光学学报

基于荧光光谱的水体分类与荧光组分识别方法

陈庆[1]，汤斌[1*]，缪俊锋[1]，周彦[3]，龙邹荣[1**]，张金富[1]，王建旭[1]，周密[1]，叶彬强[1,2]，赵明富[1]，钟年丙[1]

[1]重庆理工大学光纤传感与光电检测重庆市重点实验室，重庆 400054；
[2]重庆大学微电子与通信工程学院，重庆 400044；
[3]重庆市铜梁区环境保护局，重庆 402560

摘要 提出了一种基于MobileNetV2和VGG11组分拟合（CF-VGG11）卷积神经网络（CNN）与平行因子分析（PARAFAC）结合的水样分类和荧光组分拟合方法，通过输入单个三维荧光光谱（3D-EEM）数据来预测水样类别、溶解性有机物（DOM）质量浓度等级和荧光组分。算法以PARAFAC结果为基础建立荧光光谱数据集，分两步完成类别与组分的预测：第一步使用MobileNetV2算法对不同水样进行类别预测和DOM质量浓度分级；第二步使用CF-VGG11网络拟合荧光组分。采集地表水、工业废水处理水、污水处理厂进出口水和乡村饮用水4种类型的水样构建数据集，获得了95.83%的分类精度和98.11%的组分拟合精度。实验结果表明，所提方法可对不同水样和DOM质量浓度等级进行准确分类，拟合特定荧光组分，精确定位污染源，并能进行超标预警。

关键词 光谱学；三维荧光光谱；水污染；分类；卷积神经网络

中图分类号 X84；O433.5 **文献标志码** A **DOI：** 10.3788/AOS221518

1 引言

地表水、饮用水和废水中有机污染物的治理是目前人类社会发展迫切需要解决的社会问题之一[1]。三维荧光光谱（3D-EEM）常使用主成分分析、支持向量机、荧光区域积分和平行因子分析（PARAFAC）方法对3D-EEM进行分析[2-4]。其中，PARAFAC不仅可以估计3D-EEM中荧光团的数量，还可用于估算水质指数，证实三维荧光指数、荧光成分与地表水水质指数之间存在显著相关性[5]。3D-EEM技术是近几十年来快速发展的一种新型应用技术，因速度快、信息含量丰富和无二次污染等特点，近年来在食品、医疗和环境检测等领域中得到了快速发展[6-10]。由于3D-EEM的荧光特征会随着荧光有机物的种类和质量浓度的不同而发生改变[11]，故水环境检测领域中常将其用于鉴别污水种类，进而识别排放源。原始的3D-EEM数据由数千个激发/发射对和相应的荧光强度值组成。由于3D-EEM中存在大量干扰噪声和荧光重叠信息，故亟需一种快速准确的方法来提取和分析3D-EEM中的有用信息[12]。

3D-EEM数据由激发波长、发射波长和荧光强度组成，与图像数据结构（高度、宽度和灰度值）相同。卷积神经网络（CNN）常用于处理图像数据[13]，并已被越来越多地应用于科学和工业领域中[14-15]。然而，将CNN应用于3D-EEM的荧光组分识别近年才有学者开始关注。Wu等[16]使用CNN来鉴别假芝麻油，先用CNN网络来提取光谱特征，再使用支持向量机与偏最小二乘法分别对假芝麻油分类和掺假质量浓度进行预测。一些学者[4, 17]采用CNN对3D-EEM数据进行快速分类和组分图谱拟合，既克服了传统的PARAFAC方法的复杂耗时（数据预处理、异常值检验和拆半验证等）且需大量数据的局限性，又确保了来源识别的准确性，取得了不错的效果。综上所述，阐明了将CNN应用在3D-EEM数据中的可行性，并且具有坚实的理论基础。

基于此，本文将MobileNetV2[18]、VGG11[19] CNN与PARAFAC方法相结合。采集不同水样的3D-EEM数据，以PARAFAC的分析结果作为标签，建立不同水样多输出分类模型和荧光组分图谱预测拟合模型。其中，MobileNetV2网络用于识别不同水样来源与溶解性有机物（DOM）质量浓度等级（采用氨氮质量浓度来进行表征[20]），VGG11组分拟合网络（CF-VGG11）用于对不同水样荧光组分图进行拟合。相较其他卷积网络模型，所提模型不用为每个组件图创建

收稿日期：2022-07-22；**修回日期**：2022-08-22；**录用日期**：2022-09-22；**网络首发日期**：2022-10-02

基金项目：国家自然科学基金（61805029）、重庆市自然科学基金面上项目（cstc2020jcyj-msxmX0879）、重庆市高校创新研究群体项目（CXQT21035）、重庆市铜梁区科学技术局技术创新与应用发展专项（CCF20220623）、重庆理工大学科研启动经费资助项目（0107210299、0107200283）

通信作者：*tangbin@cqut.edu.cn；**longzourong@cqut.edu.cn

子模型，且无需考虑荧光团数量相同而无法进行分类识别的问题，从而更具普适性。实验表明，结合PARAFAC的分析结果，所提模型只需输入单个水样的3D-EEM，就能预测出水样类别和DOM质量浓度等级，并对其荧光组分进行拟合。所提方法不用重复进行PARAFAC分析，仅使用两个CNN模型就能快速获得水样类别、DOM质量浓度等级和组分图，为地表水、饮用水和废水监测等需要快速检测的场景提供了有效的技术手段。

2　原理及方法

2.1　数据前处理

3D-EEM数据的前处理包括两部分：第一部分是对原始3D-EEM数据进行处理，即扣除空白、剪除拉曼散射与瑞利散射和进行拉曼归一化，以便于PARAFAC分析模型的构建；第二部分是将处理后的3D-EEM数据转换为图像，作为输入的数据集用于训练和测试CNN模型。

本次样品并不是同一时间测量的，为了能将不同时间上测得的数据进行比较分析，需要对数据进行标准化处理，即进行拉曼归一化，将数据转换为拉曼单位，拉曼归一化公式为

$$A_{\mathrm{rp}}^{\lambda_{\mathrm{ex}}}=\int_{\lambda_{\mathrm{em}}^{(1)}}^{\lambda_{\mathrm{em}}^{(2)}} I_{\lambda_{\mathrm{em}}}\,\mathrm{d}\lambda_{\mathrm{em}}, \tag{1}$$

$$F_{\lambda_{\mathrm{ex}},\lambda_{\mathrm{em}}}=\frac{I_{\lambda_{\mathrm{ex}},\lambda_{\mathrm{em}}}}{A_{\mathrm{rp}}}, \tag{2}$$

式中：λ_{ex}为激发波长；$I_{\lambda_{\mathrm{em}}}$为在发射波长$\lambda_{\mathrm{em}}$处的拉曼峰的光谱校正强度；$A_{\mathrm{rp}}$为$\lambda_{\mathrm{em}}$在371～428 nm范围内（$\lambda_{\mathrm{ex}}$=350 nm）荧光强度的积分值；$I_{\lambda_{\mathrm{ex}},\lambda_{\mathrm{em}}}$为拉曼归一化前的光谱校正强度，其单位为arb. units；$F_{\lambda_{\mathrm{ex}},\lambda_{\mathrm{em}}}$为拉曼归一化后的荧光强度，其单位为拉曼单位。

