智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

结合Segformer与增强特征金字塔的文本检测方法

张铭泉, 张泽恩, 曹锦纲, 邵绪强

引用本文:
张铭泉, 张泽恩, 曹锦纲, 邵绪强. 结合Segformer与增强特征金字塔的文本检测方法[J]. 智能系统学报, 2024, 19(5): 1111-1125.
ZHANG Mingquan, ZHANG Zeen, CAO Jingang, et al. Text detection method combining Segformer with an enhanced feature pyramid[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(5): 1111-1125.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202301013

DOI: 10.11992/tis.202301013
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240828.0924.004

结合 Segformer 与增强特征金字塔的文本检测方法

张铭泉[1,2], 张泽恩[1,2], 曹锦纲[1,2], 邵绪强[1,2]
（1. 华北电力大学 控制与计算机工程学院, 河北 保定 071003; 2. 华北电力大学 复杂能源系统智能计算教育部工程研究中心, 河北 保定 071003）

摘　要: 针对自然场景文本检测算法中的小尺度文本漏检、类文本像素误检以及边缘定位不准确的问题, 提出一种基于 Segformer 和增强特征金字塔的文本检测模型。该模型首先采用基于混合 Transformer (mix Transformer, MiT) 的编码器生成多尺度特征图; 然后, 在具有特征金字塔结构解码器的上采样部分, 提出级联融合注意力模块, 通过全局平均池化、全局最大池化和 Ghost 模块获取全局通道信息并保留文本特征; 接着, 在解码器的特征融合部分提出两级正交融合注意力模块, 利用非对称卷积分别从水平和垂直方向进行信息增强; 最后, 利用可微分二值化对结果进行后处理。将本文方法在 ICDAR2015、ShopSign1265 和 MTWI 3 个数据集上进行实验, 相比于其他 8 种方法, 本文方法的 F 值均为最优, 分别达到了 87.8%、59.1% 和 74.8%。结果表明, 本文方法有效提高了文本检测的准确率。

关键词: 文本检测; 特征金字塔; 注意力机制; Segformer; Ghost 模块; 多尺度特征融合; 平均池化; 最大池化
中图分类号: TP391.4　　**文献标志码:** A　　**文章编号:** 1673−4785(2024)05−1111−15

中文引用格式: 张铭泉, 张泽恩, 曹锦纲, 等. 结合 Segformer 与增强特征金字塔的文本检测方法 [J]. 智能系统学报, 2024, 19(5): 1111–1125.
英文引用格式: ZHANG Mingquan, ZHANG Zeen, CAO Jingang, et al. Text detection method combining Segformer with an enhanced feature pyramid[J]. CAAI transactions on intelligent systems, 2024, 19(5): 1111–1125.

Text detection method combining Segformer with an enhanced feature pyramid

ZHANG Mingquan[1,2], ZHANG Zeen[1,2], CAO Jingang[1,2], SHAO Xuqiang[1,2]
(1. School of Control and Computer Engineering, North China Electric Power University, Baoding 071003, China; 2. Engineering Research Center of intelligent Computing for Complex Energy Systems Ministry of Education, Baoding 071003, China)

Abstract: To address the issues of small-scale text omission, text-like pixel misdetection, and inaccurate edge localization in text detection algorithms for natural scenes, we propose a text detection model based on Segformer and an enhanced feature pyramid. First, the model employs an MiT-B2-based encoder to generate multiscale feature maps. Subsequently, during the upsampling phase of the decoder, a cascaded fusion attention module is introduced, which acquires global channel information and text features through global average pooling, global max pooling, and ghost convolution. Then, a two-level orthogonal fusion attention module utilizes asymmetric convolution to enhance the information in the feature fusion section horizontally and vertically. Finally, the results are post-processed using differentiable binarization. The experiments were conducted on the ICDAR2015, ShopSign1265, and MTWI datasets. Compared with the other eight methods, the proposed method achieved the highest F-values, reaching 87.8%, 59.1%, and 74.8%%, respectively. These results demonstrate that the method effectively improves the accuracy of text detection.

Keywords: text detection; enhanced feature pyramid; attention mechanism; Segformer; ghost convolution; multiscale feature fusion; average pooling; max pooling

近年来, 由于文本具有描述性和概括性的能力, 自然场景下的文本检测在图像理解[1]、视觉搜索[2] 和自动驾驶[3] 等领域具有广泛应用, 越来越

收稿日期: 2023−01−11.　**网络出版日期:** 2024−08−28.
基金项目: 中央高校基本科研业务费专项资金项目 (2021MS092); 河北省省级科技计划项目 (22310302D).
通信作者: 曹锦纲. E-mail: caojg168@126.com.

