第 42 卷 第 24 期/2022 年 12 月/光学学报 研究论文

光学学报

CNN-SVR用于MAX-DOAS预测对流层NO_2廓线

潘屹峰[1]，田鑫[1,2]，谢品华[2,3,4*]，李昂[2]，徐晋[2]，任博[2,4]，黄骁辉[1]，田伟[1]，王子杰[1]

[1]安徽大学物质科学与信息技术研究院安徽省信息材料与智能传感实验室，安徽 合肥 230601；

[2]中国科学院安徽光学精密机械研究所环境光学与技术重点实验室，安徽 合肥 230031；

[3]中国科学院区域大气环境研究卓越创新中心，福建 厦门 361021；

[4]中国科学技术大学环境科学与光电技术学院，安徽 合肥 230026

摘要 提出一种基于卷积神经网络(CNN)和支持向量回归机(SVR)的多轴差分光学吸收光谱(MAX-DOAS)对流层NO_2垂直分布预测方法。将2019年南京站点采集的原始MAX-DOAS数据通过QDOAS软件拟合获取O_4和NO_2差分斜柱浓度，结合基于最优估算的气溶胶和痕量气体廓线反演算法——PriAM算法反演了对流层NO_2廓线，并将其作为预测模型的输出。此外，通过平均影响值方法进行预测模型输入变量的选择，确定了MAX-DOAS数据、温度、气溶胶光学厚度和低云覆盖率为模型的最佳输入变量。通过实验优化网络结构和参数，最终建立预测模型在测试集与PriAM的平均百分比误差仅为9.14%，与单独建立的CNN、SVR、反向传播模型相比，平均百分比误差分别降低了8.22%、6.00%、32.28%。因此，CNN-SVR能够利用MAX-DOAS数据对对流层NO_2廓线进行有效预测。

关键词 大气光学；卷积神经网络；支持向量回归机；多轴差分吸收光谱；对流层NO_2廓线

中图分类号 TP301 **文献标志码** A **DOI**：10.3788/AOS202242.2401001

Prediction of Tropospheric NO_2 Profile Using CNN-SVR-Based MAX-DOAS

Pan Yifeng[1], Tian Xin[1,2], Xie Pinhua[2,3,4*], Li Ang[2], Xu Jin[2], Ren Bo[2,4], Huang Xiaohui[1], Tian Wei[1], Wang Zijie[1]

[1]*Information Materials and Intelligent Sensing Laboratory of Anhui Province, Institutes of Physical Science and Information Technology, Anhui University, Hefei* 230601, *Anhui, China*;

[2]*Key Laboratory of Environmental Optics and Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei* 230031, *Anhui, China*;

[3]*CAS Center for Excellence in Regional Atmospheric Environment, Xiamen* 361021, *Fujian, China*;

[4]*School of Environmental Science and Optoelectronic Technology, University of Science and Technology of China, Hefei* 230026, *Anhui, China*

Abstract This study proposes a method based on a convolutional neural network (CNN) and support vector regression machine (SVR) for predicting the vertical distribution of NO_2 in the troposphere by multi-axis differential optical absorption spectroscopy (MAX-DOAS) technology. Taking the Nanjing site as an example, we obtain the O_4 and NO_2 differential slant column density (dSCD) according to the raw MAX-DOAS data collected by QDOAS fitting in 2019, invert the tropospheric NO_2 profile by combining the optimal estimation-based aerosol and trace gas profile inversion algorithm—PriAM, and use the profile as the output of the prediction model. In addition, the input variables of the prediction model are selected by the mean impact value method, with MAX-DOAS data, temperature, aerosol optical thickness, and low cloud coverage finally identified as the optimal input variables for the model. Furthermore, the network structure and parameters are optimized through experiments, and the average percentage error of the final CNN-SVR prediction model in the test set with PriAM is only 9.14%, which is 8.22%, 6.00%, and 32.28% lower than that of the separately constructed CNN, SVR, and backpropagation models, respectively. Therefore, CNN-SVR can effectively predict

收稿日期：2022-03-10；**修回日期**：2022-04-17；**录用日期**：2022-05-05

基金项目：国家自然科学基金(42105132,U19A2044, 42030609)、安徽省自然科学基金(2008085QD183,2008085QD182)

通信作者：*phxie@aiofm.ac.cn

tropospheric NO_2 profiles by using MAX-DOAS data.

Key words atmospheric optics; convolutional neural network; support vector regression machine; multi-axis differential absorption spectroscopy; tropospheric NO_2 profile

1 引 言

进入21世纪以来，随着人类社会经济的迅速发展，工厂以及交通排放物所占比例加大，导致以NO_2为主的氮氧化物(NO_x)在大气中的浓度显著提升，使得我国面临的光化学污染和灰霾污染问题日益严峻。因此，及时掌握空气质量状况，采取更为有效的科学手段控制空气污染，对高效获取大气中NO_2的空间信息提出了更为严格的要求[1]。

多轴差分光学吸收光谱(MAX-DOAS)技术是一种广泛使用的遥测技术，可以同时观测多种大气污染成分，是一种便捷、有效地获取大气污染物垂直分布的方法[2]，已在对流层NO_2、SO_2、HCHO等痕量气体和气溶胶廓线反演方面得到了广泛的应用[3-6]。现阶段利用MAX-DOAS技术获取痕量气体垂直分布主要基于两类反演算法，即查表法和最优估算法[7-9]，这些方法都存在一定的局限性：最优估算法依赖于先验廓线，反演过程中可能出现奇异值；查表法是一种参数化反演方法，该算法各层的分配比例和线性函数等参数定义比较粗略，空间分辨率低[10]。同时，这两种廓线反演方法需采用大气辐射传输模型(RTM)来模拟光的传输路径，这个过程是极其复杂的，而机器学习模型可以很好地弥补廓线反演方法的这个缺点。机器学习方法通过对原始数据的分析处理和各种预测算法的应用来建立输入数据和输出数据的映射关系，直接从MAX-DOAS预测对流层NO_2的垂直廓线，从而减少辐射传输模型模拟的流程，实现快速且准确的垂直廓线预测。

