计算机科学
COMPUTER SCIENCE

一种基于特征增强的场景文本检测算法

高楠, 张雷, 梁荣华, 陈朋, 付政

引用本文

高楠, 张雷, 梁荣华, 陈朋, 付政. 一种基于特征增强的场景文本检测算法[J]. 计算机科学, 2024, 51(6): 256-263.

GAO Nan, ZHANG Lei, LIANG Ronghua, CHEN Peng, FU Zheng. Scene Text Detection Algorithm Based on Feature Enhancement [J]. Computer Science, 2024, 51(6): 256-263.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于融合序列的远控木马流量检测模型

Remote Access Trojan Traffic Detection Based on Fusion Sequences

计算机科学, 2024, 51(6): 434-442. https://doi.org/10.11896/jsjkx.230400159

一种面向中文自动问答的注意力交互深度学习模型

Attentional Interaction-based Deep Learning Model for Chinese Question Answering

计算机科学, 2024, 51(6): 325-330. https://doi.org/10.11896/jsjkx.230300175

基于改进Swin Transformer的中心点目标检测算法

Center Point Target Detection Algorithm Based on Improved Swin Transformer

计算机科学, 2024, 51(6): 264-271. https://doi.org/10.11896/jsjkx.230300222

融合Transformer与多阶段学习框架的点云上采样网络

Point Cloud Upsampling Network Incorporating Transformer and Multi-stage Learning Framework

计算机科学, 2024, 51(6): 231-238. https://doi.org/10.11896/jsjkx.230300154

异质虹膜识别研究综述

Review of Heterogeneous Iris Recognition

计算机科学, 2024, 51(6): 186-197. https://doi.org/10.11896/jsjkx.231200175

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500230

一种基于特征增强的场景文本检测算法

高 楠 张 雷 梁荣华 陈 朋 付 政
浙江工业大学计算机科学与技术学院 杭州 310023

摘 要 针对自然场景下图像文本复杂背景、尺度多变等造成的漏检、误检问题,提出了一种基于特征增强的场景文本检测算法。在特征金字塔融合阶段,提出了双域注意力特征融合模块(Dual-domain Attention Feature Fusion Module,D2AAFM)。该模块能够更好地融合不同语义和尺度的特征图信息,从而提高文本信息的表征能力。同时,考虑到网络深层特征图在上采样融合过程中出现语义信息损失的问题,提出了多尺度空间感知模块(Multi-scale Spatial Perception Module,MSPM),通过扩大感受野来获取更大感受野的上下文信息,增强深层特征图的文本语义信息特征,从而有效地减少文本漏检、误检。为了评估所提算法的有效性,在公开数据集 ICDAR2015,CTW1500 以及 MSRA-TD500 上进行实验,所提方法综合指标 F 值分别达到了 82.8%,83.4%和 85.3%。实验结果表明,该算法在不同数据集上都具有良好的检测能力。

关键词:深度学习;场景文本检测;注意力机制;多尺度特征融合;空洞卷积
中图分类号 TP391

Scene Text Detection Algorithm Based on Feature Enhancement

GAO Nan,ZHANG Lei,LIANG Ronghua,CHEN Peng and FU Zheng
College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China

Abstract To address the problem of missed and false detection of image text in natural scenes due to complex backgrounds and variable scales,this paper proposes a text detection algorithm for scenes based on feature enhancement. In the feature pyramid fusion stage,a dual-domain attention feature fusion module(D2AAFM)is proposed,which can better fuse feature map information of different semantics and scales,thus improving the characterization ability of text information. At the same time,considering the problem of semantic information loss in the process of up-sampling and fusion of deeper feature maps of the network,the multi-scale spatial perception module(MSPM)is proposed to enhance the semantic features of text in higher-level feature maps by expanding the perceptual field to obtain contextual information of a larger perceptual field,thus effectively reduce the text of missed and false detection. In order to evaluate the effectiveness of the proposed algorithm,it is tested on the publicly available datasets ICDAR2015,CTW1500 and MSRA-TD500,and its overall index F-value reaches 82.8%,83.4% and 85.3%,respectively. The experimental results show that the algorithm has good detection capability on different datasets.

Keywords Deep learning,Scene text detection,Attention mechanisms,Multi-scale feature fusion,Dilated convolution

1 引言

自然场景中的文本包含着丰富的语义信息,对于场景理解有着重要的意义。因此,近年来场景文本检测技术受到了计算机视觉领域的广泛关注。然而,图像拍摄的视角、清晰度以及文本本身的尺度多变性以及分布的多样性,给自然场景文本检测带来了巨大的挑战。因此,如何准确快速地检测出场景图片中的任意形状文本成为当前重要的研究任务[1]。

目前,随着卷积神经网络(CNN)的快速发展和应用,许多优秀的场景文本检测算法被提出,并取得了良好的效果。然而,现有的基于 CNN 的文本检测算法通常使用特征金字塔(Feature Pyramid Network,FPN)提取多尺度特征信息,但是不同层级的文本特征之间存在语义差异,直接融合后无法充分表达特征信息;其次,深层特征图的语义信息虽然丰富,但在与下一级的特征融合前需要进行降维操作,会出现语义信息丢失的情况,影响最终的检测结果。基于上述已有文本检测方法中存在的问题,本文提出了一种基于特征增强的场景文本检测算法,该算法对自然场景下各类型的文本有着更好的检测能力。本文的主要贡献如下:

1)为了解决深层特征图上采样信息损失的问题,提出了

到稿日期:2023-05-31 返修日期:2023-10-25
基金项目:国家自然科学基金(61702456,62036009,U1909203);国家重点研发计划(2020YFB1707700)
The work was supported by the National Natural Science Foundation of China(61702456,62036009,U1909203) and National Key Research and Development Program of China(2020YFB1707700).
通信作者:高楠(gaonan@zjut.edu.cn)

多尺度空间感知模块(MSPM)。该模块在空洞空间金字塔模块(Atrous Spatial Pyramid Pooling，ASPP)基础上进行改进，并且引入了深度可分离卷积来减少模型计算量；通过扩大感受野，增强深层特征图语义信息，提高了网络的分类能力，减少了非文本像素造成的误检和小尺度文本实例的漏检。

