstruction algorithm. Finally, the self-attention mechanism is used to reconstruct the spectra, which successfully improves the spectral reconstruction accuracy with a mean square error (MSE) of only 0.0019 and a coefficient of determination (R^2)

as high as 0.9780, which will be a solid step towards the practicalization of computational micro-spectrometers.

2. Method

2.1. Principles of computational spectroscopy

The principle of a computational spectrometer is shown in Fig. 1. When aligned to a measurement target, its incident spectrum first passes through a designed coding matrix. It is then converted into an uncalibrated data. The incident spectrum conversion equation is as follows:

$$\int L(\lambda)T(\lambda, X, Y)\mathrm{d}\lambda = I(X, Y) \tag{1}$$

where $L(\lambda)$ denotes the incident spectrum; $T(\lambda, X, Y)$ is the designed coding matrix; X, Y denotes the filter function at the corresponding position in the coding matrix, and $I(X, Y)$ is the uncalibrated data converted by the filter function at the corresponding position.

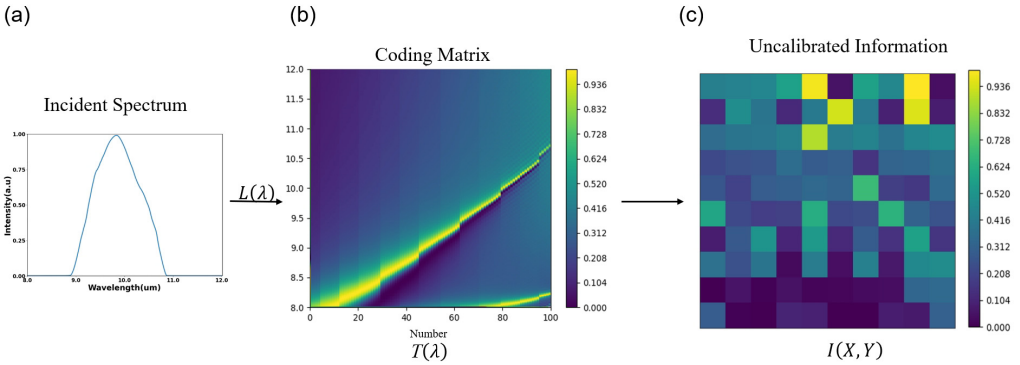

Fig. 1. Spectral conversion to undecoded process. (a) Spectral curves; (b) the constructed coding matrix; (c) the converted uncalibrated data.

2.2. Coding matrix design

In this section, the nanostructured cells are first simulated using the finite difference in time domain (FDTD) method to obtain a filtering function library. The material of the nanopillar is silicon (Si), and the substrate material is also Si. The height of the nanostructures is set to 1 μm, and the period is set to 6 μm. The simulation wavelength range is 8–12 μm, and one point is simulated at every interval of 0.02 μm, with 201 points. By gradually increasing the FDTD nanostructures from 1 to 5 μm, 401 filter curves were obtained as filter curves as a library of filter functions.

Then, the mathematical expectation is used to sieve the filter functions that are insufficient to regulate the light in the FDTD method results to maximize the retention