2.2　PARAFAC方法

PARAFAC方法[21]是基于三线性分解理论，采用交替最小二乘算法实现的一种数学模型。实际样本得到的荧光光谱数据是一个$I\times J\times K$型的三维响应矩阵$\boldsymbol{X}$，其中：K为样本数量；I为激发波长点的扫描数量；J为发射波长点的扫描数量。使用工具箱DOM Fluor（V1.7）应用PARAFAC以数学方式将数据信号分解成一组三线性项和残差数组，相应的公式为

$$x_{ijk}=\sum_{f=1}^{F} a_{if}b_{jf}c_{kf}+e_{ijk}, \tag{3}$$

式中：i=1，2，⋯，I；j=1，2，⋯，J；k=1，2，⋯，K；x_{ijk}为三维响应矩阵$\boldsymbol{X}$中的任一元素；a_{if}为相对激发矩阵$\boldsymbol{A}_{I\times F}$中的任一元素；$b_{jf}$为相对发射矩阵$\boldsymbol{B}_{J\times F}$中的任一元素；$c_{kf}$为相对质量浓度矩阵$\boldsymbol{C}_{K\times F}$中的任一元素；$\boldsymbol{e}_{ijk}$为三维残差矩阵$\boldsymbol{E}_{I\times J\times K}$中的任一元素；$F$为矩阵$\boldsymbol{A}_{I\times F}$、$\boldsymbol{B}_{J\times F}$和$\boldsymbol{C}_{K\times F}$的列数，代表样品中组分的数量。图1为PARAFAC方法的整体流程。

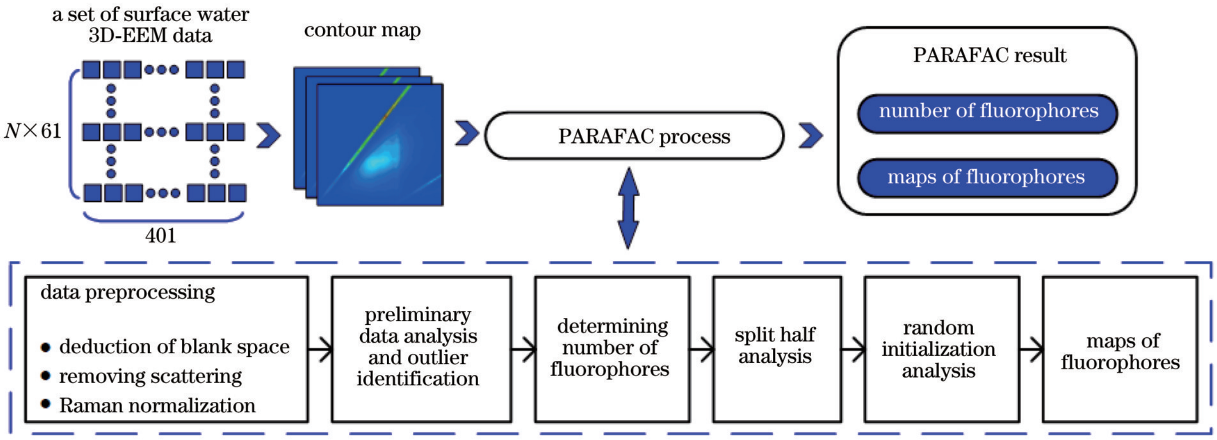

图1　PARAFAC方法的整体流程

Fig. 1　Overall flow of PARAFAC method

2.3　基于CNN快速分类识别模型的构建

建立的CNN快速分类识别模型如图2所示。对采集的水样3D-EEM前处理后的数据x_i（样品数不小于20）进行PARAFAC，从而得出荧光组分数量y_{n}和荧光组分图谱y_{m}，并将用化学法确定下来的水样类别y_i和DOM质量浓度等级y_j设置为CNN快速分类识别模型的分类标签。采用改进的MobileNetV2网络模型进行水样类别和DOM质量浓度等级多输出分类，采用CF-VGG11模型进行荧光组分图谱拟合。PARAFAC仅在训练数据集准备阶段使用，在应用阶段无需重复进行PARAFAC，可直接将新采集的水样3D-EEM（样品数不小于1）输入到训练好的CNN快速分类识别模型中，快速分类识别出水样类别（组件数量）、DOM质量浓度等级和组件图谱。

将整个网络模型训练好后，新采集的3D-EEM数据经过数据前处理后，输入到MobileNetV2分类网络和CF-VGG11拟合网络中进行分析。本研究包括三个部分：1）数据预处理（PARAFAC和CNN模型准备

输入数据集）；2）先训练 MobileNetV2 分类网络，结合 PARAFAC 的结果和化学法测量的 DOM 质量浓度等级作为 MobileNetV2 分类网络的标签，模型得出水样类别（组分数量）和 DOM 质量浓度等级的多输出分类结果，再训练 CF-VGG11 拟合网络，将 PARAFAC 分析获得的组件图谱作为 CF-VGG11 网络模型的标签，最后将 3D-EEM 数据输入到 CF-VGG11 网络的每个子网络中，输出拟合的各组件图谱；3）输入一个新 3D-EEM 数据到训练好的网络模型中，可以预测水样类别、DOM 质量浓度等级并对特定组分图进行拟合。其中，前两个部分为准备阶段，第三个部分为应用阶段。

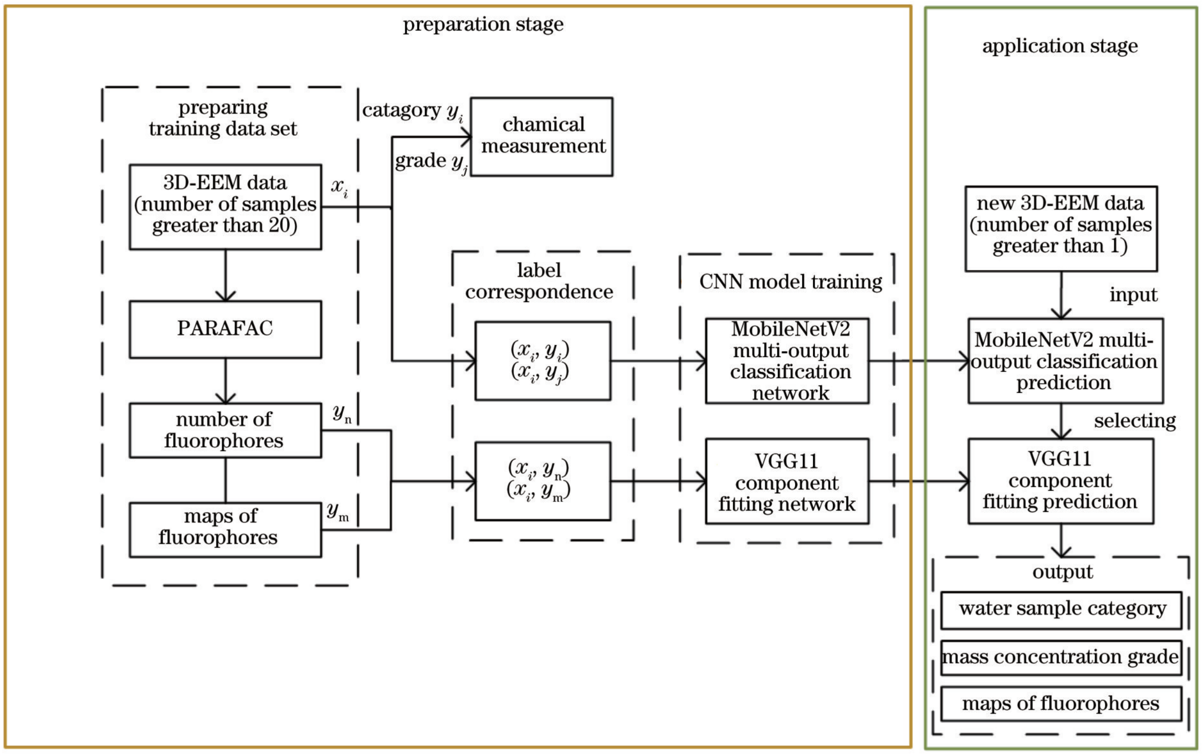
preparation stage
application stage
preparing training data set
catagory y_i
grade y_j
chamical measurement
3D-EEM data (number of samples greater than 20)
x_i
PARAFAC
number of fluorophores
y_n
maps of fluorophores
y_m
label correspondence
(x_i, y_i) (x_i, y_j)
(x_i, y_n) (x_i, y_m)
CNN model training
MobileNetV2 multi-output classification network
VGG11 component fitting network
new 3D-EEM data (number of samples greater than 1)
input
MobileNetV2 multi-output classification prediction
selecting
VGG11 component fitting prediction
output
water sample category
mass concentration grade
maps of fluorophores