2)为了解决不同语义、尺度的特征图直接融合导致的漏检、误检问题，提出了双域注意力特征融合模块(D2AFFM)，将通道注意力和空间注意力结合起来，更好地融合不同语义、不同尺度特征图的空间信息，抑制背景噪声，凸显文本特征，从而提高文本检测的准确率。

3)在多方向文本数据集 ICDAR2015、弯曲文本数据集 CTW1500 和多语言文本数据集 MSRA-TD500 上的实验证明了所提算法的合理性和有效性。

2　相关工作

随着专家学者们对深度学习的不断研究和创新，基于深度学习的场景文本检测算法蓬勃发展，检测的精度和速度都取得了很大的提升。这类算法大多是基于 CNN 框架的，根据检测网络的不同大致可以分为两种：一种是基于边界框回归的方法；另一种是基于像素分割的方法。这两种方法各有优劣，适用于不同的任务[2]。

基于边界框回归的文本检测方法通常受到目标检测方法的启发，如 Faster R-CNN[3]，YOLO[4]，SSD[5] 等。首先通过候选区域生成器产生候选文本区域，再使用卷积神经网络对这些区域进行分类和回归，得到文本和背景的概率和边界框信息。最后使用非极大值抑制方法和文本框筛选算法得到最终的文本框。Tian 等[6] 提出了文本区域连接网络(Connectionist Text Proposal Network，CTPN)，使用 RPN 生成多尺度文本候选框，再将这些候选框连接成完整的文本行或实例。该方法引入 BiLSTM 进行文本序列建模，能有效地解决长文本检测问题。Jiang 等[7] 提出了 R2CNN 网络，该方法借鉴了区域卷积神经网络(R-CNN)和 Faster R-CNN 的思想，在生成不同方向的对称文本矩形候选框的基础上，使用不同尺度池化操作进行特征融合，得到文本区域的分类和位置回归结果。该方法适用于检测任意方向的文本，且检测速度较快。Shi 等[8] 基于目标检测框架 SSD 提出了 SegLink 算法，通过添加 SegLink 连接线来实现文本行的检测，无需额外的文本分割或聚类步骤。算法通过检测局部片段并连接所有片段来检测任意长度的文本行。SegLink 算法能够有效地检测多方向和弯曲的文本行，但其速度和准确率受限于 SSD 模型的性能。Zhou 等[9] 借鉴 DenseBox 的架构和 U-Net 特性提出了 EAST 算法，它采用全卷积网络(FCN)生成多尺度特征图，对像素级的文本块进行预测，通过回归文本包围框的属性实现文本检测。Zhang 等[10] 提出的 LOMO 网络结构主要在 EAST 算法基础上额外增加了一个迭代优化模块(Iterative Refinement Module，IRM)和任意形状表达模块(Shape Expression Module，SEM)，首先通过 EAST 算法生成文本区域的四边形检测框，然后分别使用 IRM，SEM 模块优化检测框，同时重建更加精准的文本区域。Liao 等[11] 提出的 TextBxes 在 SSD 目标检测方法的基础上，修改 anchor 的长宽比和卷积内核的形状大小，能够较好地检测图像中具有极端尺度比的文本。TextBoxes＋＋[12] 在 TextBoxes 模型的基础上进行了改进，主要是增加了文本框的角度预测，以便准确地检测任意方向的文本。尽管基于回归的方法在检测上取得了很好的结果，但它仍然不适用于检测任意形状的文本。

基于像素级分割的检测方法本质上就是将文本框的检测转化为文本与背景的像素级别的语义分割问题，这类方法通常使用全卷积网络 FCN[13] 对场景图片进行特征提取和文本区域的分割。为了提高检测的准确性，通常使用多尺度特征融合的方式丰富特征图信息。接着，通过后处理算法对文本区域进行精细定位，以区分不同的文本实例，最终输出文本位置坐标和相关属性。Zhang 等[14] 首次采用了基于文本像素分类的检测方法，该方法使用一个全卷积神经网络来预测文本区域的分割图，利用 MSER 算法从候选区域中提取字符位置信息。最后，通过对字符进行后处理操作连接字符区域，生成最终的文本检测结果。Long 等[15] 提出了一种针对曲线文本的检测算法 TextSnake，该算法采用自适应曲线拟合方式，利用多个不同大小且带有方向的圆盘覆盖标注文本，预测圆盘的中心坐标、大小和方向，可以适应各种弯曲的文本形状，从而准确定位场景中的文本，提高文本检测的准确率。相比传统矩形框方法，TextSnake 能够更准确地检测弯曲文本位置信息。He 等[16] 提出了一种级联的卷积网络 CCTN，分别是粗略检测网络和精修检测网络。CCTN 通过先进行粗略分类，再进行精细分割的方式来检测场景文本。Xie 等[17] 提出的 STD 网络首先将场景图片像素划分为文本、非文本和边界 3 类，然后引入 FCN 网络对图片进行像素级分类来区分和检测不同文本。Deng 等[18] 提出的 PixelLink 通过预测像素点与相邻的 8 个像素点之间的连接关系，来区分不同文本。Wang 等[19] 提出了 SPCNet，该算法基于 Mask R-CNN 模型，提出了 text context module 和 rescore 机制，有效地抑制了错误样本的检测，可以灵活地检测各种形状的文本。Li 等[20] 提出了一种渐进式尺度扩展算法 PSENet，该网络预测文本区域内多个尺度的分割图，然后使用宽度优先搜索算法(Breadth First Search，BFS)逐级扩展为多尺度的文本区域。PAN 算法[21] 提取并融合不同层级的特征图，通过分割网络预测文本区域、核以及相似向量来适应任意方向的文本，提出像素聚合(Pixel Aggregation，PA)算法对文本实例进行重建。Liao 等提出了可微分二值化(Differentiable Binarization，DB)[22] 的文本检测算法，该算法在网络中嵌入二值化操作，通过分割网络获取概率图和阈值图，并将两者结合起来，使用启发式技术(如像素聚类)重构文本实例。尽管基于分割的算法在检测任意方向的文本实例上有很大优势，但存在难以区分相邻的文本实例的缺点，而且伴随着复杂的后处理过程，难以满足实时的性能要求。