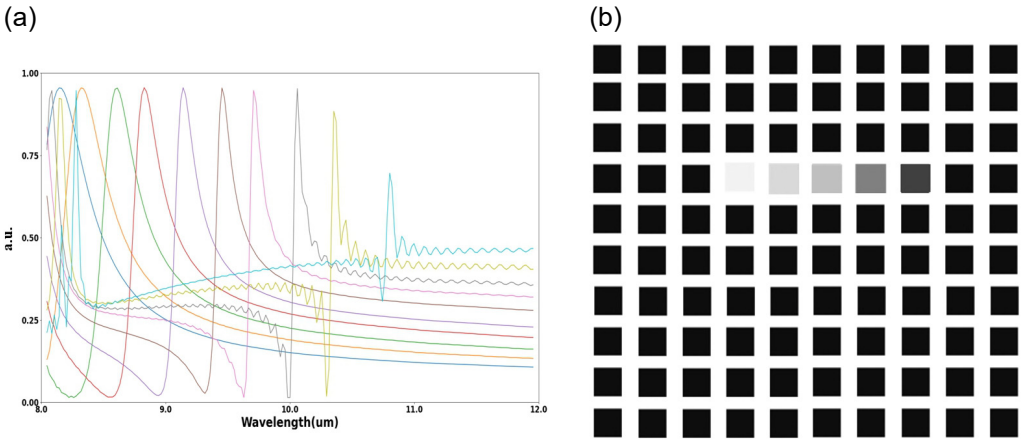

Fig. 2. Filter function arrangement. (a) Filter function; (b) encoding matrix.

of the filter functions that can respond to the light changes. The correlation between each filter function and other filter functions is calculated using the mutual correlation algorithm for the filter function library after the initial screening. Then, the new filter function library is sorted according to the strength of the correlation. The specific sorting method arranges the filter functions in Fig. 2(a) according to the strength of the correlation from left to right and from top to bottom in Fig. 2(b). The correlation formula is as follows:

$$\rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{D(X)D(Y)}} = \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)}\sqrt{E(Y^2) - E^2(Y)}} \tag{2}$$

Where X, Y respectively denote different curves; the value of Cov is the covariance of X, Y curves; its calculation result ρ is a scalar and the range $\rho \in [-1, 1]$; when $\rho > 0$, it means positive correlation, and the closer to 1 means the stronger correlation; $\rho < 0$, it means negative correlation and the closer to –1 means the stronger correlation; when the correlation coefficient is closer to 0, the correlation is weaker, and equals to 0 when two variables are not correlation.

2.3. Data set construction

The spectral curves were first modeled using Gaussian basis functions to generate a series of different spectral curves, with multiple peaked spectral curves generated by superimposing multiple Gaussian functions or adding Gaussian noise. In the end, a total of 500,000 spectral curves were obtained.

After the spectral curves were generated, they were converted to uncalibrated data as input using the designed coding matrix, totaling 500,000 sets of datasets for spectral reconstruction model training and validation. They were randomly divided into training and validation sets in the ratio of 8:2. In addition, 1,000 sets of actual data were

collected as an additional validation set to verify the performance of the algorithmic model in a realistic environment.

2.4. Algorithmic modeling of self-attention mechanisms

This section proposes a method based on the self-attention mechanism network model for spectral reconstruction to learn the mapping relationship from undecoded data to spectral curves. The model of the spectral reconstruction algorithm based on the self -attention mechanism is shown in Fig. 3.

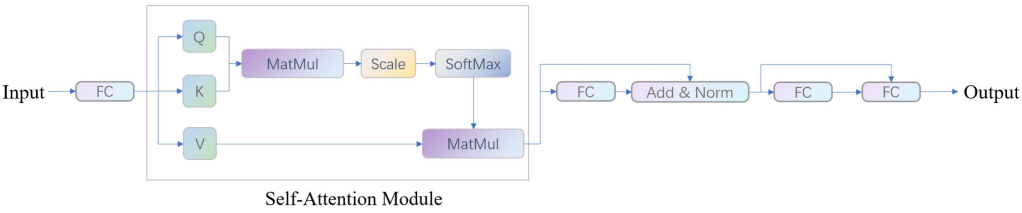

Fig. 3. Model of spectral reconstruction algorithm based on self-attention mechanism.

The model consists of three main parts: a feature extraction part, a nonlinear mapping part, and a reconstruction part.

1. Feature extraction part. Let $X \in R^{10\times 10}$ denote the input data in the training set, where (10, 10) is the format of this input data, and $Y \in R^{1\times 201}$ corresponding to X denotes that the corresponding number of spectral bands is 201, and a dimensionality reduction process needs to be done on X before entering the model to convert X into the format of $X \in R^{1\times 100}$.

2. Nonlinear mapping section. Nonlinear mapping aims to obtain information on the spectral features of undecoded features. First, the undecoded-spectral features of different dimensions are extracted using Q and K, and a new matrix is obtained by doing matrix operations on them to compute a weight for the features.