图 2　CNN 快速分类识别模型

Fig. 2　CNN fast classification and recognition model

MobileNetV2 网络可以在移动设备或嵌入式设备上运行深度学习模型，结合实际光谱数据类型，本实验参照 MobileNetV2 模型进行网络结构的改进和参数优化。基于 MobileNetV2 的 3D-EEM 分类网络如图 3 所示。MobileNetV2 网络的第一层为全卷积层，随后是 4 层线性瓶颈结构层，其中使用了新的激活函数 ReLu6，最后使用全连接层（FC）。为避免过拟合、提高泛化性能和提升训练速度，在训练过程中使用比例为 0.5 的随机失活（Dropout）和批标准化（BatchNorm）。在 MobileNetV2 模型中输入尺寸为 1×224×224 的水样等高特征图像，采用交叉熵函数作为模型训练的损失函数，将准确率（A）作为分类精度，利用 Adam 优化器对模型训练进行优化。交叉熵损失函数 L 与准确率的计算公式为

$$L=-\sum_{n=1}^{N} y_n \ln p_n \,, \tag{4}$$

$$A=\frac{V_{\mathrm{TP}}+V_{\mathrm{TN}}}{V_{\mathrm{TP}}+V_{\mathrm{TN}}+V_{\mathrm{FP}}+V_{\mathrm{FN}}} \,, \tag{5}$$

式中：N 为种类数量；y_n 若为类别 n 的标签，则 $y_n=1$，否则为 0；p_n 为类别为 n 的神经网络的输出概率；V_{TP} 为真正例的数量；V_{TN} 为真负例的数量；V_{FP} 为假正例的数量；V_{FN} 为假负例的数量。损失函数的下降采用 Adam 优化器，将其学习率设置为 0.00003。为了使多输出分类模型能有序地进行参数更新和实现快速收敛，训练样本被分成多个批次，其中批处理样本数目设置为 32。

使用 MobileNetV2 网络对 3D-EEM 数据进行分类后，就可以得到各水样的类别（组分数量）和 DOM 质量浓度等级。然后，以改进的 CF-VGG11 网络作为三维荧光组分映射分析的模型，对三维荧光样品的组分图谱进行拟合。图 4（a）为单 CF-VGG11 网络结构，输入为大小为 60×60 的各类水样灰度图像，输出为长度为 3600 的向量，输出向量可变成大小为 60×60 的矩阵。CF-VGG11 网络由 5 个卷积层和 3 个全连接层构成，第一个和第二个全连接层之后有一个 Dropout 层，用于防止过拟合。CF-VGG11 网络通过均方误差（MSE）损失函数进行训练，通过余弦相似度验证训练效果。余弦相似度的计算公式为

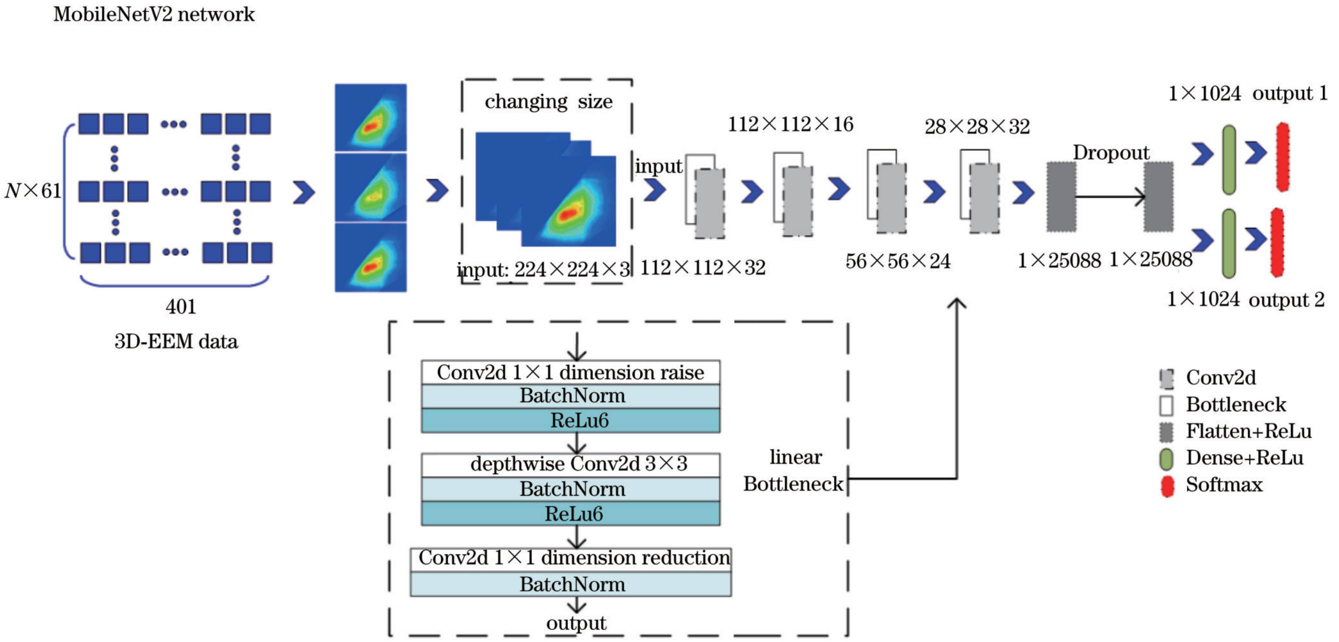

图 3　基于 MobileNetV2 的 3D-EEM 分类网络

Fig. 3　3D-EEM classification network based on MobileNetV2

$$\cos\theta=\frac{\sum_{i=1}^{n}(M_i\times N_i)}{\sqrt{\sum_{i=1}^{n}M_i^2}\sqrt{\sum_{i=1}^{n}N_i^2}}\text{，}\tag{6}$$