3　本文算法介绍

3.1　整体算法框架

本文提出的场景文本检测算法整体框架结构如图 1 所示，可以分为 4 个部分，特征提取阶段、特征融合、预测部分、后处理部分。本文使用 ResNet50 作为特征提取层，对输入

图像进行多尺度特征提取，得到 4 个特征图 C_1-C_4，它们分别来自骨干网络 Conv2-Conv5 卷积层。为了降低网络模型的复杂度和计算量，对 4 个特征图进行了降维处理，将维度统一降到 256。经过降维处理后，C_1，C_2，C_3，C_4 的大小分别为 160×160×256，80×80×256，40×40×256 和 20×20×256。特征融合阶段包括两个主要模块，即多尺度空间感知模块和双域注意力特征融合模块。

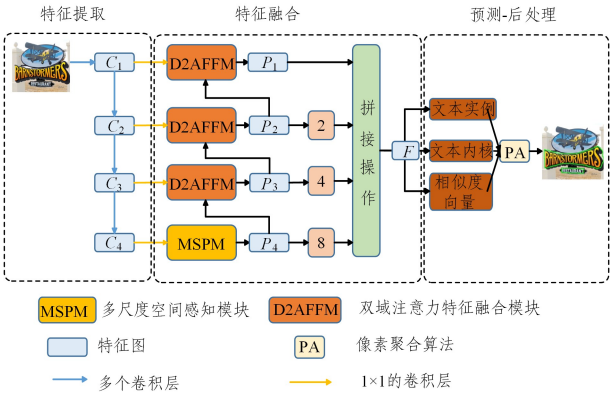


图 1 本文算法网络结构图

Fig. 1 Network structure of the proposed algorithm

首先，使用多尺度空间感知模块对最高层特征图 C_4 进行处理，以增强高层语义信息从而减少特征图上采样过程中可能出现的信息损失。该模块生成的特征图代替原始的 C_4 特征图参与后续的 FPN 特征融合过程。然后，使用引入双域注意力特征融合模块的 FPN 对提取的多尺度特征图进行融合，得到 P_1-P_4 多尺度特征图。接下来，对这些特征图进行相应大小的上采样操作，进行拼接得到最终的特征图 $\boldsymbol{F}$，便于后续的文本检测。最后，使用检测头进行预测得到文本实例、文本内核以及像素相似度向量，紧接着使用后处理方式(PA)进行文本实例的重建，得到最终的文本分割结果。

3.2 多尺度空间感知模块

通常情况下，骨干网络提取图像的多尺度特征图。浅层特征图包含更多的细节信息，深层特征图则包含更丰富的语义信息。在进行多尺度特征融合时，深层特征图往往会在上采样过程中丢失部分语义信息，影响像素的分类能力。因此，研究者在 DeepLab 系列网络中提出了空洞空间金字塔模块(Atrous Spatial Pyramid Pooling，ASPP)[23]，该模块使用不同大小的空洞卷积对输入特征图进行并行采样，以获得不同尺寸的感受野，增强深层特征图的语义信息。ASPP 模块通常由一个 1×1 卷积、3 个不同膨胀因子的 3×3 空洞卷积，以及一个 Pooling 操作组成，膨胀因子一般设置为 6，12，18。然而，ASPP 模块中多层复杂的空洞卷积叠加容易丢失图像中小尺度的细节信息，从而导致分割结果不准确。

为此，本文提出了一种改进的 ASPP 网络结构，称为多尺度空间感知模块，整体结构如图 2 所示。该模块通过密集连接的方式将原有的独立 ASPP 分支改为级联结构，以实现更密集的像素采样。在原有的 3 个并行的空洞卷积分支基础上，增加了一个串联结构，将输出较小的空洞卷积和主干网络的输出级联起来，再送入膨胀因子较大的空洞卷积，以获得更好的特征提取效果。同时，考虑到原有 ASPP 的膨胀因子较大，不利于提取高层语义信息，采用了以 2 的倍数为间隔的膨胀因子，并引入了一个 Image pooling 分支。该分支通过全局平均池化提取全局信息，作为空洞卷积提取特征的补充。这个输出特征不仅具有很大的感受野，而且还包含多尺度的上下文信息。在多尺度空间感知模块中，我们采用了深度可分离卷积的计算方式。这种计算方式既减少了计算量和参数量，又能够获取到更多的语义信息，从而提高了特征提取能力。