卷积神经网络(CNN)和支持向量回归机(SVR)是新颖的机器学习模型，已在光学领域得到广泛的应用。目前已有研究将CNN应用到近红外光谱领域进行预测和分类，均取得极佳的效果[11-12]，这也为本研究将CNN应用到MAX-DOAS数据检测中提供了理论支持。与传统的神经网络算法相比，CNN具有更强的建模能力，可以从复杂的MAX-DOAS数据中提取有效的特征数据，并学习特征数据的内部结构以更好地进行预测。Malmgren-Hansen等[13]首次介绍了使用CNN从IASI探测数据中反演大气廓线，用深度卷积网络完成对多维轮廓的预测。支持向量机得益于核函数和少数起到决定作用的支持向量，适合解决各种回归预测问题。本文基于长期MAX-DOAS观测数据，提出一种CNN-SVR混合模型，并利用该模型对对流层NO_2廓线进行预测，利用CNN和SVR算法建立将MAX-DOAS数据和气象数据等作为输入数据、对流层NO_2廓线作为输出数据的模型，实现将光谱数据和气象数据等输入模型，就能得到一条准确的对流层廓线的目的。

2 实验数据与基本原理

2.1 样本数据采集

用于建模的数据来自2019年中国科学院安徽光学精密机械研究所(AIOFM)在南京站点(32°7′12″N，118°57′36″E)测得的MAX-DOAS多仰角光谱数据。仪器方位角为310°(规定正北方向为0°)，测量仰角时以11个角度为一组，分别为1°、2°、3°、4°、5°、6°、8°、10°、15°、30°、90°。完整仰角序列的光谱保留用作本实验的输入数据，若有仰角缺失，则被判定为“有缺失”数据组，并剔除该组数据。

南京站点MAX-DOAS中使用的光谱仪是Maya2000 Pro。该光谱仪的波长范围为290～420 nm，对应2068条通道，光谱分辨率为0.4～0.7 nm。将光谱仪置于恒温箱中，并将温度控制在20 ℃[14]。本研究使用最小二乘法处理所观测的光谱[4]，利用QDOAS(V2.111)软件进行光谱反演，NO_2反演的波长范围为330～375 nm，对应的光谱数据通道数量为613～1297条。本实验的输入数据集即为每组测量仰角在该通道数范围内的数据。

气象数据(温度、湿度、风速、风向)对NO_2的浓度有着重要影响[15]，气溶胶光学厚度(AOD)、低云覆盖率也会影响基于最优估算的气溶胶和痕量气体廓线反演算法——PriAM算法反演对流层NO_2廓线的结果[16-18]。因此，在建立预测模型时，将这些参数也作为输入数据进行敏感性分析，求解最优敏感因子。AOD、低云覆盖率的数据来自欧洲中期天气预报中心哥白尼大气监测服务(CAMS)网站(https://ads.atmosphere.copernicus.eu)。目前已有大量研究证明CAMS预测的AOD的准确性，且CAMS相对于美国宇航局(NASA)开发的MERRA-2预测的中国华东地区的AOD准确性有相当大的提升[19]。风向、风速、温度、相对湿度的数据来自全球天气精准预报网(https://www.wunderground.com/history/daily/cn/nanjing/ZSNJ)。

输出数据集是采用MAX-DOAS廓线反演算法PriAM反演的NO_2廓线，具体获取流程如下：

1) NO_2和O_4差分斜柱浓度(dSCD)的获取。基于Lambert-Beer定律，利用DOAS算法反演NO_2和O_4的dSCD。将一个测量循环开始的天顶光谱作为夫琅禾费参考光谱，以去除太阳夫琅禾费结构及平流层的影响。利用QDOAS软件反演NO_2廓线的配置和本实验选取的330～375 nm波段范围均是采用先前文献[14,16-17]报道的标准配置。

2)NO_2廓线的反演。PriAM算法是由AIOFM和MPIC(马克斯·普朗克化学研究所)共同研发的基于非线性最优估计的痕量气体和气溶胶垂直廓线两步反

演算法。首先，将同一波段由MAX-DOAS反演的O_4 dSCD输入PriAM，以反演气溶胶廓线。其次，利用PriAM痕量气体廓线反演算法，将气溶胶廓线和痕量气体dSCD输入到该算法中，得到对流层痕量气体垂直廓线[14, 18-20]。

将每组原始光谱数据经过QDOAS拟合和PriAM算法反演之后得到对流层NO_2廓线数据。考虑到天气因素对PriAM算法反演对流层NO_2廓线的影响，剔除雨雪天气反演出的对流层NO_2廓线数据。剔除这部分数据的原因是雨雪天气下PriAM算法反演得到的对流层NO_2廓线数据存在很大的误差，将这些错误的数据输入到模型中，会影响机器学习预测模型的精度。对流层NO_2廓线的高度范围为0～4 km，垂直分辨率为200 m（最低一层为50 m），故NO_2气体的浓度数据一共有21个格点。根据反演得到的廓线数据和对应的通道内原始光谱数据，利用神经网络算法建立预测模型。如Hornik等[21]所述，只要训练数据集足够大，神经网络就可以逼近未知的函数关系，在统计学上表示为反演得到的量。本研究选取2019年南京站点采集的8225组数据作为样本，而选取连续一年的观测数据可以有效地考虑NO_2的季节性变化对模型验证以及误差分析的影响。

2.2 数据预处理

MAX-DOAS数据是连续型数据，而风向、季节是离散型数据，这些参数在作为模型输入数据前，需要对离散型数据进行处理。将风向分为北（N）、东北（EN）、西北（WN）、东（E）、东南（ES）、南（S）、西南（WS）、西（W），季节分为春、夏、秋、冬。使用独热（One-Hot）编码处理离散型变量，以合理地计算特征距离[22]。

为了提升模型精度以及收敛速度，使得本研究设计的模型能快速收敛，避免神经元饱和，对输入和输出NO_2廓线采用数据归一化进行预处理。归一化后的数据缩放到[0，1]之间，x_t为归一化后的数据值，归一化公式为

$$x_t=(x-x_{min})/(x_{max}-x_{min}), \tag{1}$$

式中：x为输入数据；x_{min}为输入数据的最小值；x_{max}为输入数据的最大值。

2.3 输入变量的敏感性分析

将利用CNN提取的MAX-DOAS数据的特征向量、温度、风向、风速、相对湿度、季节、AOD、低云覆盖率作为SVR的输入参数。模型输入参数的好坏决定了模型的精度，本研究为了提高建模效率和实验精度，需要对输入变量进行筛选。选择使用Dombi等[23]提出的平均影响值（MIV）方法，考虑到该方法原理简单、计算方便，且本研究的输入参数没有清晰的变量筛选依据，故使用应用最广泛的MIV方法。该方法流程如图1所示。