A normalization operation is then used to calculate the weights of Q and K between 0 and 1. Then, SoftMax converts a set of actual values into a probability distribution.

Finally, a matrix is computed using the data in dimension V with the data above, and residual joins are added to facilitate convergence. In this model, the undecoded data undergoes a self-attention block to get the spectral information features.

3. Reconstruction section. The reconstruction part of the spectrum focuses on reconstructing 1×100 information into 1×201 spectral information. A fully connected layer approach was used where each fully connected layer had incremental neurons added to it to reach the last layer of 201 neurons and finally, the reconstructed spectrum was successfully obtained.

2.5. Training and validation

The Adam optimization algorithm is used to optimize the weights of the parameters. The learning rate is initially 0.001, and as the training proceeds, the network converges

using a learning rate decay strategy. Every 100 epochs, the learning rate will be half the original, the batch size is 100, the maximal training period is 2,000 rounds, and the loss function adopts the mean squared error (MSE).

2.6. Spectral reconstruction performance indicators

Two related indicators were used as reconstruction indicators. The R^2 similarity function is used to evaluate the following metrics:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - Y_i)^2}{\sum_{i=1}^{n} (Y_i - \overline{Y}_i)^2} \tag{3}$$

Where y_i is the reconstructed spectrum's intensity value, Y_i is the simulated spectrum's intensity value, and $\overline{Y}$ is the simulated spectrum's average intensity value; as well as MSE:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - Y_i)^2 \tag{4}$$

Where y_i is the intensity value of the reconstructed spectrum and Y_i is the intensity value of the simulated spectrum.

3. Result

3.1. Coding matrix results

This section demonstrates the filter function library obtained from the screening. The steps are shown in Fig. 4. Figure 4(a) shows the filter function library simulated by

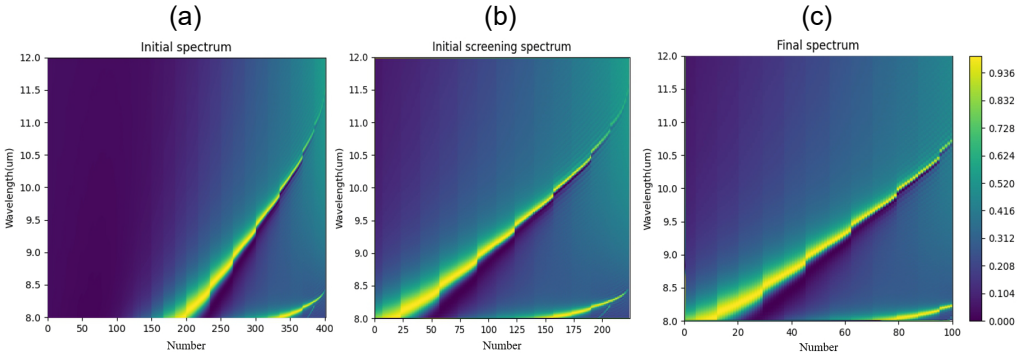

Fig. 4. Different filter function libraries. (a) Filter function library consisting of 401 filter functions obtained by the FDTD method; (b) filter function library consisting of 224 filter functions after an initial screening with an expectation threshold of 0.15; (c) filter function library consisting of 100 filter functions filtered by cross-correlation.

FDTD, with a total of 401 filter functions, from which it can be seen that the filter function on the left side has a change in the optical response close to 0 and no apparent fluctuations. Therefore, the expectation threshold is set to 0.15 to sieve out the function filter with no change in the light response, thus obtaining the filter function library shown in Fig. 4(b), with 224 filter functions and different responses to the spectral changes.

Finally, the cross-correlation algorithm is utilized further to filter the filter function library after the initial filtering, and 100 filter functions are obtained, as shown in Fig. 4(c). In Fig. 4(c), the variation of optical response grows stepwise with the waveband, making the function library more sparse than the library after the initial screening, which facilitates the expression of features in the subsequent spectral reconstruction and thus improves the reconstruction accuracy.