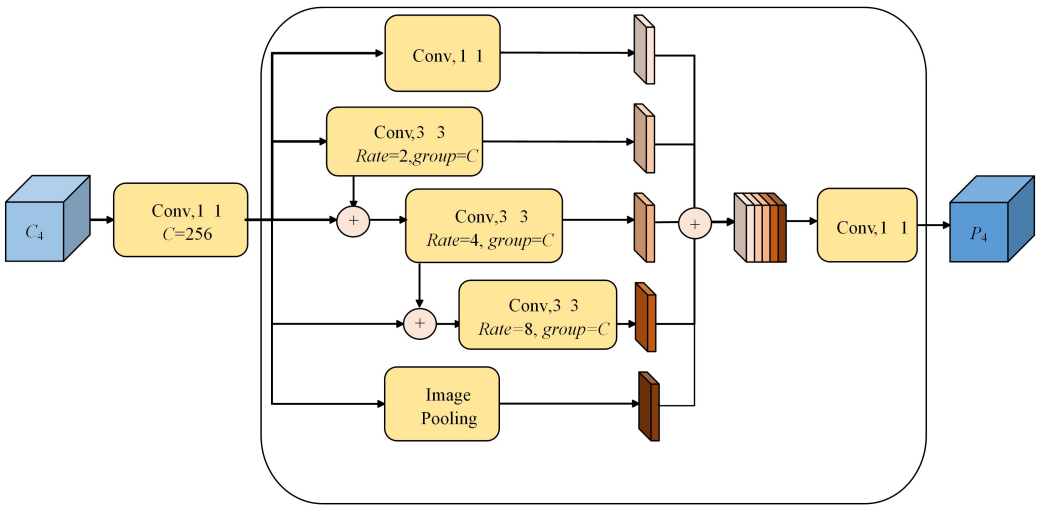


图 2 多尺度空间感知模块

Fig. 2 Multi-scale spatial perception module

3.3 双域注意力特征融合模块

本文提出了一种双域注意力特征融合模块，用于特征金字塔网络进行多尺度特征融合，增强文本显著性特征的感知，挖掘不同语义和尺度特征图之间的潜在联系，提升文本检测的准确度，如图 3 所示。具体实现如下：首先，对高层特征图 $\boldsymbol{P}_i(2\leqslant i\leqslant 4)$ 进行上采样操作得到与浅层特征图 $\boldsymbol{C}_{i-1}(2\leqslant i\leqslant 4)$ 相同尺寸的特征图 $\boldsymbol{P}_i'(2\leqslant i\leqslant 4)$，将两个特征图进行逐像素点相加操作，得到初步融合的特征图 $\boldsymbol{Z}$。之后，特征图 $\boldsymbol{Z}$ 分别经过通道域注意力模块和空间域注意力模块的处理，生成了注意力重新加权的特征图。这一过程旨在捕获不同通道

之间的依赖关系并学习图像中各个区域位置的权重信息，以获得更具判别性的特征图。最后，将其与原来输入的浅层特征图进行自适应加权相加得到融合最后的特征图 $\boldsymbol{P}_{i-1}$。

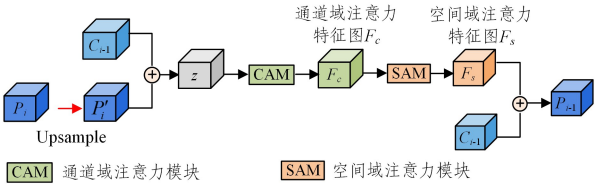


图 3　双域注意力特征融合模块

Fig. 3　Dual-domain attention feature fusion module

3.3.1　通道域注意力

通道域注意力模块如图 4 所示。为了提取更丰富的语义特征信息，首先针对特征图 $\boldsymbol{Z}$ 进行全局最大池化和平均池化操作，得到双重池化特征图 $\boldsymbol{F}_{\mathrm{GMP}}$ 和 $\boldsymbol{F}_{\mathrm{GAP}}$。接下来，将这两个池化特征图输入一个包含两个瓶颈层的多层感知机（MLP）中，其中通道数分别为 $\boldsymbol{C}/16$ 和 $\boldsymbol{C}$，得到双通道域特征图 $\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$。这一步的目的是使用缩放映射学习更多的特征信息，进一步提高特征表征的能力。然后将 $\boldsymbol{C}_1$ 和 $\boldsymbol{C}_2$ 沿通道维度进行元素相加并使用 Sigmoid 函数归一化，得到通道域维度方向的注意力权重向量 $\boldsymbol{M}_C$。最后将该权重向量通过逐元素相乘的方式分配给特征图 $\boldsymbol{Z}$，以此得到经过通道域注意力重新加权后的特征图 $\boldsymbol{F}_C$。此过程涉及的操作如下：

$$\boldsymbol{C}_1 = f(f'(F_{\mathrm{GMP}})) \tag{1}$$

$$\boldsymbol{C}_2 = f(f'(F_{\mathrm{GAP}})) \tag{2}$$

$$\boldsymbol{M}_c(F) = \mathrm{Sigmoid}(FC(add(\boldsymbol{C}_1, \boldsymbol{C}_2))) \tag{3}$$

$$\boldsymbol{F}_c = F \otimes \boldsymbol{M}_c(F) \tag{4}$$

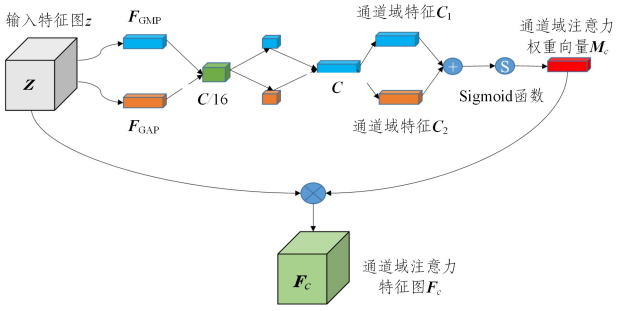


图 4　通道域注意力模块

Fig. 4　Channel domain attention module

3.3.2　空间域注意力

空间域注意力模块如图 5 所示，首先分别使用最大全局池化和平均池化处理输入特征图 $\boldsymbol{F}_c$，得到两个池化特征图 F'_{GMP} 和 F'_{GAP}。然后对两个池化特征图进行并行处理。其中一个分支采用沿通道维度拼接操作，另一分支将对应元素相加，分别得到双空间域特征 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$，将 $\boldsymbol{M}_1$ 和 $\boldsymbol{M}_2$ 沿通道维度进行 concat 操作，获得含有丰富空间域信息的特征图 $\boldsymbol{M}_3$。再使用 3×3 卷积对 $\boldsymbol{M}_3$ 进行特征聚合和通道缩减，得到最终的空间域特征向量 $\boldsymbol{M}$。接着将空间域特征向量 $\boldsymbol{M}$ 通过 Sigmoid 函数进行权重归一化，得到权重向量 $\boldsymbol{M}_S$，然后将该权重向量与输入特征图对应元素相乘，以此得到经过空间域注意力重新加权后的特征图 $\boldsymbol{F}_S$。

$$F(F') = Concat(F'_{\mathrm{GMP}}, F'_{\mathrm{GAP}}), F'_{\mathrm{GMP}} \oplus F'_{\mathrm{GAP}}) \tag{5}$$

$$\boldsymbol{M}_s(F) = \mathrm{Sigmoid}(Conv(F)) \tag{6}$$

$$F_s = F' \otimes \boldsymbol{M}_s(F) \tag{7}$$

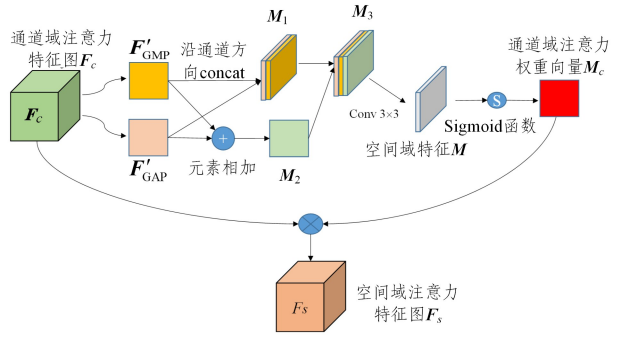


图 5　空间域注意力模块

Fig. 5　Spatial domain attention module

3.4　后处理方式

本文采用的后处理方式为像素聚合算法（PA），这类算法主要基于聚类的思想，通过网络预测得到含有文本的内核，并视其为聚类中心，将其他文本像素视为聚类对象，计算文本实例中文本像素与文本内核之间的距离从而进行文本实例的重建。像素聚合算法的步骤如下：