按照图1所示的流程，计算不同输入变量的MIV值占全部输入变量MIV之和的百分比，如图2所示。

MIV的绝对值大小决定了输入变量对预测结果

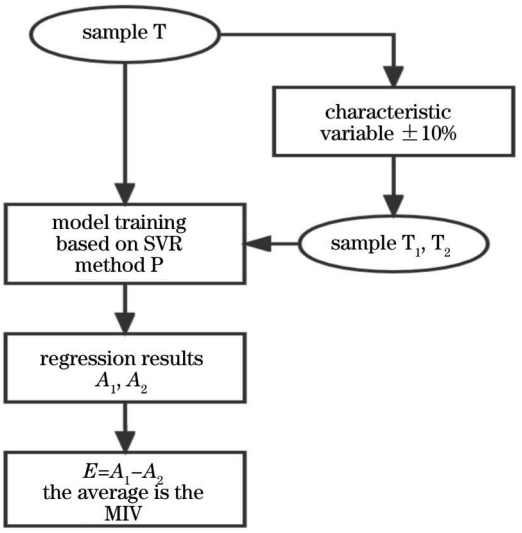

图1 MIV方法流程框图

Fig. 1 Flow chart of MIV method

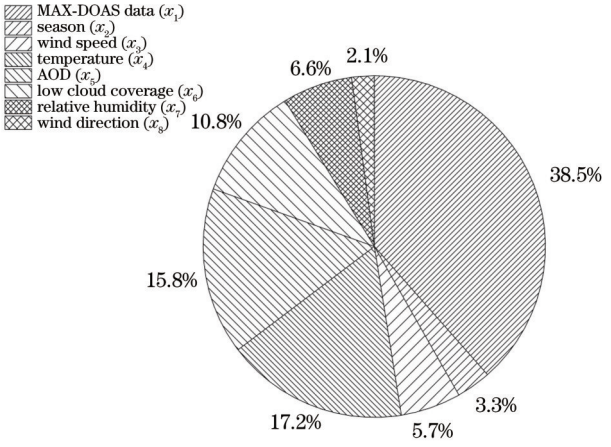

图2 不同输入变量的MIV值占全部输入变量MIV之和的百分比

Fig. 2 Percentage of MIV value of different input variables in the sum of all input variables

的重要性，本研究得到的重要性依次为：MAX-DOAS数据（x_1）、温度（x_4）、AOD（x_5）、低云覆盖率（x_6）、相对湿度（x_7）、风速（x_3）、季节（x_2）、风向（x_8）。根据MIV的绝对值从低到高的顺序依次剔除输入变量，观察不同体系中预测精度的变化，最高预测精度对应的即最优指标体系。以均方根误差（RMSE）为评价指标，其计算公式为

$$E_{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_a-y_p)^2}, \tag{2}$$

式中：y_a为实际值；y_p为预测值。不同输入变量下的RMSE如表1所示。

从表1可以看到：输入变量体系F1～F5的RMSE值逐渐减小，说明在逐步剔除风向、季节、风速、相对湿度等变量的过程中模型的RMSE一直在降低；在体系F5～F8中逐步剔除低云覆盖率、AOD、温度，模型的RMSE却越来越大。当输入变量为体系F5时，模型

表1 不同输入变量体系下的RMSE的值

Table1 RMSE values under different input variable systems

System	Input variable system	RMSE
F1	x_1, x_4, x_5, x_6, x_7, x_3, x_2, x_8	0.512
F2	x_1, x_4, x_5, x_6, x_7, x_3, x_2	0.487
F3	x_1, x_4, x_5, x_6, x_7, x_3	0.452
F4	x_1, x_4, x_5, x_6, x_7	0.415
F5	x_1, x_4, x_5, x_6	0.322
F6	x_1, x_4, x_5	0.401
F7	x_1, x_4	0.498
F8	x_1	0.586

RMSE最小，即输入敏感因子中，MAX-DOAS数据、温度、AOD和低云覆盖率的影响明显高于风速、风向、季节和相对湿度。因此，本研究最终确定MAX-DOAS数据、温度、AOD和低云覆盖率为所提模型的最佳输入变量。

2.4 CNN、SVR和BP神经网络

考虑到MAX-DOAS的结构是一个一维向量，本研究利用CNN良好的自学习功能，通过CNN的卷积层提取MAX-DOAS独特的特征，池化层进行降维的同时保留MAX-DOAS的特征，以防止过拟合，并利用全连接层来处理得到的特征，使其输到输出层。通过CNN 3种不同网络层的依次作用，可以有效地提取MAX-DOAS的特征，避免复杂的光谱预处理。

SVR是目前机器学习回归预测算法中的翘楚。用二元集合$(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)$，$x_n\in\mathbf{R}$，$y_n\in\mathbf{R}$，表示测试集。$\boldsymbol{x}_n$为由AOD、温度、季节、MAX-DOAS数据、低云覆盖率等构成的特征矩阵。输出数据为对应的对流层NO_2廓线y_n。构造和计算超平面的距离是SVR的核心，超平面函数$f(\boldsymbol{x})$和距离L_i的计算公式分别为

$$f(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_n+b, \tag{3}$$

$$L_i-f(\boldsymbol{x})=L_i-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}-b, \tag{4}$$

式中：$\boldsymbol{w}^{\mathrm{T}}$为权重向量；$b$为偏移量；$\boldsymbol{x}$为输入数据的特征向量。

SVR为了构建最优模型、得到最优超平面，引入了松弛因子R_i和R_i'，以减小回归误差，即

$$\min_{\boldsymbol{w},b}\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}+P\sum_{i=1}^{n}(R_i+R_i'), \tag{5}$$

$$\text{s.t.}\begin{cases}L_i-f(\boldsymbol{x})\leqslant\varepsilon+R_i\\ f(\boldsymbol{x})-L_i\leqslant\varepsilon+R_i'\\ R_i\geqslant 0,R_i'\geqslant 0,i=1,2,\cdots,n\end{cases}, \tag{6}$$

式中：P为惩罚系数，且$P>0$；ε为不敏感损失函数。

所提SVR模型引入径向基函数(RBF)的核函数来解决样本数据线性不可分的问题，通过设置惩罚系数P和核函数系数γ，最终得到最优超平面的决策函数，即

$$f(\boldsymbol{x})=\sum_{i=1}^{N_{sv}}L_i\boldsymbol{g}_iK(\boldsymbol{x},x_i)+b, \tag{7}$$