3.2. Constructed datasets

In this section, 100,000 spectral curves were first simulated using Gaussian basis functions. And since there are various Gaussian white noises in the real environment, the noise was added to each simulated spectral curve by 5%, 10%, 15%, and 20%. Thus, a total of 500,000 spectral curves were simulated, as shown in Fig. 5, which will be used as the output of the algorithm model. Figure 5(a) shows the original spectral curve, and Fig. 5(b)-(e) shows the spectral curves obtained by adding different Gaussian noises on top of this curve. The figure shows that the larger the percentage of noise, the more interfering information the curve has, and thus, the more interference it produces in the subsequent reconstruction results.

Below is the uncalibrated data obtained from the spectral curves transformed by the designed coding matrix, which will be used as inputs to the algorithmic model, and its corresponding spectral curves will be used as outputs of the model for training. The uncalibrated data in Fig. 6(a)-(e) correspond to the spectral curves in Fig. 5(a)-(e), respectively.

3.3. Algorithmic reconstruction results for self-attention mechanisms

To verify the magnitude of the improvement in spectral reconstruction accuracy of the model used in this paper relative to other algorithms, the least square, compressed sensing, deep neural network (DNN), and sequence-to-sequence (Seq2Seq) algorithms are chosen as comparison algorithms. A comparison of the results of these four comparison algorithms and the algorithms used in this paper is shown in the Table. As can be seen from the Table, the reconstruction accuracy of the proposed method in this study is significantly improved compared to the least squares and compression-aware algorithms, with the R^2 improved by 0.108 and 0.106. The MSE was reduced by 0.1181 and 0.0981, respectively, and compared to the fully connected network and the coding-decoding network, the R^2 improved by 0.068 and 0.038. The MSE reduced by 0.00681 and 0.00381, respectively.