1）在预测网络中文本核的分割结果中找到每一个文本内核。

2）设置文本内核与像素的最小距离，对于每一个文本内核，遍历其相邻文本像素并计算它的相似度向量，如果小于阈值，则将该文本像素与相应的文本内核合并。

3）重复步骤 2），直到没有符合条件的邻近文本像素。

3.5　损失函数

文本检测的损失函数可以表示为：

$$L = \boldsymbol{L}_{\mathrm{tex}} + \alpha \boldsymbol{L}_{\mathrm{ker}} + \beta(\boldsymbol{L}_{\mathrm{agg}} + \boldsymbol{L}_{\mathrm{dis}}) \tag{8}$$

其中，$\boldsymbol{L}_{\mathrm{tex}}$ 和 $\boldsymbol{L}_{\mathrm{ker}}$ 分别为文本区域分割和文本内核分割的损失，α 和 β 是用来平衡损失函数的超参数，本文中默认设置为 0.5 和 0.25。为了解决文本和非文本像素不平衡问题，采用了 Dice loss 函数来优化分割文本区域 $\boldsymbol{P}_{\mathrm{tex}}$ 和 $\boldsymbol{P}_{\mathrm{ker}}$，其损失函数计算式如下：

$$\boldsymbol{L}_{\mathrm{tex}} = 1 - \frac{2\sum_i \boldsymbol{P}_{\mathrm{tex}}(i)\boldsymbol{G}_{\mathrm{tex}}(i)}{\sum_i \boldsymbol{P}_{\mathrm{tex}}(i)^2 + \sum_i \boldsymbol{G}_{\mathrm{tex}}(i)^2} \tag{9}$$

$$\boldsymbol{L}_{\mathrm{ker}} = 1 - \frac{2\sum_i \boldsymbol{P}_{\mathrm{ker}}(i)\boldsymbol{G}_{\mathrm{ker}}(i)}{\sum_i \boldsymbol{P}_{\mathrm{tex}}(i)^2 + \sum_i \boldsymbol{G}_{\mathrm{tex}}(i)^2} \tag{10}$$

其中，$\boldsymbol{P}_{\mathrm{tex}}(i)$ 和 $\boldsymbol{G}_{\mathrm{tex}}(i)$ 分别表示第 i 个像素的分割结果和文本区域的标注边框，$\boldsymbol{P}_{\mathrm{ker}}(i)$ 和 $\boldsymbol{G}_{\mathrm{ker}}(i)$ 分别表示第 i 个像素的预测值和文本内核的真值。本文采用在线困难样本挖掘方法平衡样本，计算 $\boldsymbol{L}_{\mathrm{tex}}$ 时忽略易识别的非文本像素，计算 $\boldsymbol{L}_{\mathrm{ker}}$，$\boldsymbol{L}_{\mathrm{agg}}$ 和 $\boldsymbol{L}_{\mathrm{dis}}$ 时只考虑文本像素。

$\boldsymbol{L}_{\mathrm{agg}}$ 损失衡量的是同一文本实例内像素点之间的距离是否在一定范围内，$\boldsymbol{L}_{\mathrm{dis}}$ 用于处理不同文本实例内核，以确保它们之间的距离不会太小。其计算式为：

$$\boldsymbol{L}_{\text{agg}}=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|T_i|}\sum_{p\in T_i}\ln(\boldsymbol{D}(p,\boldsymbol{K}_i)+1) \tag{11}$$

$$\boldsymbol{L}_{\text{dis}}=\frac{1}{N(N-1)}\sum_{i=1}^{N}\sum_{1j=1j\neq i}^{N}\ln(\boldsymbol{D}(\boldsymbol{K}_i,\boldsymbol{K}_j)+1) \tag{12}$$

其中，$\boldsymbol{N}$ 表示文本实例总数，$\boldsymbol{T}_i$ 表示第 i 个文本实例，$\boldsymbol{K}_i$ 表示第 i 个文本内核，$\boldsymbol{D}(p,\boldsymbol{L}_i)$ 代表文本实例 $\boldsymbol{T}_i$ 内的像素 p 与相对应文本内核 $\boldsymbol{K}_i$ 的距离，$\boldsymbol{D}(\boldsymbol{K}_i,\boldsymbol{K}_j)$ 代表核 $\boldsymbol{K}_i$ 与核 $\boldsymbol{K}_j$ 之间的距离。该公式计算每个文本实例的文本内核与其他文本内核之间的距离，并将这些距离进行了累加。

4 实验结果与分析

4.1 数据集和实验环境

ICDAR2015[24]：该数据集中的图像大多数是由谷歌眼镜在自然场景下拍摄的图片，图像中的大部分文本都非常小、模糊且被遮挡。数据集被划分为训练集和测试集，其中训练集1 000张图片，测试集500张图片，文本区域用矩形框标注。

CTW1500[25]：该数据集的图像来自于不同的场景，如街道、超市、商场等，其中一些图像还包含有遮挡、模糊和低对比度等复杂情况。数据集中的文本形式也非常多样，如水平、弯曲、垂直和多方向等。它包含1 500张图片，其中训练集1 000张图片，测试集500张图片，由14个点坐标的十四边形来表示曲线的文本实例。