式中：$\boldsymbol{g}_i$为支持向量；$K(\cdot)$为核函数；N_{sv}为支持向量的数量。

反向传播(BP)神经网络是应用最广泛的神经网络模型，该模型具有较强的自学习能力、泛化能力和很强的容错能力。该模型通过误差的BP来逐步降低模型训练中的误差，最终获得最优的BP神经网络模型。目前BP神经网络在预测近地面NO_2的浓度中，基于其强大的学习和泛化能力，模型设计简单(通常使用三层BP神经网络)，可快速预测近地面NO_2的浓度，并取得不错的效果，所建立的模型具有一定的推广、概括能力。尽管BP神经网络目前还未曾应用于对流层NO_2廓线的反演，但考虑到BP神经网络在预测近地面NO_2的浓度取得成功，故本研究建立了BP神经网络模型来对比其他预测模型。

3 CNN-SVR混合模型分析与讨论

3.1 CNN-SVR混合模型

本研究将CNN和SVR相结合，提出一种CNN-SVR混合预测模型。该模型利用了CNN提取数据特征的能力和SVR在解决高维特征回归问题的高效性，将CNN中间层的输出特征作为提取到的最终特征并和经过MIV筛选的温度、AOD、低云覆盖率一起传入SVR模型中预测对流层NO_2廓线。

首先，基于MIV算法的筛选确立了MAX-DOAS数据、温度、AOD、低云覆盖率作为模型的输入变量，输出指标数据为对流层NO_2廓线。将MAX-DOAS数据作为输入变量输入CNN模型中，对原始MAX-DOAS数据进行归一化处理，消除量纲的影响。将MAX-DOAS数据输入到CNN卷积层，通过多次卷积提取有效特征，通过全连接层输出最终特征。以均方误差(MSE)的最小值为多次迭代的目标函数，利用梯度下降算法进行优化，从而寻找最优解，此时将全连接层的输出数据特征以矢量形式存储。将CNN提取的MAX-DOAS数据特征作为SVR的输入，同时将温度、AOD、低云覆盖率数据也作为SVR的输入，从而确定SVR的参数。CNN网络层的模型参数设置如表2所示，所设计的CNN-SVR混合模型结构如图3所示。

为了减少由光谱平移和旋转造成的干扰，池化层采用最大池化方法采样。均方误差用来衡量PriAM反演的对流层NO_2廓线与预测对流层NO_2廓线的欧氏距离，并作为模型的损失函数。优化器使用动量梯度下降(SGD+Momentum)，学习率为10^{-3}，批量大小(Batchsize)根据训练集的样本数量设置为64，以保证内存利用率，加快处理速度[24]。SVR模型采用RBF核函数，为了避免陷入局部最优解，最终确定$P=100$、$\gamma=0.01$分别作为SVR模型的惩罚因子和核参数[25]。

模型及参数确定后在测试集开展对流层NO_2廓线预测。采用平均绝对值误差(MAPE)[式(8)]和对称平均绝对百分比误差(SMAPE)[式(9)]来评估对

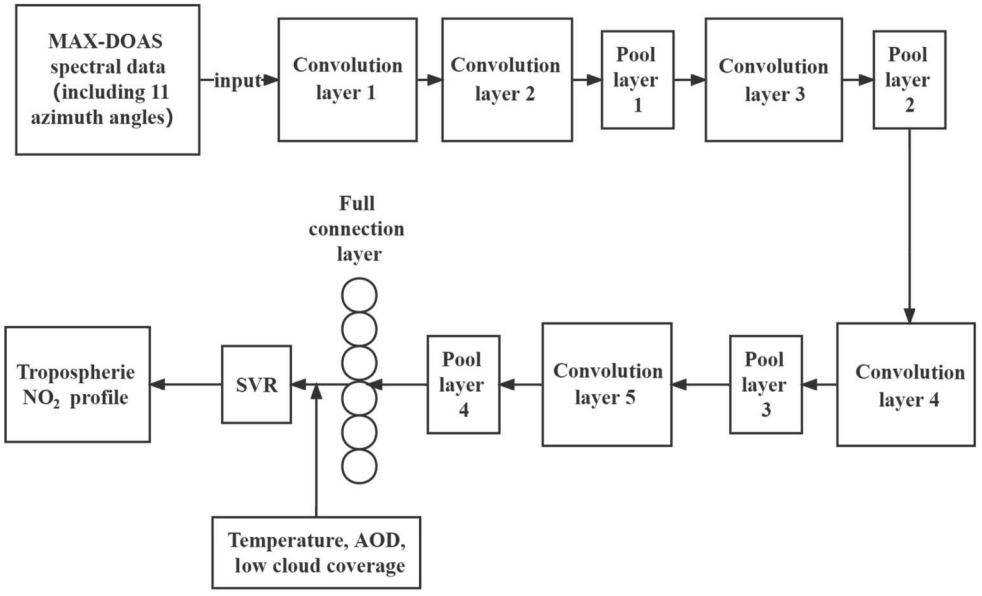

图 3 CNN-SVR混合模型框架

Fig. 3 Frame of CNN-SVR hybrid model

表 2 CNN参数设置

Table 2 Parameter setting of CNN

Network layer	Model parameter setting
Input layer	685×11 spectral data matrix
Convolution layer 1	64 1×1 convolution kernels; kernel_size: 5
Convolution layer 2	128 1×1 convolution kernels; kernel_size: 5
Pool layer 1	MaxPool; kernel_size: 1; stride: 2
Convolution layer 3	128 1×1 convolution kernels; kernel_size: 5
Pool layer 2	MaxPool; kernel_size: 1; stride: 2
Convolution layer 4	256 1×1 convolution kernels; kernel_size: 5
Pool layer 3	MaxPool; kernel_size: 1; stride: 2
Convolution layer 5	512 1×1 convolution kernels; kernel_size: 5
Pool layer 4 (adaptive pooling layer)	Output one-dimensional vector
Full connection layer	Output 21 neurons