式中：n为总样本个数；向量$\boldsymbol{M}$和向量$\boldsymbol{N}$分别为拟合组分图谱和组分图谱；θ为两个向量的夹角，夹角越小，表明两个组分图谱的相似度越高。采用Adam优化器，将其学习率设置为0.00001，批处理样本数目设置为8。图4（b）为整个拟合网络的结构，以大小为1×60×60的灰度图像作为输入，并输出一个大小为3600的矢量（通道为1、长度为60、宽度为60），该矢量代表组分图的荧光强度值。采用PARAFAC结果作为采集的地表水、工业废水处理水、污水处理厂进出口水和乡村饮用水4类水样的CF-VGG11网络数据集的标签，这4组水样具有4个组分数量，再使用组分图（C1～C4）作为CF-VGG11训练的拟合标签，分别从C1～C4组分出发对每种类型的3D-EEM的CF-VGG11网络分量进行编号，最后输出大小为60×60矩阵拟合的各类水样组分图谱。

3　实验部分

3.1　实验样品

在重庆市铜梁区采集不同水样的3D-EEM数据，第一组包含51个地表水样本，第二组包含127个工业废水处理水样本，第三组包含37个污水处理厂进出口水样本，第四组包含58个乡村饮用水样本。数据集来自同一地区，水样来源不同，但它们具有相似的3D-EEM数据特征。采集样品的详细信息如表1所示。

3.2　荧光测量

使用上海棱光技术有限公司的F98荧光光谱仪对各类水样的荧光特性进行研究。设置实验扫描激发波长为250～550 nm（步长为5 nm），发射波长为250～650 nm（步长为1 nm）。激发光和发射光的狭缝宽度为10 nm，扫描速度为30000 nm/min，增益电压为750 V，空白水样为超纯水。用荧光光谱仪扫描的样品得到对应的激发-发射矩阵，每个激发-发射矩阵由61个激发波长和401个发射波长构成，其维度为61×401。这些矩阵以轮廓图的形式表示（横轴表示激发波长，纵轴表示发射波长，轮廓表示荧光强度值$F_{\lambda_{\text{ex}},\lambda_{\text{em}}}$）。

4　结果与讨论

4.1　PARAFAC的结果

对采集的3D-EEM数据先进行数据前处理，用于生成PARAFAC的数据集与CNN模型的训练集和测试集。如表1所示采集了273个样品的3D-EEM数据进行PARAFAC，每组样本的数量超过20个，满足PARAFAC对数据量的要求。

每个分类数据集中的3D-EEM数量不平衡，DB、FS、WS和XCYY分别有51、127、37、58个3D-EEM图谱。为了减少数量不均衡带来的隐含偏差，将DB、XCYY的3D-EEM复制一次，WS的3D-EEM复制两次。使这4个类别的数据量相似，即DB、FS、WS和XCYY分别有102、127、111、116个3D-EEM。最后，将这4个数据集整合为用于分类的数据集（456个样本）。验证PARAFAC结果显示这些3D-EEM数据样本都为4个荧光组分。将具有相同荧光组分的3D-EEM数据样本整合为一个数据集，从而获得4个分类数据集。利用PARAFAC方法对各类水样的数据集分别进行分析。以DB水样为例，PARAFAC方法的分析

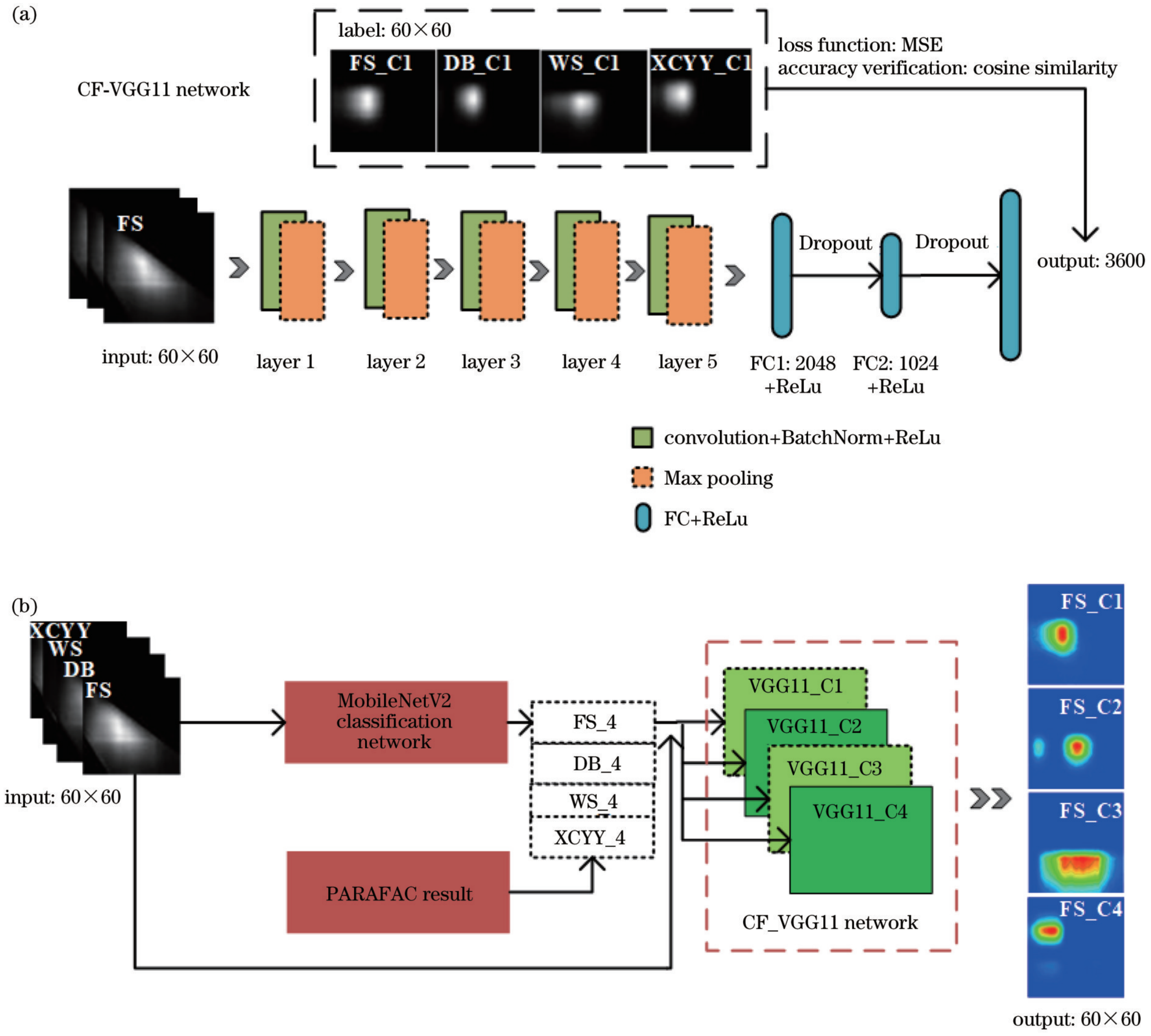

图 4　基于 CF-VGG11 的 3D-EEM 拟合网络。(a)单层 CF-VGG11 网络结构；(b) 整个拟合网络的结构

Fig. 4　3D-EEM fitting network based on CF-VGG11. (a) Structural diagram of single-layer CF-VGG11 network; (b) structural diagram of whole fitting network

表 1　水样采集

Table 1　Water sample collection

Type	Label	Number
Surface water	DB	51
Treatment water of industrial wastewater	FS	127
Inlet and outlet water of sewage treatment plant	WS	37
Rural drinking water	XCYY	58

过程如图 5 所示。对所有数据集进行了 PARAFAC，表 2 和图 6 显示了 PARAFAC 的结果。如表 2 所示将 PARAFAC 结果上传到 OpenFluor 数据库中[22]进行比较，得到各类水样荧光成分的可能物质。对比结果显示，某些组分在数据库中比对只有一个结果，但所有组件的相似度对比得分均超过 95%。