MASD-TD500[26]：一个由微软亚洲研究院(MSRA)提供的多尺度、多语言数据集，该数据集包含500个图像，其中300个用于训练，200个用于测试。这些图像是从不同的场景中获取的，包括街道、商店、学校等。每个图像都包含一个或多个文本区域，每个区域都标注有文本的边界框和对应的文本字符串。

本文实验算法均在Ubuntu系统下进行，GPU型号为NVIDIA GeForce RTX 3090，实验环境为CUDA11.2+anaconda3+python3+Torch1.10.1。该模型使用在ImageNet数据集下的ResNet50的初始权重作为基础网络，随后使用各目标数据集的官方训练数据进行相应的训练。采用随机梯度下降算法SGD对网络进行训练，动量为0.9，权重衰减系数为5×10^{-4}，初始化学习率为0.001。网络训练中，每迭代200次学习率降为原来的十分之一，Batch size设置为8，训练次数设置为600。

4.2 消融实验

为了验证本文所提出模块的有效性，在ICDAR2015和CTW1500数据集上进行消融实验，结果如表1所列。

表1 本文算法在ICDAR2015和CTW1500数据集上的消融实验

Table 1 Ablation experiment of the proposed algorithm on ICDAR2015 and CTW1500 datasets

(%)

骨干网络(backone)	MSPM	D2AAFM	ICDAR2015			CTW1500		
			R	*P*	F_1	*R*	*P*	F_1
Resnet-50	×	×	76.3	82.9	79.5	77.5	84.6	80.9
Resnet-50	√	×	79.3	83.6	81.4	79.6	84.8	82.1
Resnet-50	×	√	78.7	**85.7**	82.0	78.6	**86.7**	82.5
Resnet-50	√	√	**80.5**	85.3	**82.8**	**80.6**	86.4	**83.4**

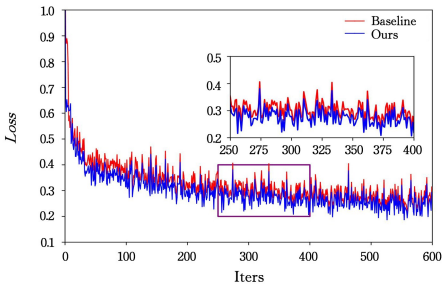
BaseLine：表1第1行为基线实验。

MSPM：如表1第2行所列，在添加本文所提出的MSPM后，其性能在数据集ICDAR2015和CTW1500上均有提升，其中召回率分别提升了3.0%和2.1%，F_1指标分别提升了1.9%和1.2%。实验结果表明增大感受野可以捕获更多的文本语义特征，有效地检测出更多不易检测的文本实例，提升文本检测的召回率。

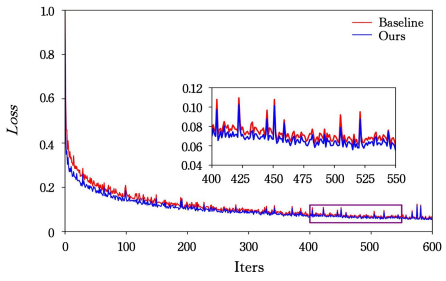
D2AAFM：如表1第3行所列，实验结果表明，在特征金字塔网络中引入双域注意力特征融合模块的本文算法，相比基线实验在准确率和召回率方面均取得了一定的提升。与基线对比，D2AAFM在数据集ICDAR2015上F_1指标提升2.5%，在数据集CTW1500上F_1指标提升1.6%。由此可以看出，双域注意力融合模块可以有效地融合不同尺度、语义的特征图，提升了文本信息的表征能力，从而减少了文本误检。

MSPM+D2AAFM：如表1第4行所示，实验结果表明，将多尺度空间感知模块和双域注意力特征融合模块联合，较好地平衡了准确度和召回率。与基线相比，在两个数据集上准确率、召回率以及F_1指标都有提升，其中F_1指标分别提升了3.3%和2.5%。

为了评估损失函数的收敛性，本文研究对比了基线模型和添加了MSPM以及D2AAFM模块的模型在公开数据集ICDAR2015和CTW1500上的训练损失函数。图6给出了模型在训练过程中随着iters的loss变化以及局部loss的放大情况，可以看出，相较于基准实验，本文提出的模型平均损失降幅更小。



(a)ICDAR2015



(b)CTW1500

图6 本文算法在两个数据集上的损失曲线

Fig. 6 Loss curves of the proposed algorithm on two datasets

为了进一步验证本文提出的双域注意力特征融合模块对特征融合效果的影响，在保持其他条件不变的情况下，分别使用空间注意力融合模块(SAFM)和通道注意力融合模块(CAFM)替换所提出的双域注意力特征融合模块，并对其

进行性能对比。实验结果如表 2 所列，可以看出，本文提出的双域注意力特征融合模块在文本检测性能方面有显著的优势。这是因为通道注意力融合模块和空间注意力融合模块分别通过建模通道间相关性和引入高级语义信息的方式，有效提高了特征的表征能力，从而进一步提升了文本检测的性能。

表 2　双域注意力与单域注意力的性能分析

Table 2　Performance analysis of dual-domain attention and single-domain attention

(%)

Attention	ICDAR2015			CTW1500		
	R	P	F_1	R	P	F_1
SAFM	79.6	84.7	82.1	79.8	85.4	82.5
CAFM	80.2	85.1	82.6	80.1	86.1	83.0
D2AAFM	**80.5**	**85.3**	**82.8**	**80.6**	**86.4**	**83.4**

4.3　对比实验

1）多方向场景文本检测

为了验证本文算法在多方向场景文本上的有效性，在多方向场景文本数据集 ICDAR2015 上进行了对比实验，同时与多个主流算法进行了对比。对比实验结果如表 3 所列，本文算法的准确率 P 值达到 85.3%，召回率 R 值为 80.5%，综合指标 F_1 值为 82.8%。相比 CTPN，SegLink，EAST，TextSnake 等场景文本算法，本文算法在各项指标上都有不同程度的提升。本文方法相比 PanF_1 指标提升了 2.5%，相比 Dbnet 准确度降低 1.5%，但召回率提升了 2.1%，整体的 F_1 提升了 0.5%。通过以上结果可以看出，本文算法在处理多方向的水平文本方面具有较好的性能。