流层NO_2廓线预测精度。MAPE最大的优点是以百分比表示，可用于比较不同比例预测模型的准确性，易于解释预测模型的误差，但是MAPE的值可能超过100%且无上限。而SMAPE具有下限(0%)和上限(200%)，使得误差在一定范围内而不存在无穷大值。

$$E_{\mathrm{MAPE}}=\frac{1}{n}\sum_{i=1}^{N}\left|\frac{y_{\mathrm{a}}-y_{\mathrm{p}}}{y_{\mathrm{p}}}\right|\times 100\%, \tag{8}$$

$$E_{\mathrm{SMAPE}}=\frac{1}{n}\sum_{i=1}^{N}\frac{\left|y_{\mathrm{a}}-y_{\mathrm{p}}\right|}{\left(\left|y_{\mathrm{a}}\right|+\left|y_{\mathrm{p}}\right|\right)/2}\times 100\%。 \tag{9}$$

所有的原始数据样本被随机划分到训练集和测试集。为了研究CNN-SVR模型的不同训练集占全样本的比例对CNN-SVR建模效果的影响，采用相同的规则划分训练集和测试集。训练集样本占所有样本的50%、60%、70%、80%、90%。MAPE和SMAPE的测试集经过交叉验证的结果被用来评估预测的准确性。不同训练集样本下CNN-SVR模型测试集的预测结果如表3所示。当训练集的比例为90%时，CNN-SVR在测试集上最佳，其MAPE为9.14%，SMAPE为8.52%。

表 3 不同训练集样本下CNN-SVR模型测试集的预测结果

Table 3 Prediction results of CNN-SVR model test set under different training set samples

Ratio of training set	SMAPE /%	MAPE /%
0.5	21.97	23.41
0.6	18.11	19.33
0.7	15.13	16.24
0.8	11.55	12.41
0.9	8.52	9.14

3.2 不同预测模型对比

为了确立最优的对流层NO_2廓线预测模型，对比了构建的CNN模型、SVR模型、BP神经网络和CNN-SVR混合模型。

从表4所示的4种模型(CNN-SVR、BP、CNN、SVR)在训练数据集和测试数据集上的MAPE和

SMAPE可以看出，CNN-SVR混合模型的预测结果误差最小。与CNN模型、SVR模型、BP模型相比，CNN-SVR混合模型的MAPE分别降低了8.22%、6.00%、32.28%，SMAPE分别降低了7.59%、5.76%、30.69%。因此，使用CNN-SVR混合模型预测对流层NO_2廓线的效果最优，具有更高的预测精度。

表4 4种预测模型下的MAPE和SMAPE

Table 4 MAPE and SMAPE under four prediction models

Model	MAPE /%		SMAPE /%	
	Training error	Test error	Training error	Test error
CNN	12.59	17.63	11.61	16.11
SVR	11.82	15.14	10.92	14.28
BP	35.68	41.42	33.63	39.21
CNN-SVR	7.93	9.14	7.25	8.52

4种模型（CNN-SVR、BP、CNN、SVR）的训练集和测试集预测的NO_2廓线与PriAM反演的真实NO_2廓线之间的决定系数R^2如表5所示。CNN-SVR模型的决定系数较其他3种模型的结果高，CNN-SVR模型在训练集的R^2为0.95，在测试集的R^2为0.93，表明CNN-SVR拟合优度最佳。

表5 4种预测模型下的R^2

Table 5 R^2 under four prediction models

Model	R^2 in modeling set	R^2 in testing set
CNN	0.86	0.83
SVR	0.89	0.87
BP	0.81	0.77
CNN-SVR	0.95	0.93

3.3 对比验证

为了验证建立的CNN-SVR模型的预测能力，在春、夏、秋、冬4个季节中随机各取1个晴天，利用CNN-SVR模型预测这4天的NO_2廓线，并与PriAM反演的NO_2廓线进行对比，结果如图4所示。总体来说，CNN-SVR预测的对流层NO_2廓线与PriAM反演的NO_2廓线基本一致，且时间序列也高度一致，说明CNN-SVR模型可以很好地预测对流层NO_2廓线，其预测的近地面NO_2浓度值与PriAM反演结果最大相差约为15%，在0.6 km以上相较于PriAM几乎没有