4.2　模型性能

将用于 CNN 快速分类识别模型的训练集和测试集比例按 8∶2 来划分，即 365 个样本用于训练，91 个样本用于测试。利用交叉熵损失函数和训练正确率对 MobileNetV2 分类网络进行了评价。CF-VGG11 网络结构包含 4 个 VGG11 网络模型结构，如图 4 所示，其中每个子模型基于 3D-EEM 获得与 4 个数据集对应的组分图谱。使用 CF-VGG11 模型来拟合组分图，输出大小为 60×60 的矢量。传统的目标检测和目标识别算法不能判断结果，故采用余弦相似度表征模型拟合结果的精度，并采用均方值误差表征模型的损失值。

模型训练结果如表 3 所示。MobileNetV2_1 网络模型用于水样类别分类，对训练集和测试集的分类准确率最高可分别达到 99.65% 和 98.61%，其损失值分别为 0.2500% 和 2.0000%。MobileNetV2_2 网络模型用于 DOM 质量浓度等级分类，对训练集和测试集的分类准确率最高可分别达到 99.30% 和 95.83%，训练集和测试集的损失值分别为 2.3800% 和 12.6000%。MobileNetV2 多输出模型对样品分类的

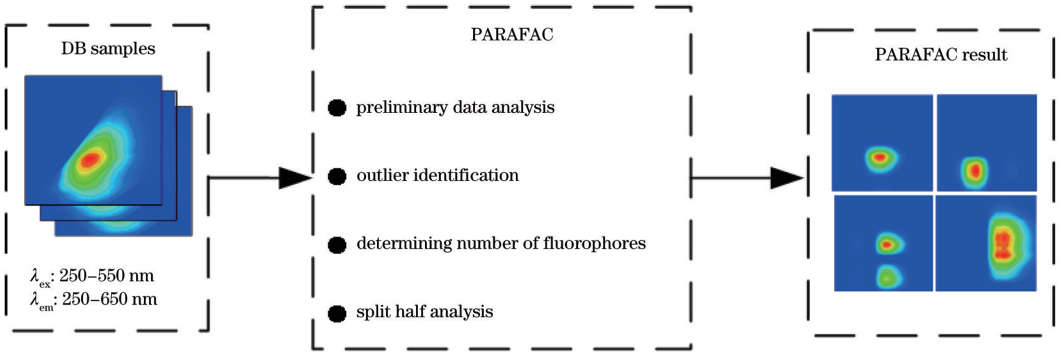
DB samples
λ_{ex}: 250–550 nm
λ_{em}: 250–650 nm
PARAFAC
preliminary data analysis
outlier identification
determining number of fluorophores
split half analysis
PARAFAC result

图5　PARAFAC方法的分析过程

Fig. 5　Analysis process of PARAFAC method

表2　从PARAFAC组件和OpenFluor数据库的比对结果看EEM的光谱特征

Table 2　Spectral characteristics of EEM seen from comparison results of PARAFAC components and OpenFluor database

Type	Component	λ_{ex} /nm	λ_{em} /nm	Fluorescent substance	Number of OpenFluor matches
DB	C1	370	464	Humic acid[23]	11
	C2	315	384	Microbial humus[24]	21
	C3	290, 395	460	Terrestrial humus[25]	4
	C4	360, 395	526	Soil fulvic acid[26]	1
FS	C1	350	428	Waste water collection tracer[27]	17
	C2	275, 400	480	Terrestrial humus[25]	1
	C3	360, 395	526	Soil fulvic acid[26]	1
	C4	315	384	Microbial humus[28]	7
WS	C1	<270, 365	465	Terrestrial humus[29]	2
	C2	345	409	Humus like substance[30]	1
	C3	295	369	Microbial humus[31]	8
	C4	275, 420	488	Microbial humus[32]	3
XCYY	C1	345	429	Anthropogenic humus[33]	9
	C2	390	454	Fulvic acid and humus[34]	1
	C3	<270, 365	465	Terrestrial humus[35]	4
	C4	360, 395	526	Soil fulvic acid[26]	1

效果很好，而对DOM质量浓度等级分类预测的效果相对差一些，但在可接受范围内。CF-VGG11网络模型用于各水样荧光组分拟合，对训练集和测试集的拟合准确率最高可分别达到99.60%和98.11%，其损失值分别为0.0879%和0.0455%，表明CF-VGG11模型拟合荧光图谱的性能极佳。

为了表明训练好的MobileNetV2和CF-VGG11模型具有良好的分类和拟合性能，选择一部分未经过训练的3D-EEM数据进行测试，将CF-VGG11模型拟合的组分与PARAFAC的结果进行比较。如图7所示，两类水样各有4组分荧光图谱，荧光成分如表2所示。PARAFAC结果中横坐标为发射波长，纵坐标为激发波长，荧光强度为拉曼归一化后的荧光强度。CF-VGG11模型拟合图谱中横坐标由($\lambda_{em}-250$)/6.67求得，纵坐标由($\lambda_{ex}-250$)/5.00求得。从图7可以看出，模型拟合的结果与PARAFAC方法得到的结

表3　模型训练结果

Table 3　Results of model training　　unit: %

Network model	Accuracy of training set	Accuracy of test set	Loss value of training set	Loss value of test set
MobileNetV2_1	99.65	98.61	0.2500	2.0000
MobileNetV2_2	99.30	95.83	2.3800	12.6000
CF-VGG11	99.60	98.11	0.0879	0.0455