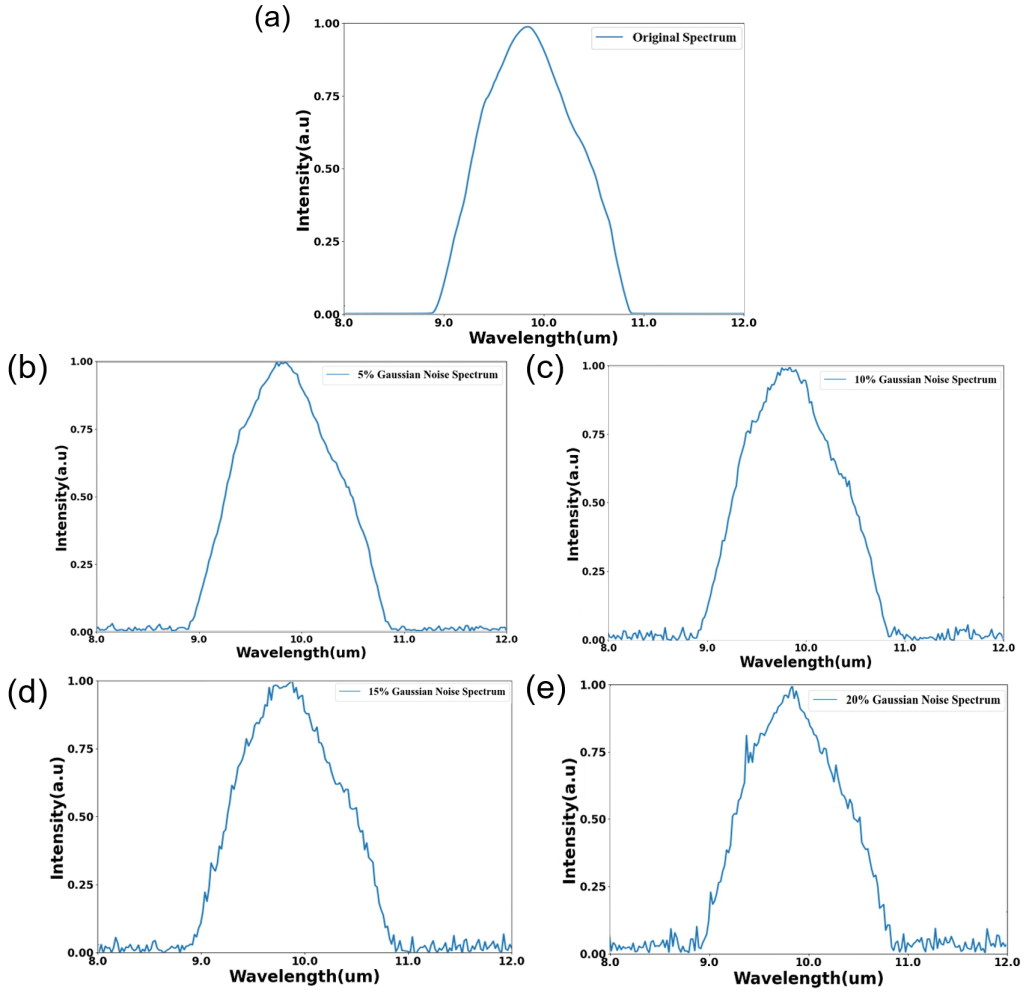

Fig. 5. Simulated spectral curves and spectral curves with different noises. (a) Original spectra; (b) adding 5% noise coefficient spectra; (c) adding 10% noise coefficient spectra; (d) adding 15% noise coefficient spectra; (e) adding 20% noise coefficient spectra.

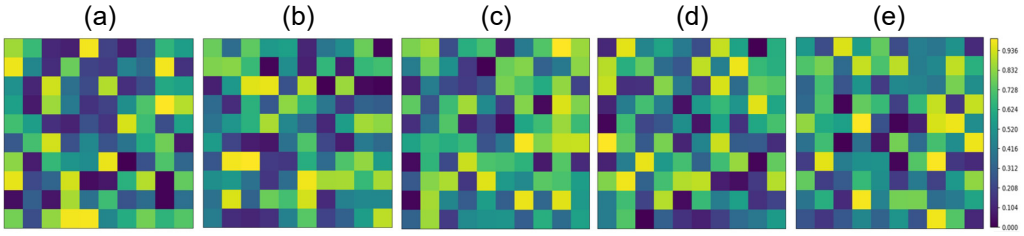

Fig. 6. Simulated uncalibrated data. (a) Uncalibrated data for the original spectra; (b) uncalibrated data with 5% noise factor added; (c) uncalibrated data with 10% noise factor added; (d) uncalibrated data with 15% noise factor added; (e) uncalibrated data with 20% noise factor added.

T a b l e. Comparison of reconstruction results of different algorithms.

Different algorithms	R^2	MSE
Least square	0.87	0.12
Compressed sensing	0.89	0.10
DNN	0.91	0.07
Seq2Seq	0.94	0.04
Self-attention	0.9780	0.0019

Noise needs to be analyzed since it can cause the results to deviate from the real situation and make the reconstruction accuracy much less accurate. The reconstruction results for different noises based on the model of the self-attention algorithm are shown in Fig. 7(a)-(d), and Fig. 7 shows the results for noise parameters of 10% and 20%. It can be seen that the algorithm achieves good denoising results.