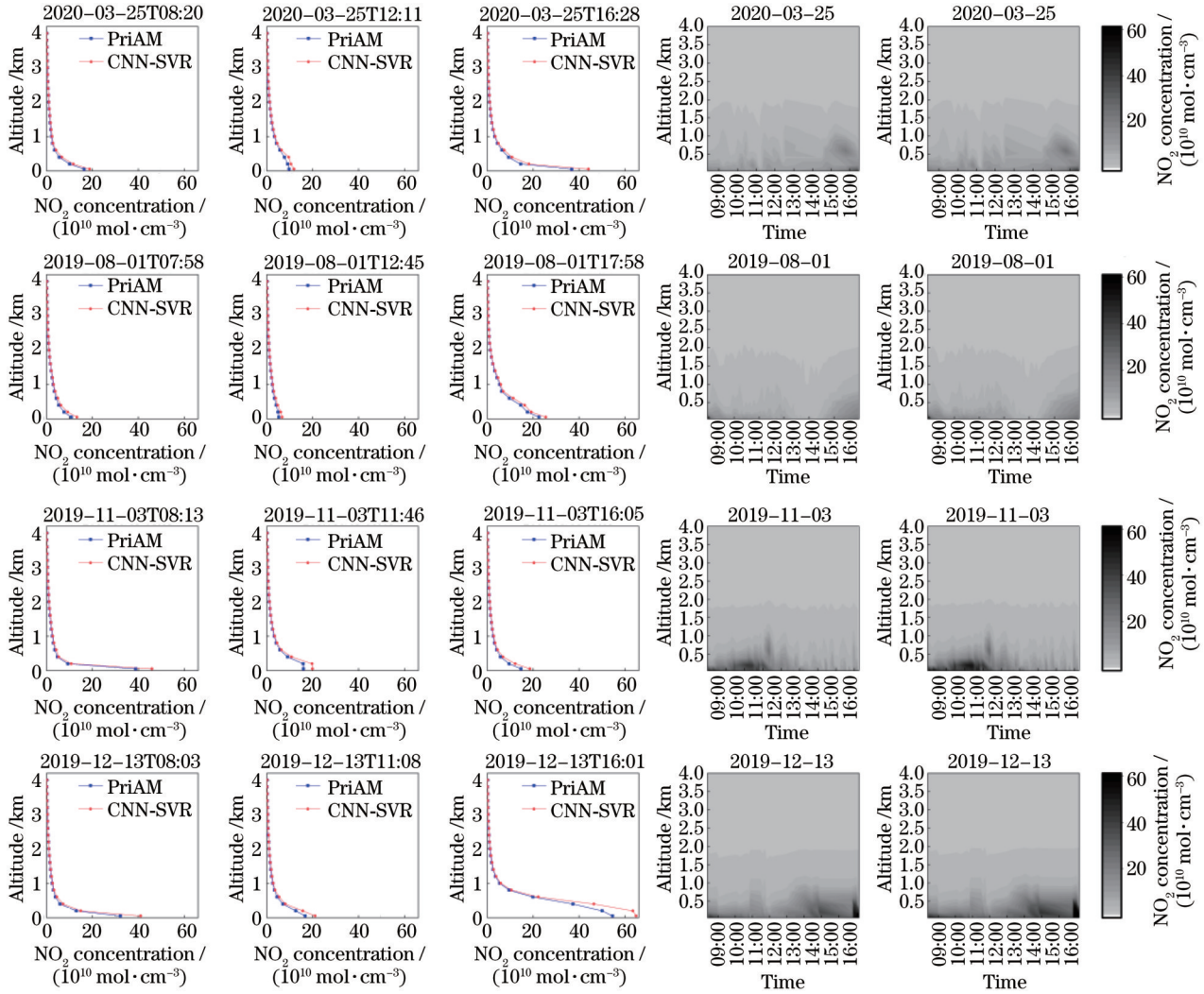

图4 CNN-SVR模型预测与MAX-DOAS反演的NO_2廓线对比

Fig. 4 Comparison of NO_2 profiles predicted by CNN-SVR model and retrieved by MAX-DOAS

差异。

为了进一步对比算法的可靠性和优劣性，利用2019年7月至2020年6月的MAX-DOAS数据对其NO_2廓线进行预测。选择该时段数据的主要原因是：1）2019年7—12月数据包括预测模型的建模集数据，可能导致预测数据的误差有一定偏向性，因此再利用2020年1—6月的月平均廓线计算其误差来验证模型的普适性；2）对流层NO_2的浓度受到季节因素的影响，使用连续一年的数据可以更好地验证模型，避免了某一季节的因素导致预测数据误差偏高或偏低。CNN-SVR模型预测的以及PriAM反演的NO_2的月平均廓线对比如图5所示。利用两种方法得到的NO_2廓线具有较好的一致性，表明CNN-SVR模型可以很好地预测对流层NO_2廓线。总体上CNN-SVR模型预测的NO_2浓度较PriAM反演的结果高约9.4%。反演误差主要集中在1 km及以下，该区域廓线的平均MAPE为18%。这是因为输出的是21个不同高度的对流层NO_2廓线，CNN-SVR预测模型在回归的过程中为了维持总体的误差，保证绝大多数高度可以预测准确，在预测中没有偏向将1 km以下的数据预测准确。