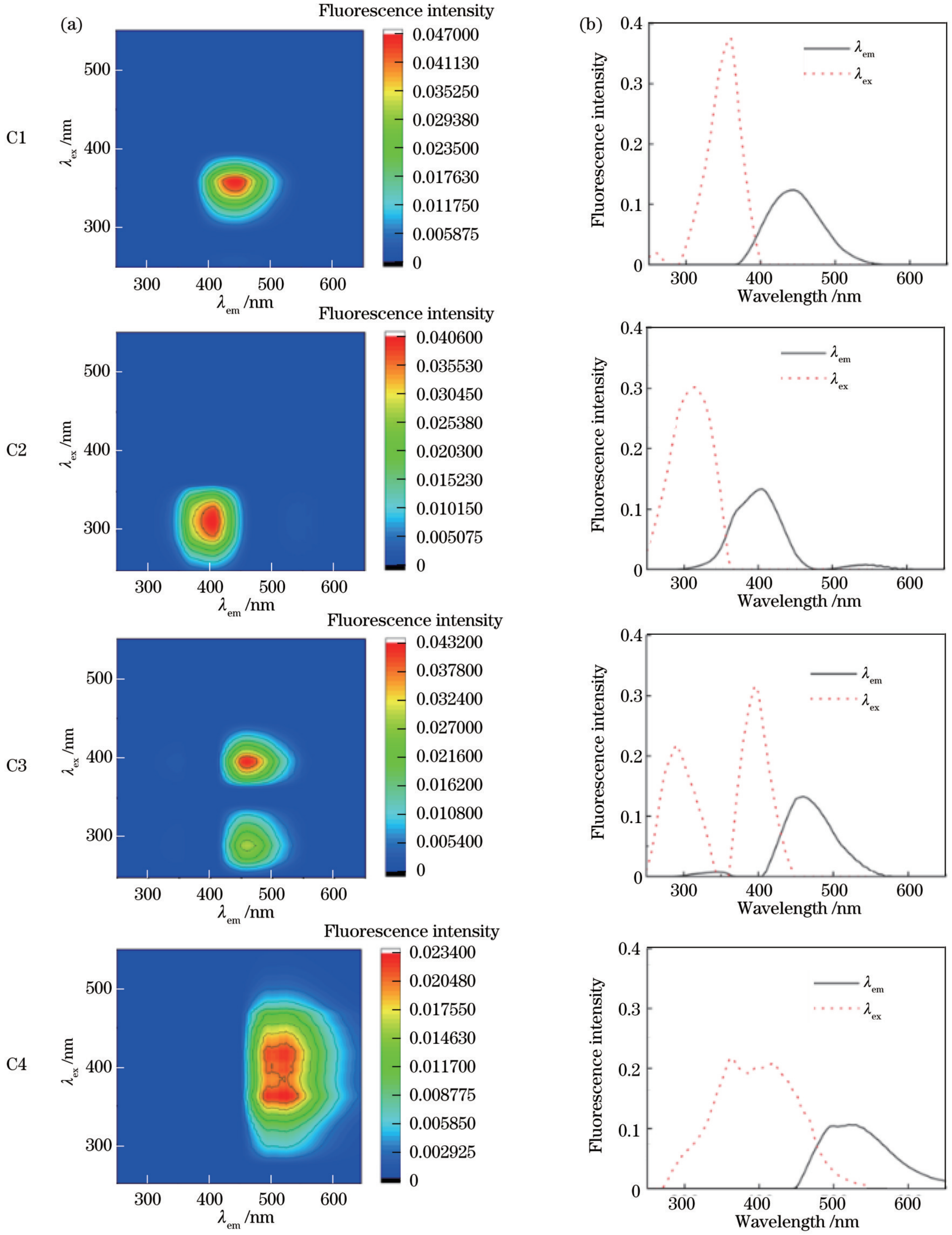

图 6 PAFARAC 得到的 DB 分析结果。(a)组件图谱;(b)加载组件图谱对应的拆半验证结果

Fig. 6 Analysis results of DB obtained by PAFARAC. (a) Component maps; (b) results of split half verification corresponding to loading component maps

果非常相似,不同的是模型拟合的结果中有几个噪声点,这些噪声点对 3D-EEM 的分量影响不大。因此,可以得出,MobileNetV2 和 CF-VGG11 可以很好地对水样的 3D-EEM 进行分类和拟合。

4.3 所提方法与 PARAFAC 方法的比较

所提模型是将 PARAFAC 方法与 CNN 相结合,先使用 PARAFAC 方法来确定正确的组分数量和组分图谱,再将组分作为标签对 CNN 模型进行训练,从而实现对水质 3D-EEM 数据分类和组分图拟合的目的。此外,MobileNetV2 和 CF-VGG11 网络模型与 PARAFAC 方法相比具有一定优势,训练好的模型只需要输入一个 3D-EEM 数据样本就可以得到对应的

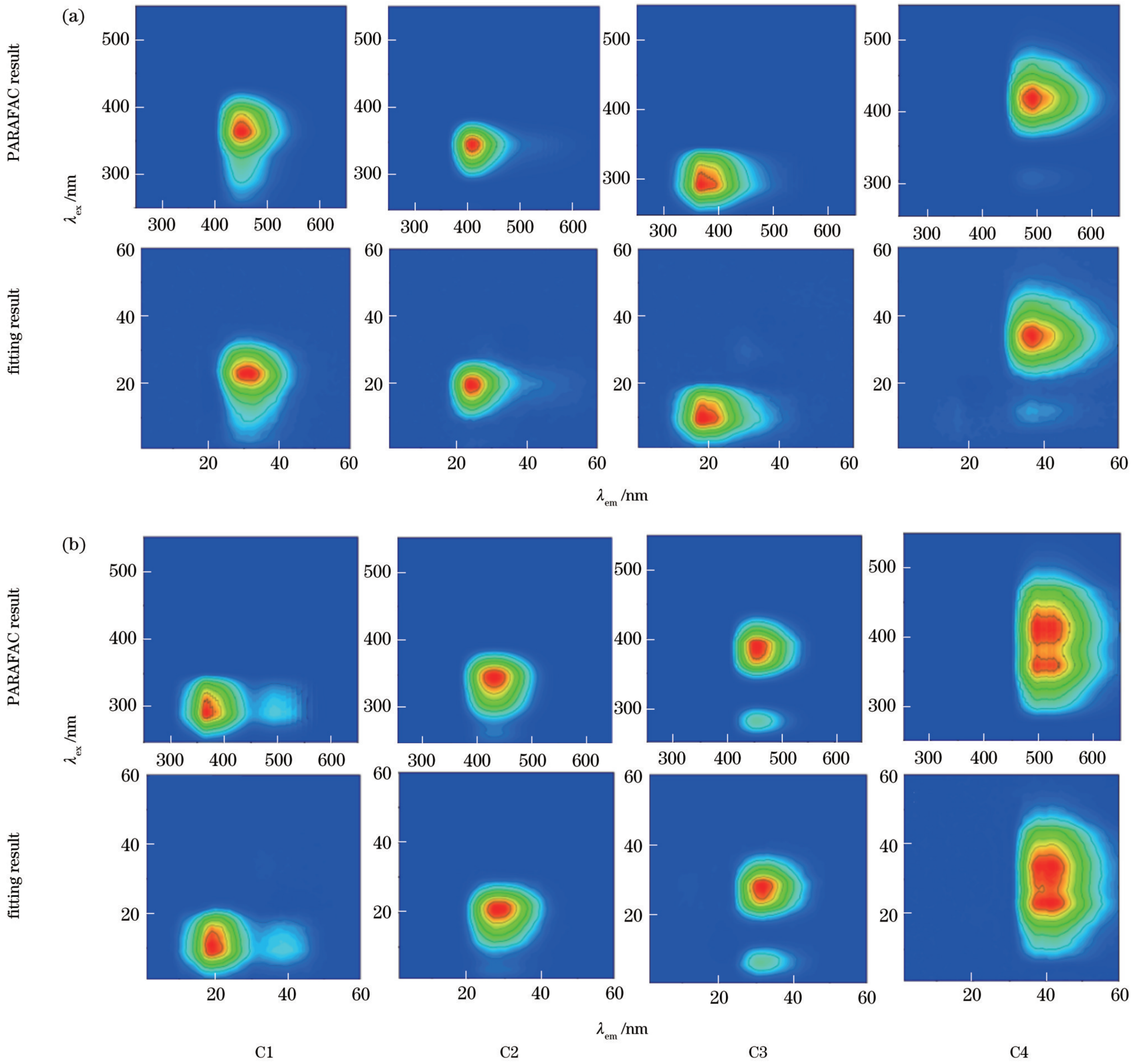

图 7　单个 3D-EEM 谱的 PARAFAC 结果和 CNN 模型拟合组分图谱。(a) WS 水样；(b)XCYY 水样

Fig. 7　PARAFAC result of 3D-EEM spectrum and fitting component map obtained by CNN model. (a) WS water sample; (b) XCYY water sample

水样类别、DOM 质量浓度等级和组分图，在快速识别水样 3D-EEM 分析场景中具有巨大的应用价值。CNN 快速分类识别模型与 PARAFAC 方法的比较结果如表 4 所示。近年来，许多学者尝试用地表水的 3D-EEM 数据进行快速分析用于表征其 DOM 污染。基于所提模型能够实现对各类水样的快速分析。