表 3　不同算法在 ICDAR2015 数据集上的实验结果

Table 3　Experimental results of different algorithms on ICDAR2015 dataset

(%)

Methods	R	P	F_1
CTPN	51.6	74.2	60.9
PixelLink	81.7	82.9	82.3
EAST	73.5	83.6	78.2
SegLink	76.8	73.1	75.0
TextSnake	80.5	84.9	82.6
Psenet	79.7	81.5	80.6
Pan	77.8	82.9	80.3
Dbnet	78.4	**86.8**	82.3
Huang et al.	75.5	86.2	80.5
Lu et al.	**82.1**	83.2	82.7
ours	80.5	85.3	**82.8**

2）弯曲场景文本检测

为了验证算法在弯曲文本上的检测情况，本文在弯曲场景文本数据集 CTW1500 上进行了对比实验，验证了本文方法在弯曲文本检测上的优越性。如表 4 所列，本文算法准确率 P 值达到 86.4%，召回率 R 值为 80.6%，F_1 指标为 83.4%。可以看出，本文方法在准确度上取得了最好的结果。在 F_1 指标上，本文算法相比 Psenet 算法提升了 5.4%，相比 Dbnet 算法提升了 2.4%，表现出对于弯曲文本数据集很好的鲁棒性。

表 4　不同算法在 CTW1500 数据集上的实验结果

Table 4　Experimental results of different algorithms on CTW1500 dataset

(%)

Methods	R	P	F_1
CTPN	53.8	60.4	56.9
EAST	49.1	78.7	60.4
ABCNet	79.1	83.8	81.4
SegLink	40.0	42.3	41.1
TextSnake	**85.3**	67.9	75.6
Psenet	75.6	80.6	78.0
Pan	79.3	85.2	82.2
Dbnet	77.5	84.8	81.0
Shi	78.9	83.8	81.3
ours	80.6	**86.4**	**83.4**

3）多尺度场景文本检测

为了验证本文算法在多尺度文本上的检测能力，在多尺度场景文本数据集 MSRA-TD500 上进行了对比实验，如表 5 所列。在 MSRA-TD500 数据集上，本文算法准确率 P 值达到 86.2%，召回率 R 值为 84.5%，综合指标 F_1 值为 85.3%。相较于 SegLink 和 TextSnake，召回率分别提升 14.5% 和 10.6%；相较于 Zhang 等提出的算法准确率提升 4.8%；相较于 PAN 算法 F_1 指标提升 1.2%。由此可见，在多尺度场景下，本文提出的文本检测网络相比许多现有算法均取得了更好的准确率和召回率，证明了其在处理多尺度文本检测任务上的有效性。

表 5　不同算法在 MSRA-TD500 数据集上的实验结果

Table 5　Experimental results of different algorithms on MSRA-TD500 dataset

(%)

Mehtods	R	P	F_1
EAST	67.4	87.3	76.1
SegLink	70.0	86.0	77.0
TextSnake	73.9	83.2	78.1
PixelLink	73.2	83.0	77.8
Lyu et al.	76.2	87.6	81.5
Pan	83.8	84.4	84.1
Zhang et al.	75.3	81.4	78.2
ours	**84.5**	**86.2**	**85.3**

4.4　性能分析实验

为了验证本文所提算法的实时性，在 CTW1500 数据集上对不同的骨干网络以及不同尺寸输入图像进行探究。骨干网络选取轻量级网络 ResNet18 和 ResNet50，输入图像大小分别为 640，736，960，性能分析实验结果如表 6 所列，可以看出，本文算法在检测弯曲文本上有着优秀的性能。当 ResNet50 为特征提取网络时，输入图像尺寸为 960 像素时，F_1 值为 83.4%，达到最大，检测速度为 10.2 Frames/s。输入图像尺寸为 736 像素时，F_1 值从 83.4% 降至 83.1%，下降了 0.3%，但检测速度上升了 5.9 Frames/s。输入图像尺寸为 640 像素时，模型的 F_1 值降低到 82.1%，检测速度上升为 26.3 Frames/s。当 ResNet18 作为骨干特征提取网络时，在输入图像尺寸为 640 像素时，尽管 F_1 值降低到了 81.2%，但检测速度上升至 35.1 Frames/s。此实验表明当输入分辨率不高的文本图像时，本文算法可以满足实时检测文本的需求。

表 6 本文算法在 CTW1500 数据集上的性能分析

Table 6 Performance analysis of the proposed algorithm on CTW1500 data set

骨干网络	图像长边/像素	F_1/%	检测速度/(Frames·s^{-1})
Resnet50	960	83.4	10.2
	736	83.1	16.1
	640	82.1	26.3
Resnet18	640	81.2	35.1

4.5 实验结果展示

图 7 给出了本文提出的场景文本检测算法在多个数据集上的部分检测结果。从图中可以看出，本文提出的场景文本检测算法在多个数据集上表现出了更准确、更完整的文本检测效果，并且能够有效地抑制背景噪声，减少误检现象。这些结果说明了本文算法在文本检测任务上的优越性。

(a)ICDAR2015 (b)CTW 1500 (c)MSRA-TD 500

图 7 本文算法可视化结果示例

Fig. 7 Example of visualization results of the proposed algorithm

结束语 针对自然场景文本背景复杂、文本形状多变等因素造成的文本检测难题，本文提出了一种基于特征增强的场景文本检测算法。其中，多尺度空间感知模块对骨干网络高层语义信息进行处理，增强网络对文本特征信息的理解，从而更好地提高文本像素的分类能力，减少误检现象；另一方面，双域注意力特征融合模块使语义和尺度不一致的特征更精准地融合，提高了不同尺度文本的定位能力，减少了过度分割现象的产生，并提高了弯曲文本检测的准确率。通过在多个公开数据集上的实验及对比分析，证明了本文方法的有效性。后续工作会将文本信息引入网络当中，构建端到端文本识别网络，进一步提升检测的准确度以及进行相应的文本识别领域的探索。

参 考 文 献

[1] QIN Y, ZHANG Z. Summary of Scene Text Detection and Recognition[C]// Proceedings of IEEE Conference on Industrial Electronics and Applications. Kristiansand: IEEE, 2020: 85-89.