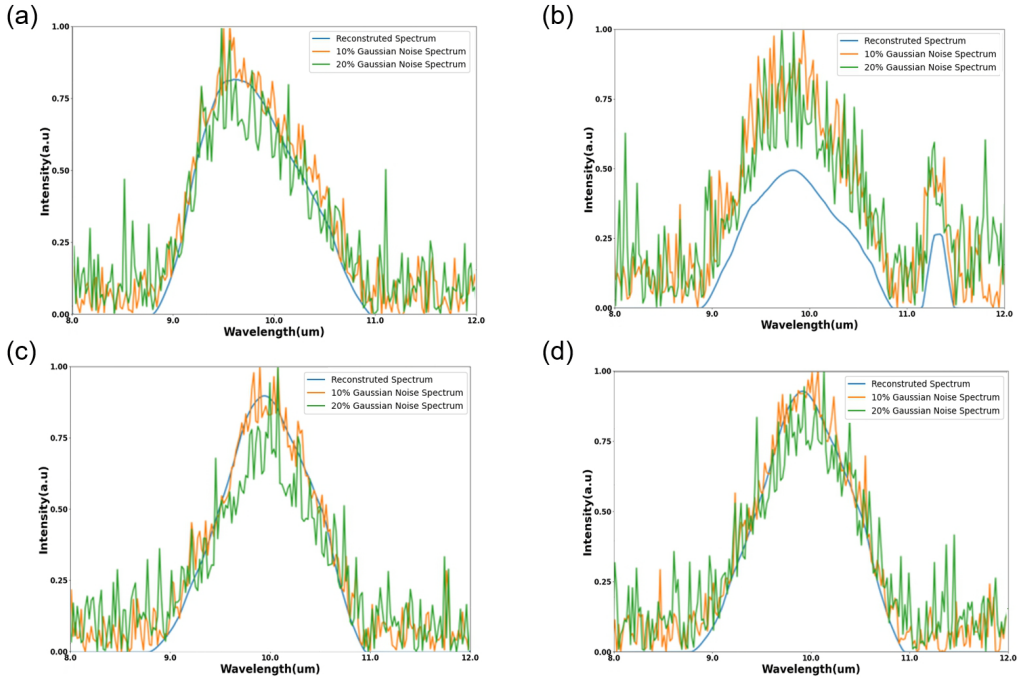

Fig. 7. Reconstruction results with 10% and 20% Gaussian noise added. (a)-(d) Reconstruction results for different curve noises, respectively.

The results of the model used in this thesis for the reconstruction of the actual collected spectra are shown in Fig. 8(a)-(f), and the overall average result of its measurement of 1000 data is $R^2 = 0.9027$, MSE = 0.0085. The results show that using the mutual correlation algorithm to filter the filtering function and the citation of the self-attention mechanism network also has a yield.