表 4　PARAFAC 方法与所提模型的比较

Table 4　Comparison between PARAFAC method and proposed model

Model	Data quantity requirement	Operation environment	Time cost	Analysis process
PARAFAC	≥20	MATLAB	High	Complex
MobileNetV2+CF-VGG11	≥1	Python	Low	Simple

5　结　　论

提出了一种基于荧光光谱的 CNN 快速分类识别算法来对不同水样进行类别预测和 DOM 质量浓度分级，并快速预测 3D-EEM 中重叠的荧光成分。依靠 PARAFAC 进行初步数据准备，用 MobileNetV2 网络进行水样类别和 DOM 质量浓度等级分类，可实现水污染溯源和超标预警。用 CF-VGG11 网络对水样荧

光组分图进行拟合。实验结果表明，基于PARAFAC结果建立的快速分类识别网络模型，只需输入单个水样的3D-EEM数据，就能快速预测出水样类别和DOM质量浓度等级，并对其特定荧光组分进行拟合。所提模型无需重复地进行复杂的PARAFAC，为水污染实时三维荧光分析场景提供了一种新的技术手段。

参考文献

[1] Duan P F, Wei M J, Yao L G, et al. Relationship between non-point source pollution and fluorescence fingerprint of riverine dissolved organic matter is season dependent[J]. Science of the Total Environment, 2022, 823: 153617.

[2] Maqbool T, Qin Y L, Ly Q V, et al. Exploring the relative changes in dissolved organic matter for assessing the water quality of full-scale drinking water treatment plants using a fluorescence ratio approach[J]. Water Research, 2020, 183: 116125.

[3] Xu R Z, Cao J S, Feng G Y, et al. Fast identification of fluorescent components in three-dimensional excitation-emission matrix fluorescence spectra via deep learning[J]. Chemical Engineering Journal, 2022, 430: 132893.

[4] Song F H, Wu F C, Feng W Y, et al. Fluorescence regional integration and differential fluorescence spectroscopy for analysis of structural characteristics and proton binding properties of fulvic acid sub-fractions[J]. Journal of Environmental Sciences, 2018, 74: 116-125.

[5] Wang X P, Zhang F, Kung H T, et al. Evaluation and estimation of surface water quality in an arid region based on EEM-PARAFAC and 3D fluorescence spectral index: a case study of the Ebinur Lake Watershed, China[J]. CATENA, 2017, 155: 62-74.

[6] 丁志群，王金霞，赵洪霞，等．基于三维荧光光谱技术的食用油快速分析研究[J]. 光子学报，2015，44(6): 0630004.
Ding Z Q, Wang J X, Zhao H X, et al. Rapid analysing edible oil using three dimensional fluorescence spectroscopy[J]. Acta Photonica Sinica, 2015, 44(6): 0630004.

[7] 王书涛，展书杰，刘诗瑜，等．三维荧光光谱结合ICSO-SVM对性激素的分类鉴别[J]. 光学学报，2021，41(10): 1030004.
Wang S T, Zhan S J, Liu S Y, et al. Classification and identification of sex hormones by three-dimensional fluorescence spectroscopy combined with ICSO-SVM[J]. Acta Optica Sinica, 2021, 41(10): 1030004.

[8] 陈晓玉，杜雅欣，刘亚茹，孔德明．三维荧光光谱结合2DPCA-SSA-GRNN对柴油占比的检测[J]. 中国激光，2022，49(18): 1811002.
Chen X Y, Du Y X, Liu Y R, et al. Detection of diesel proportion using three-dimensional fluorescence spectrum and 2DPCA-SSA-GRN[J]. Chinese Journal of Lasers, 2022, 49(18): 1811002.

[9] 吴鹏，倪敬书，洪海鸥，等．基于荧光光谱法的皮肤胆固醇快速无创检测技术[J]. 中国激光，2021，48(3): 0307002.
Wu P, Ni J S, Hong H O, et al. Rapid non-invasive technology for skin cholesterol detection based on fluorescent spectrometry [J]. Chinese Journal of Lasers, 2021, 48(3): 0307002.

[10] 张洋，何腾超，刘林，等．基于离散三维荧光光谱的糖尿病识别方法研究[J]. 光学学报，2022，42(1): 0117002.
Zhang Y, He T C, Liu L, et al. Diabetes recognition method based on discrete three-dimensional fluorescence spectrum[J]. Acta Optica Sinica, 2022, 42(1): 0117002.

[11] 刘传旸，柴一荻，徐宪根，等．南方某河水质荧光指纹特征及污染溯源[J]. 光谱学与光谱分析，2021，41(7): 2142-2147.
Liu C Y, Chai Y D, Xu X G, et al. Aqueous fluorescence fingerprint characteristics and discharge source identification of a river in Southern China[J]. Spectroscopy and Spectral Analysis, 2021, 41(7): 2142-2147.

[12] Li L, Wang Y, Zhang W J, et al. New advances in fluorescence excitation-emission matrix spectroscopy for the characterization of dissolved organic matter in drinking water treatment: a review [J]. Chemical Engineering Journal, 2020, 381: 122676.

[13] Smith J T, Yao R Y, Sinsuebphon N, et al. Fast fit-free analysis of fluorescence lifetime imaging via deep learning[J]. Proceedings of the National Academy of Sciences of the United States of America, 2019, 116(48): 24019-24030.

[14] Tolkach Y, Dohmgörgen T, Toma M, et al. High-accuracy prostate cancer pathology using deep learning[J]. Nature Machine Intelligence, 2020, 2(7): 411-418.

[15] He R, Wu X, Sun Z N, et al. Wasserstein CNN: learning invariant features for NIR-VIS face recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(7): 1761-1773.

[16] Wu X J, Zhao Z L, Tian R L, et al. Identification and quantification of counterfeit sesame oil by 3D fluorescence spectroscopy and convolutional neural network[J]. Food Chemistry, 2020, 311: 125882.

[17] Ruan K, Zhao S, Jiang X Q, et al. A 3D fluorescence classification and component prediction method based on VGG convolutional neural network and PARAFAC analysis method [J]. Applied Sciences, 2022, 12(10): 4886.

[18] Nagrath P, Jain R, Madan A, et al. SSDMNV2: a real time DNN-based face mask detection system using single shot multibox detector and MobileNetV2[J]. Sustainable Cities and Society, 2021, 66: 102692.

[19] Iglovikov V, Shvets A. TernausNet: U-Net with VGG11 encoder pre-trained on ImageNet for image segmentation [EB/OL]. (2018-01-17) [2022-03-06]. https://arxiv. org/abs/1801.05746.

[20] Tang J M, Liang S X, Sun H W, et al. Analysis of dissolved organic matters in Fu River of Baoding using three dimensional fluorescence excitation-emission matrix[J]. Spectroscopy and Spectral Analysis, 2014, 34(2): 450-454.

[21] Wünsch U J, Murphy K R, Stedmon C A. The one-sample PARAFAC approach reveals molecular size distributions of fluorescent components in dissolved organic matter[J]. Environmental Science & Technology, 2017, 51(20): 11900-11908.

[22] Murphy K R, Stedmon C A, Wenig P, et al. OpenFluor: an online spectral library of auto-fluorescence by organic compounds in the environment[J]. Analytical Methods, 2014, 6(3): 658-661.