[2] CHEN M M, IBRAYI M, HAMDULL A. Research of Scene Text Detection Algorithms[C]// Proceedings of International Conference on Intelligent Robotics and Control Engineering. Tianjin: IEEE, 2022: 108-112.

[3] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-time Object Detection[C]// Proceedings of International Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 779-788.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C]// Proceedings of International Conference on Computer Vision and Pattern Recognition. Amsterdam: Springer, 2016: 21-37.

[6] TIAN Z, HUANG W, HE T, et al. Detecting Text in Natural Image with Connectionist Text Proposal Network[C]// Proceedings of European Conference on Computer Vision. Amsterdam: Springer, 2016: 56-72.

[7] JIANG Y, ZHU X, WANG X, et al. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection[C]// Proceedings of International Conference on Pattern Recognition. Vienna: IEEE, 2018: 3610-3615.

[8] SHI B, BAI X, BELONGIE S. Detecting Oriented Text in Natural Images by Linking Segments[C]// Proceedings of International Conference on Computer Vision and Pattern Recognition. HI: IEEE, 2017: 3482-3490.

[9] ZHOU X, YAO C, WEN H, et al. EAST: An Efficient and Accurate Scene Text Detector[C]// Proceedings of International Conference on Computer Vision and Pattern Recognition. HI: IEEE, 2017: 2642-2651.

[10] ZHANG C, LIANG B, HUANG Z, et al. Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes[C]// Proceedings of International Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2019: 10544-10553.

[11] LIAO M, SHI B, BAI X, et al. Textboxes: A Fast Text Detector with A Single Deep Neural Network[C]// Proceedings of the AAAI Conference on Artificial Intelligence. San Francisco: AAAI, 2017: 4161-4167.

[12] LIAO M, SHI B, BAI X. Textboxes++: A Sin-gle-shot Oriented Scene Text Detector[J]. IEEE Transactions on Image Processing, 2018, 27(8): 3676-3690.

[13] LONG J, SHELHAMER E, DARRELL T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 39(4): 640-651.

[14] ZHANG Z, ZHANG C, SHEN W, et al. Multi-oriented text detection with fully convolutional networks[C]// Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 4159-4167.

[15] LONG S, RUAN J, ZHANG W, et al. Textsnake: A Flexible Representation for Detecting Text of Arbi-trary Shapes[C]// Proceedings of European Conference on Computer Vision. Munich: AAAI, 2018: 20-36.

[16] HE T, HUANG W, QIAO Y, et al. Accurate Text Localization in Natural Image with Cascaded Convolutional Text Network [J]. arXiv: 1603.09423, 2016.

[17] XIE E, ZANG Y, SHAO S, et al. Scene Text Detection with Supervised Pyramid Context Network[C]// Proceedings of the

AAAI Conference on Artificial Intelligence. AAAI，2019：9038-9045.

[18] DENG D，LIU H，LI X，et al. Pixellink：Detecting Scene Text Via Instance Segmentation[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI，2018：20-36.

[19] WANG Q，GAO J，ZHANG M，et al. SPCNet：Scale Position Correlation Network for End-to-End Visual Tracking[C]// Proceedings of International Conference on Pattern Recognition. Beijing：IEEE，2018：1803-1808.

[20] WANG W，XIE E，LI X，et al. Shape Robust Text Detection with Progressive Scale Expansion Net-work[C]// Proceedings of International Conference on Computer Vision and Pattern Recognition. Long Beach：IEEE，2019：9328-9337.

[21] WANG W，XIE E，SONG X，et al. Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network [C]// Proceedings of International Conference on Computer Vision. Seoul：IEEE，2019：8440-8449.

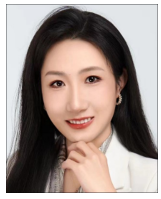
[22] LIAO M，WAN Z，YAO C，et al. Real-time Scene Text Detection with Differentiable Binarization[C]// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI，2020：11474-11481.

[23] CHEN L C，PAPANDREOU G，KOKKINOS I，et al. Deeplab：Semantic image segmentation with deep convolutional nets，atrous convolution，and fully connected crfs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017，40(4)：834-848.

[24] KARATZAS D，GOMEZ B L，NICOLAOU A，et al. Icdar 2015 Competition on Robust Reading[C]// Proceedings of International Conference on Document Analysis and Recognition. Tunis：IEEE，2015：1156-1160.

[25] LIU Y，JIN L，ZHANG S. Detecting Curve Text in the Wild：New Dataset and New Solution[J]. arXiv：1712. 02170，2017.

[26] YAO C，BAI X，LIU W Y，et al. Detecting Texts of Arbitrary Orientations in Natural Images[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Providence：IEEE，2012：1083-1090.



GAO Nan，Ph. D，born in 1983，is a member of CCF（No. 83932F）. Her main research interests include cross modal generation and retrieval，natural language processing，medical image processing，etc.

（责任编辑：何杨）

CCF 公益日@重庆丨走进企业赋能新兴互联网业态发展

2024 年 5 月 12 日上午，CCF 重庆会员活动中心在重庆西部科学城举行了 CCF 公益日分会场活动，CCF 重庆分部秘书长武春岭带领 CCF 会员数十人走进重庆互联网新兴企业“购免荟”，线上线下活动人数达 315 人。

本次活动是受重庆市女企业家协会秘书长孙总委托，为企业家白女士新投资互联网项目“购免荟”进行技术诊断。首先企业方技术人员用了 1 个小时介绍了项目整体设计和运营模式，CCF 重庆分部成员刘宇、王璐烽、岳守春、曹小平等教授专家经过现场提问和质询，对项目的技术方案提了很多建设性建议和解决方案，得到企业方高度赞扬。

本次活动现场，除了重庆市女企业家协会秘书长孙琳女士，还有 7 位企业家亲临现场，另外重庆市女企业家协会 280 余人通过视频会议线上参与了活动。最后 CCF 重庆分部秘书长武春岭教授进行总结，他认为这次活动不仅解决了企业发展中的技术问题，同时，也为高校人才培养提供了更多产教融合机会，校企双方均获益满满，希望今后多举办 CCF 走进企业公益活动，为地方产业发展助力。

据 CCF 微信公众号