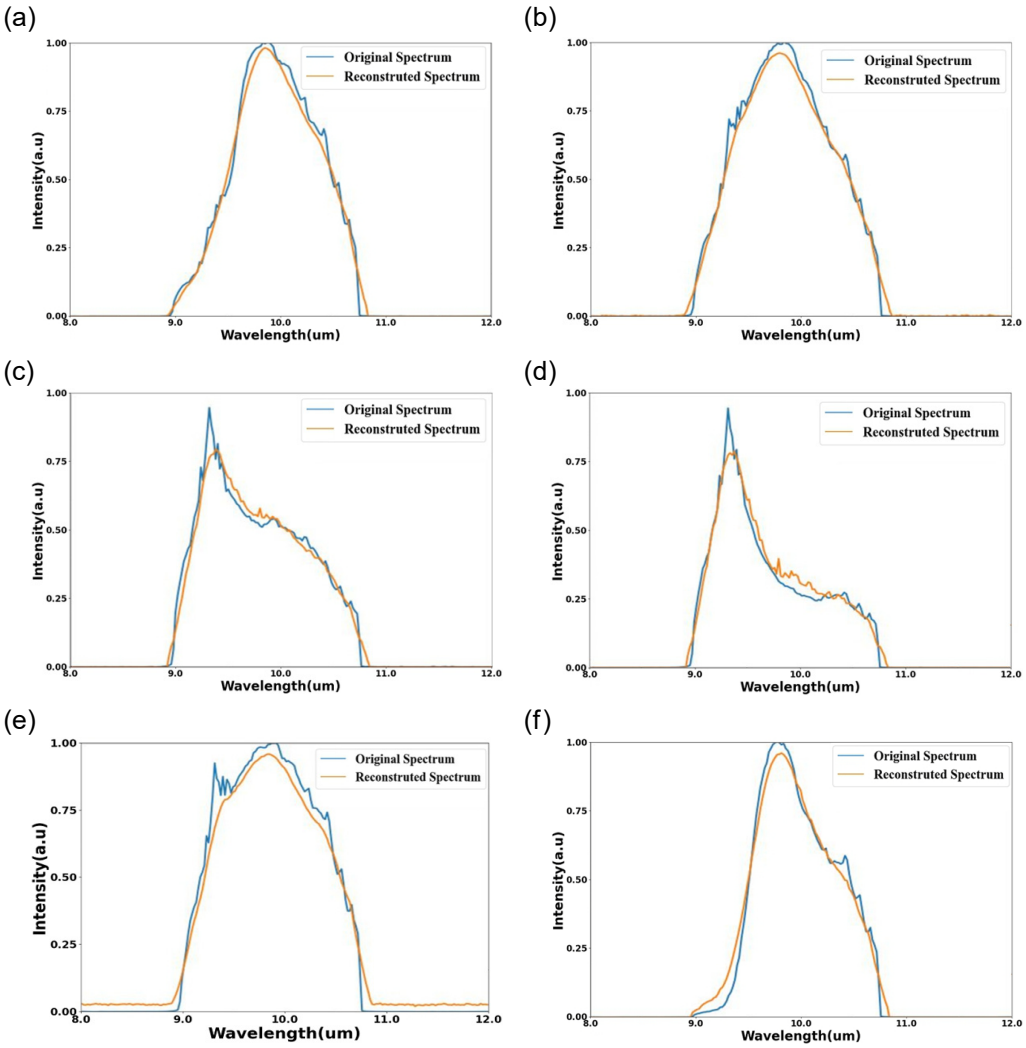

Fig. 8. Actual spectral curve reconstruction results. (a)-(f) Different actual spectral reconstruction results.

4. Conclusions

In this paper, we design a set of spectral coding matrices based on the FDTD method, expectation algorithm, and mutual correlation algorithm, which improves the precision and accuracy of spectral reconstruction of the micro-spectrometer by introducing the algorithm of self-attention mechanism in deep learning. Firstly, the filter function and the mutual correlation function simulated by the FDTD method were used to construct the coding matrix. Then, the dataset was constructed using this coding matrix and the Gaussian function. Finally, the spectral reconstruction was carried out based on the self -attention mechanism and compared with different algorithms. As a result, the method

has dramatically improved the accuracy of spectral reconstruction with a coefficient of determination R^2 of 0.9780 and a root mean square error MSE of 0.0019, which brings great significance to the improvement of the accuracy of spectral reconstruction and makes an essential contribution to the practical application of computational mini-spectrometers.

References

[1] Yang Z., Albrow-Owen T., Cui H., Alexander-Webber J., Gu F., Wang X., Wu T.-C., Zhuge M., Williams C., Wang P., Zayats A.V., Cai W., Dai L., Hofmann S., Overend M., Tong L., Yang Q., Sun Z., Hasan T., *Single-nanowire spectrometers*, Science **365**(6457), 2019: 1017-1020. https://doi.org/10.1126/science.aax8814

[2] Li W., Liu Y., Ling L., Sheng Z., Cheng S., Yi Z., Wu P., Zeng Q., Tang B., Ahmad S., *The tunable absorber films of grating structure of AlCuFe quasicrystal with high Q and refractive index sensitivity*, Surfaces and Interfaces **48**, 2024: 104248. https://doi.org/10.1016/j.surfin.2024.104248

[3] Li W., Liu M., Cheng S., Zhang H., Yang W., Yi Z., Zeng Q., Tang B., Ahmad S., Sun T., *Polarization independent tunable bandwidth absorber based on single-layer graphene*, Diamond and Related Materials **142**, 2024: 110793. https://doi.org/10.1016/j.diamond.2024.110793

[4] Liang S., Xu F., Li W., Yang W., Cheng S., Yang H., Chen J., Yi Z., Jiang P., *Tunable smart mid infrared thermal control emitter based on phase change material VO_2 thin film*, Applied Thermal Engineering **232**, 2023: 121074. https://doi.org/10.1016/j.applthermaleng.2023.121074

[5] Ma J., Wu P., Li W., Liang S., Shangguan Q., Cheng S., Tian Y., Fu J., Zhang L., *A five-peaks graphene absorber with multiple adjustable and high sensitivity in the far infrared band*, Diamond and Related Materials **136**, 2023: 109960. https://doi.org/10.1016/j.diamond.2023.109960