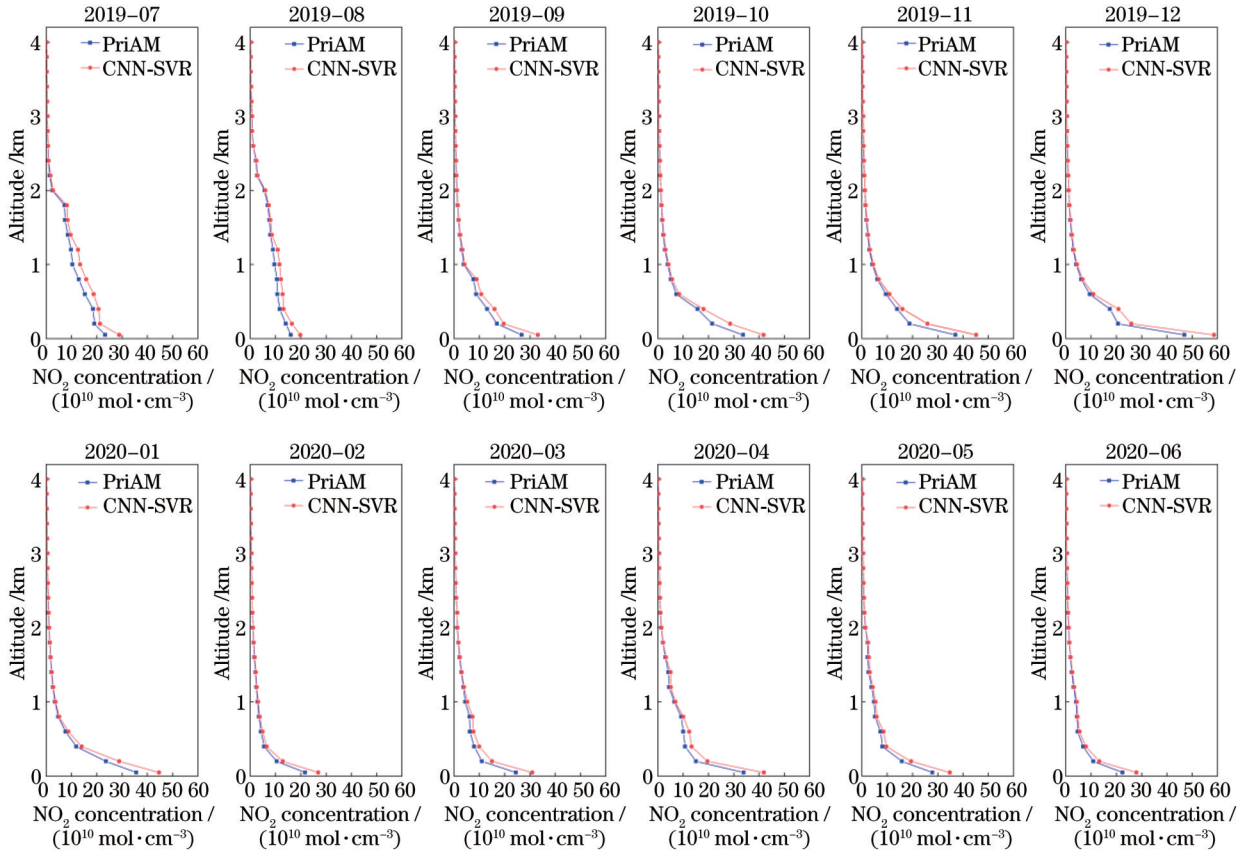

图5　2019年7月至2020年6月的NO_2月平均廓线

Fig. 5　Monthly average profile of NO_2 from July 2019 to June 2020

3.4　不同AOD的敏感性分析

由于气溶胶对于痕量气体廓线反演的影响较大，AOD的增大会一定程度地增加PriAM算法的平滑误差。研究发现，气溶胶AOD的错误对NO_2反演的影响尤为剧烈，气溶胶AOD的增加会在一定程度降低反演廓线的灵敏度，因此本研究开展了AOD对于CNN-SVR模型预测敏感性的分析。从2019年7月至2020年6月的数据也发现了在不同的AOD条件下，CNN-SVR预测结果与PriAM反演数据的平均百分比误差存在一定差异，结果如图6所示。当AOD大于3.0时，CNN-SVR与PriAM的平均MAPE最大（12.08%）；当AOD在0～0.5时，平均MAPE最小（6.72%）。不同AOD下CNN-SVR模型预测的MAPE范围如表6所示。可以明显看出，AOD在0～0.5范围内，预测MAPE的最大值为7.63%，而AOD大于3.0的MAPE最小值高达9.71%，最大预测MAPE更是AOD范围在0～0.5的2倍多。表6和图6的数据进一步证明了AOD的大小会影响PriAM反演的准确性，这是因为CNN-SVR模型的训练集中包含了一定数量的、由AOD的影响造成PriAM反演存在误差的数据，这些数据会对CNN-SVR模型最终的结果造成负面影响，从而降低模型精度。

4　结　　论

将CNN和SVR用于MAX-DOAS预测对流层NO_2廓线，提出一种有效的CNN-SVR模型，实现了通过MAX-DOAS原始光谱数据来实值回归预测对流层NO_2廓线，并取得了较好效果。通过MAX-DOAS数

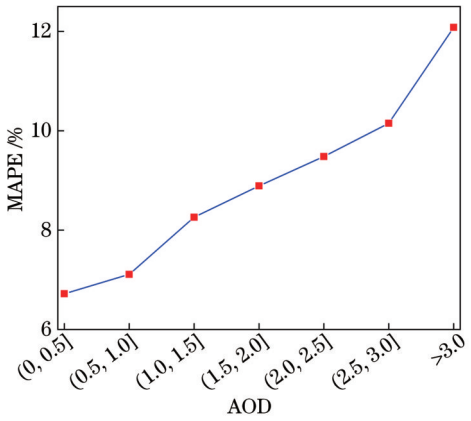

图 6 不同 AOD 范围下 CNN-SVR 模型预测误差

Fig. 6 Prediction error of CNN-SVR model under different AOD ranges

表 6 不同 AOD 范围下 CNN-SVR 模型预测误差的范围

Table 6 Prediction error range of CNN-SVR model under different AOD ranges

AOD	Prediction MAPE range /%
(0, 0.5]	(3.17, 7.63)
(0.5, 1.0]	(5.13, 8.21)
(1.0, 1.5]	(6.73, 9.16)
(1.5, 2.0]	(6.15, 11.19)
(2.0, 2.5]	(7.59, 11.80)
(2.5, 3.0]	(7.21, 13.17)
>3.0	(9.71, 16.23)

据、温度、AOD、低云覆盖率和 PriAM 反演的真实对流层 NO_2 廓线进行模拟实验，实验结果证明了该模型的优越性。CNN 可以提取更复杂的特征，提高计算速度；采用 SVR 算法对提取的特征进行预测，可提高预测精度。结果表明，与 CNN 模型、SVR 模型和 BP 模型相比，CNN-SVR 混合模型的误差减小了 8%～32%，能够快速准确地预测目标值，性能更加稳定。同时，选取了小时廓线、日均廓线和月均廓线结果进行对比，发现 CNN 模型预测的廓线更接近 MAX-DOAS 反演的真实 NO_2 廓线，说明了将 MAX-DOAS 数据输入 CNN-SVR 模型预测 NO_2 廓线的方法是有效的，尤其是 CNN 可以从原始光谱中学习关键模式，显著降低对特征工程的需求。多个处理层的组合提高了拟合和特征提取能力，使其适用于学习 MAX-DOAS 数据潜在特征，利用 MAX-DOAS 数据特征进行对流层 NO_2 廓线的实值回归预测。不同季节随机几天的对比结果表明 CNN-SVR 模型能较好地预测对流层 NO_2 廓线，预测结果与 PriAM 反演结果高度一致。利用建立的 CNN-SVR 模型对 2019 年 7 月至 2020 年 6 月南京站点的 NO_2 廓线进行预测，发现 CNN-SVR 模型能较好地预测对流层 NO_2 廓线，是一种有效的对流层 NO_2 廓线预测方法。同时分析了不同 AOD 大小对 CNN-SVR 模型预测精度的影响，AOD 越大，模型预测精度越低。

如今机器学习模型对于各种应用均取得不错的效果，但是在模型的设置中需要相当多的技能和经验来选择合适的参数，极度依赖精确的训练集输出数据和合适的输入数据。为了更准确地预测对流层 NO_2 廓线，需要找到更为适合的输入因子，这是预测对流层 NO_2 廓线的难点，也是未来基于 MAX-DOAS 数据通过 CNN-SVR 模型预测对流层 NO_2 廓线可进一步开展的工作。

参 考 文 献

[1] 周海金，刘文清，司福祺，等. 被动多轴差分吸收光谱技术监测大气 NO_2 垂直廓线研究[J]. 光学学报，2011，31(11)：1101007.
Zhou H J, Liu W Q, Si F Q, et al. Retrieval of atmospheric NO_2 vertical profile from multi-axis differential optical absorption spectroscopy[J]. Acta Optica Sinica, 2011, 31(11): 1101007.

[2] Platt U, Stutz J. Differential absorption spectroscopy [M]//Platt U, Stutz J. Differential optical absorption spectroscopy. Physics of earth and space environments. Heidelberg: Springer, 2008: 135-174.

[3] Frieß U, Sihler H, Sander R, et al. The vertical distribution of BrO and aerosols in the Arctic: measurements by active and passive differential optical absorption spectroscopy[J]. Journal of Geophysical Research: Atmospheres, 2011, 116(D14): D00R04.

[4] Hendrick F, Müller J F, Clémer K, et al. Four years of ground-based MAX-DOAS observations of HONO and NO_2 in the Beijing area[J]. Atmospheric Chemistry and Physics, 2014, 14(2): 765-781.