[23] Lambert T, Bouillon S, Darchambeau F, et al. Shift in the chemical composition of dissolved organic matter in the Congo River network[J]. Biogeosciences, 2016, 13(18): 5405-5420.

[24] Gullian-Klanian M, Gold-Bouchot G, Delgadillo-Díaz M, et al. Effect of the use of *Bacillus* spp. on the characteristics of dissolved fluorescent organic matter and the phytochemical quality of *Stevia rebaudiana* grown in a recirculating aquaponic system[J]. Environmental Science and Pollution Research, 2021, 28(27): 36326-36343.

[25] Kothawala D N, Murphy K R, Stedmon C A, et al. Inner filter correction of dissolved organic matter fluorescence[J]. Limnology and Oceanography: Methods, 2013, 11(12): 616-630.

[26] Chai L W, Huang M K, Fan H, et al. Urbanization altered regional soil organic matter quantity and quality: insight from excitation emission matrix (EEM) and parallel factor analysis (PARAFAC)[J]. Chemosphere, 2019, 220: 249-258.

[27] Murphy K R, Hambly A, Singh S, et al. Organic matter fluorescence in municipal water recycling schemes: toward a unified PARAFAC model[J]. Environmental science & technology, 2011, 45(7): 2909-2916.

[28] Chen B F, Huang W, Ma S Z, et al. Characterization of chromophoric dissolved organic matter in the littoral zones of

eutrophic lakes Taihu and Hongze during the algal bloom season [J]. Water, 2018, 10(7): 861.

[29] Gao Z Y, Guéguen C. Size distribution of absorbing and fluorescing DOM in Beaufort Sea, Canada Basin[J]. Deep Sea Research Part I: Oceanographic Research Papers, 2017, 121: 30-37.

[30] Dainard P G, Guéguen C, Yamamoto-Kawai M, et al. Interannual variability in the absorption and fluorescence characteristics of dissolved organic matter in the Canada Basin polar mixed waters[J]. Journal of Geophysical Research: Oceans, 2019, 124(7): 5258-5269.

[31] Wauthy M, Rautio M, Christoffersen K S, et al. Increasing dominance of terrigenous organic matter in circumpolar freshwaters due to permafrost thaw[J]. Limnology and Oceanography Letters, 2018, 3(3): 186-198.

[32] Pitta E, Zeri C. The impact of combining data sets of fluorescence excitation-emission matrices of dissolved organic matter from various aquatic sources on the information retrieved by PARAFAC modeling[J]. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 2021, 258: 119800.

[33] Jutaporn P, Armstrong M D, Coronell O. Assessment of C-DBP and N-DBP formation potential and its reduction by MIEX® DOC and MIEX® GOLD resins using fluorescence spectroscopy and parallel factor analysis[J]. Water Research, 2020, 172: 115460.

[34] Ren W X, Wu X D, Ge X G, et al. Characteristics of dissolved organic matter in lakes with different eutrophic levels in southeastern Hubei Province, China[J]. Journal of Oceanology and Limnology, 2021, 39(4): 1256-1276.

[35] Kothawala D N, von Wachenfeldt E, Koehler B, et al. Selective loss and preservation of lake water dissolved organic matter fluorescence during long-term dark incubations[J]. Science of the Total Environment, 2012, 433: 238-246.

Water Sample Classification and Fluorescence Component Identification Based on Fluorescence Spectrum

Chen Qing[1], Tang Bin[1*], Miao Junfeng[1], Zhou Yan[3], Long Zourong[1**], Zhang Jinfu[1], Wang Jianxu[1], Zhou Mi[1], Ye Binqiang[1,2], Zhao Mingfu[1], Zhong Nianbing[1]

[1]*Chongqing Key Laboratory of Fiber Optic Sensor and Photodetector, Chongqing University of Technology, Chongqing* 400054, *China*;

[2]*School of Microelectronics and Communication Engineering, Chongqing University, Chongqing* 400044, *China*;

[3]*Tongliang District Environmental Protection Bureau of Chongqing, Chongqing* 402560, *China*

Abstract

Objective The treatment of organic pollutants in surface water, drinking water, and wastewater is one of the urgent social problems to be solved in the development of human society. Three-dimensional excitation-emission matrix (3D-EEM) fluorescence spectroscopy technology has been widely used to detect fluorescence components in surface water, sewage, and other samples. There are a lot of interference noises and fluorescence overlap information in the original 3D-EEM data, so there is an urgent need for a fast and accurate method to extract and analyze the useful information in 3D-EEM spectra. At present, parallel factor analysis (PARAFAC) is commonly used to decompose the overlapping fluorescence signals in 3D-EEM, but the analysis process of this method is complex, and the data set is strict, which greatly limits the on-line monitoring and analysis of samples. In this study, according to the results of PARAFAC, we propose a convolutional fast classification and recognition network model, which can quickly obtain water sample types, mass concentration grades, and fluorescent component maps by using only two convolutional neural network (CNN) models. As a result, it provides effective technical means for rapid detection of scenes such as surface water, drinking water, wastewater monitoring, and so on.

Methods In this study, a method of water sample classification and fluorescence component fitting based on MobileNetV2, VGG11 component fitting (CF-VGG11) CNN, and PARAFAC is proposed. The 3D-EEM data of four types of water samples including surface water (DB), treated industrial wastewater (FS), sewage treatment plant inlet and outlet water (WS), and rural drinking water (XCYY) are collected, and the multi-output classification model of different water samples and the prediction and fitting model of fluorescence component maps are established with the results of PARAFAC as labels. The prediction of types and components is completed in two steps. In the first step, the MobileNetV2 algorithm is used to predict and classify different water samples. The second step is to use the CF-VGG11 network to fit the fluorescence component map.

Results and Discussions The data sets of all kinds of water samples are analyzed by PARAFAC, and four fluorescence components are shown (Fig. 6). Then, the PARAFAC results are uploaded to the OpenFluor database to obtain possible substances of various types of fluorescence components in water samples (Table 2). The similarity comparison scores of all components are more than 95%. Combined with the PARAFAC results as network labels, the MobileNetV2 classification network and CF-VGG11 component fitting network obtain a classification accuracy of 95.83% and a component fitting accuracy of 98.11%, respectively (Table 3). In order to show that the trained model has good classification and fitting performance, a part of untrained 3D-EEM data is selected for the test, and the results show that MobileNetV2 and CF-VGG11 can classify and fit the 3D-EEM of water samples very well (Fig. 7), and MobileNetV2 and CF-VGG11 network models have certain advantages compared with PARAFAC in terms of time cost, data requirement, and analysis process (Table 4).

Conclusions In this study, a fast CNN classification and recognition algorithm based on fluorescence spectrum is proposed to predict the types and mass concentrations of different water samples, as well as the overlapping fluorescence components in 3D-EEM. This study relies on PARAFAC for preliminary data preparation and MobileNetV2 network for classification of water sample types and mass concentration grades, which can achieve water pollution traceability and exceedance warning, and the CF-VGG11 network is used to fit the fluorescence component map of water samples. The results show that the fast classification and identification network model based on the results of PARAFAC can quickly predict the types and mass concentration grades of water samples and fit their specific fluorescence components by inputting 3D-EEM data of a single water sample, and there is no need to repeat the complex PARAFAC. Therefore, this study provides certain theoretical support for detecting water pollution by three-dimensional fluorescence spectrometry and is of a certain practical significance.

Key words spectroscopy; three-dimensional excitation-emission matrix fluorescence spectroscopy; water pollution; classification; convolution neural network