[6] Shangguan Q., Zhao Y., Song Z., Wang J., Yang H., Chen J., Liu C., Cheng S., Yang W., Yi Z., *High sensitivity active adjustable graphene absorber for refractive index sensing applications*, Diamond and Related Materials **128**, 2022: 109273. https://doi.org/10.1016/j.diamond.2022.109273

[7] Schuler L.P., Milne J.S., Dell J.M., Faraone L., *MEMS-based microspectrometer technologies for NIR and MIR wavelengths*, Journal of Physics D: Applied Physics **42**(13), 2009: 133001. https://doi.org/10.1088/0022-3727/42/13/133001

[8] Malinen J., Rissanen A., Saari H., Karioja P., Karppinen M., Aalto T., Tukkiniemi K., *Advances in miniature spectrometer and sensor development*, Proceedings of the SPIE, Vol. 9101, Next-Generation Spectroscopic Technologies VII, 2014: 91010C. https://doi.org/10.1117/12.2053567

[9] Ebermann M., Neumann N., Hiller K., Seifert M., Meinig M., Kurth S., *Tunable MEMS Fabry -Pérot filters for infrared microspectrometers: A review*, Proceedings of the SPIE, Vol. 9760, MOEMS and Miniaturized Systems XV, 2016: 97600H. https://doi.org/10.1117/12.2209288

[10] Crocombe R.A., *Portable spectroscopy*, Applied Spectroscopy **72**(12), 2018: 1701-1751. https://doi.org/10.1177/0003702818809719

[11] Wolffenbuttel R.F., *MEMS-based optical mini- and microspectrometers for the visible and infrared spectral range*, Journal of Micromechanics and Microengineering **15**, 2005: S145-S152. https://doi.org/10.1088/0960-1317/15/7/021

[12] Kurokawa U., Choi B.I., Chang C.C., *Filter-based miniature spectrometers: Spectrum reconstruction using adaptive regularization*, IEEE Sensors Journal **11**(7), 2011: 1556-1563. https://doi.org/10.1109/jsen.2010.2103054

[13] Zhang S., Dong Y., Fu H., Huang S.-L., Zhang L., *A spectral reconstruction algorithm of miniature spectrometer based on sparse optimization and dictionary learning*, Sensors **18**(2), 2018: 644. https://doi.org/10.3390/s18020644

[14] Yang Z., Albrow-Owen T., Cai W., Hasan T., *Miniaturization of optical spectrometers*, Science **371**(6528), 2021: eabe0722. https://doi.org/10.1126/science.abe0722

[15] XIONG J., CAI X., CUI K., HUANG Y., YANG J., ZHU H., ZHENG Z., XU S., HE Y., LIU F., FENG X., ZHANG W., *One-shot ultraspectral imaging with reconfigurable metasurfaces*, Optica Open, Preprint, 2020. https://doi.org/10.48550/arXiv.2005.02689

[16] YANG J., CUI K., CAI X., XIONG J., ZHU H., RAO S., XU S., HUANG Y., LIU F., FENG X., ZHANG W., *Ultraspectral imaging based on metasurfaces with freeform shaped meta-atoms*, Laser & Photonics Reviews **16**(7), 2022: 2100663. https://doi.org/10.1002/lpor.202100663

[17] SONG H., MA Y., HAN Y., SHEN W., ZHANG W., LI Y., LIU X., PENG Y., HAO X., *Deep-learned broadband encoding stochastic fifilters for computational spectroscopic instruments*, Advanced Theory and Simulations **4**(3), 2021: 2000299. https://doi.org/10.1002/adts.202000299

[18] NIE S., GU L., ZHENG Y., LAM A., ONO N., SATO I., *Deeply learned filter response functions for hyperspectral reconstruction*, [In] *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018: 4767-4776. https://doi.org/10.1109/CVPR.2018.00501

[19] FRINTROP S., *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, Springer Berlin Heidelberg, Vol. 3899, 2006. https://link.springer.com/book/10.1007/11682110

Received March 24, 2024
in revised form May 27, 2024