[5] Kanaya Y, Irie H, Takashima H, et al. Long-term MAX-DOAS network observations of NO_2 in Russia and Asia (MADRAS) during the period 2007-2012: instrumentation, elucidation of climatology, and comparisons with OMI satellite observations and global model simulations[J]. Atmospheric Chemistry and Physics, 2014, 14(15): 7909-7927.

[6] Li X, Brauers T, Hofzumahaus A, et al. MAX-DOAS measurements of NO_2, HCHO and CHOCHO at a rural site in Southern China[J]. Atmospheric Chemistry and Physics, 2013, 13(4): 2133-2151.

[7] Hönninger G, von Friedeburg C, Platt U. Multi axis differential optical absorption spectroscopy (MAX-DOAS)[J]. Atmospheric Chemistry and Physics, 2004, 4(1): 231-254.

[8] 吴丰成，谢品华，李昂，等. 基于多轴差分吸收光谱技术的查找表法反演气溶胶消光廓线研究[J]. 光学学报，2013，33(6)：0601002.
Wu F C, Xie P H, Li A, et al. Research of aerosol extinction inverted with look-up table method based on multi-axis differential optical absorption spectroscopy[J]. Acta Optica Sinica, 2013, 33(6): 0601002.

[9] 刘文清. “双碳”目标下大气环境光学监测技术发展机遇[J]. 光学学报，2022，42(6)：0600001.
Liu W Q. Opportunities and challenges for development of atmospheric environmental optics monitoring technique under "double carbon" goal[J]. Acta Optica Sinica, 2022,

42(6): 0600001.

[10] Irie H, Takashima H, Kanaya Y, et al. Eight-component retrievals from ground-based MAX-DOAS observations[J]. Atmospheric Measurement Techniques, 2011, 4(6): 1027-1044.

[11] 王璨, 武新慧, 李恋卿, 等. 卷积神经网络用于近红外光谱预测土壤含水率[J]. 光谱学与光谱分析, 2018, 38(1): 36-41.
Wang C, Wu X H, Li L Q, et al. Convolutional neural network application in prediction of soil moisture content [J]. Spectroscopy and Spectral Analysis, 2018, 38(1): 36-41.

[12] 唐永生, 陈争光. 卷积神经网络和近红外光谱的土壤pH值预测[J]. 光谱学与光谱分析, 2021, 41(3): 892-897.
Tang Y S, Chen Z G. Soil pH prediction based on convolution neural network and near infrared spectroscopy [J]. Spectroscopy and Spectral Analysis, 2021, 41(3): 892-897.

[13] Malmgren-Hansen D, Nielsen A A, Laparra V, et al. Transfer learning with convolutional networks for atmospheric parameter retrieval[C]//IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, July 22-27, 2018, Valencia, Spain. New York: IEEE Press, 2018: 2111-2114.

[14] Tian X, Xie P H, Xu J, et al. Long-term observations of tropospheric NO_2, SO_2 and HCHO by MAX-DOAS in Yangtze River Delta area, China[J]. Journal of Environmental Sciences, 2018, 71: 207-221.

[15] 熊险平, 何璇, 赵玉广. NO_2浓度变化及气象因子清除能力分析[J]. 环境科学导刊, 2020, 39(4): 27-33.
Xiong X P, He X, Zhao Y G. The analysis of NO_2 concentration changes and removal ability of meteorological factors[J]. Environmental Science Survey, 2020, 39(4): 27-33.

[16] Ren B, Xie P H, Xu J, et al. Use of the PSCF method to analyze the variations of potential sources and transports of NO_2, SO_2, and HCHO observed by MAX-DOAS in Nanjing, China during 2019[J]. Science of the Total Environment, 2021, 782: 146865.

[17] Wang Y, Lampel J, Xie P H, et al. Ground-based MAX-DOAS observations of tropospheric aerosols, NO_2, SO_2 and HCHO in Wuxi, China, from 2011 to 2014[J]. Atmospheric Chemistry and Physics, 2017, 17(3): 2189-2215.

[18] 王杨, 李昂, 谢品华, 等. 多轴差分吸收光谱技术测量NO_2对流层垂直分布及垂直柱浓度[J]. 物理学报, 2013, 62(20): 200705.
Wang Y, Li A, Xie P H, et al. Measuring tropospheric vertical distribution and vertical column density of NO_2 by multi-axis differential optical absorption spectroscopy[J]. Acta Physica Sinica, 2013, 62(20): 200705.

[19] Fu D S, Liu M Q, Yang D Z, et al. Influences of atmospheric reanalysis on the accuracy of clear-sky irradiance estimates: comparing MERRA-2 and CAMS [J]. Atmospheric Environment, 2022, 277: 119080.

[20] 田鑫, 徐晋, 谢品华, 等. 基于多轴差分吸收光谱技术测量对流层HCHO垂直分布[J]. 光谱学与光谱分析, 2019, 39(8): 2325-2331.
Tian X, Xu J, Xie P H, et al. Retrieving tropospheric vertical distribution in HCHO by multi-axis differential optical absorption spectroscopy[J]. Spectroscopy and Spectral Analysis, 2019, 39(8): 2325-2331.

[21] Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators[J]. Neural Networks, 1989, 2(5): 359-366.

[22] Buckman J, Roy A, Raffel C, et al. Thermometer encoding: one hot way to resist adversarial examples[C]//6th International Conference on Learning Representations 2018, April 30-May 3, 2018, Vancouver, BC, Canada. [S.l.: s.n.], 2018.

[23] Dombi G W, Nandi P, Saxe J M, et al. Prediction of rib fracture injury outcome by an artificial neural network[J]. The Journal of Trauma, 1995, 39(5): 915-921.

[24] 张祥东, 王腾军, 朱劭俊, 等. 基于扩张卷积注意力神经网络的高光谱图像分类[J]. 光学学报, 2021, 41(3): 0310001.
Zhang X D, Wang T J, Zhu S J, et al. Hyperspectral image classification based on dilated convolutional attention neural network[J]. Acta Optica Sinica, 2021, 41(3): 0310001.

[25] 朱黎明, 孙刚, 陈多龙, 等. 基于支持向量机估算大气光学湍流廓线的研究[J]. 光学学报, 2022, 42(1): 0101001.
Zhu L M, Sun G, Chen D L, et al. Atmospheric optical turbulence profile estimation using support vector machine [J]. Acta Optica Sinica, 2022, 42(1): 0101001.