多的应用场景需要利用图像中的文本信息。虽然近年来文本检测技术取得了巨大的进展，但由于文本实例的不同尺度、不规则的形状和极端的纵横比，场景文本检测仍然具有挑战性[4]。

得益于深度学习技术的发展，目前主流的文本检测方法主要分为 2 类：基于候选区域的检测方法和基于分割的检测方法[5]。基于候选区域的检测方法大部分在以 SSD (single shot multibox detector)[6] 和 Faster R-CNN (faster region-based convolutional neural networks)[7] 为基础的目标检测算法上改进。R2CNN (rotational region convolutional neural networks)[8] 在 Faster R-CNN 的基础上产生不同方向的候选框，通过对各个方向的候选区域特征使用不同的池化尺寸进行特征融合，可以检测任意角度旋转的文本。TextBoxes[9] 则以 SSD 为基础，针对自然场景文本长宽比大的特性，设置了适应性的锚点和长条形的卷积核，局限性在于该方法只能检测水平方向的文本。随后，TextBoxes++[10] 提出用四边形或旋转矩形代替 TextBoxes 中的矩形来表示文本区域的思路，提升了对旋转文本检测的准确性，这类方法难以处理密集文本和任意方向的文本。

基于分割的检测方法首先利用像素级的语义分割将图像分为背景和文本区域 2 类，随后通过后处理算法得到精确的文本区域。He 等[11] 首次将文本像素分类预测用于自然场景文本检测任务当中，利用 MSER (maximally stable extremal regions) 检测算子在文本区域内提取候选字符，然后通过后处理操作连接字符区域生成文本行检测结果。虽然取得了一定效果，但是该方法对密集文本分割效果较差。PSENet (progressive scale expansion network)[12] 的主体是 ResNet (residual network)[13] 和 FPN (feature pyramid networks)[14] 的结构，对每个文本实例采用不同尺度的内核，并逐步扩展最小尺度的内核来重构单个文本实例，能准确分割相邻文本，但该方法的后处理很复杂，模型的前向预测效率比较低。PAN (pixel aggregation network)[15] 算法设计了轻量化的特征提取核融合的网络，除了预测文本区域和文本核之外，还引入了像素相似向量，使后处理方式变成可学习的，检测效率得到提高。为了进一步简化后处理过程，DBNet[16] 使用近似的可微分二值化 (differentiable binarization, DB) 算法代替固定阈值算法，并通过 Vatti 剪切算法缩小标注边界得到概率图，接着利用概率映射恢复文本实例的分离边界，该方法实现了对任意形状的文本检测，但对相邻特征图直接融合会出现空间信息减少，造成文本边缘定位不准确的情况。文献 [17] 针对现有算法未充分利用高层语义信息和空间信息，提出了一种基于增强特征金字塔网络的场景文本检测算法，在多种数据集上取得了较好的效果。文献 [18] 设计了图像级上下文信息模块以捕获全局图像信息和语义级上下文信息模块以学习目标区域信息，两者信息融合增强网络特征信息保证检测的准确性。文献 [19] 提出一种注意力监督策略下的文本检测算法，利用注意力掩膜监督生成下一级特征图，最后处理优化后得到最终的文本检测结果。

基于分割的文本检测模型通常包括骨干网络、特征融合模块和后处理 3 部分，通过对图像进行像素级预测，在一定程度上提高了对任意方向文本检测的准确度。但仍存在以下问题：1) 使用 ResNet 等卷积神经网络作为骨干网络，忽视了全局特征，存在语义信息丢失、特征提取不充分的问题，限制了整体网络的文本边界定位能力。2) 使用特征金字塔融合相邻特征图时，由于低尺度特征图的上采样操作会引入噪声，直接拼接会使空间特征丢失，容易出现小尺度文本漏检和背景像素误检的情况。3) 在对特征金字塔部分得到的多张特征图按通道拼接时，不同深度的文本特征之间存在差异，直接融合后会出现背景信息大量冗余导致局部信息缺失的问题，造成相邻文本区域边界定位不准确。

为了解决骨干网络特征提取不充分、上采样操作导致空间特征丢失以及特征融合时背景信息冗余的问题，本文采用基于分割的检测思想，提出了一种基于 Segformer 的端到端文本检测模型。该模型主要由编码器、解码器和后处理 3 部分组成，首先，引入 Segformer 中的 MiT-B2 作为编码器的主体部分，一方面，利用多头自注意力可以解决在卷积神经网络中图像的全局语义信息利用不充分的问题。其次，将编码器输出的 4 个不同尺度特征图输入到以增强特征金字塔网络 (enhance feature pyramid networks, EFPN) 为主体的解码器中进行特征融合。针对连续上采样操作引入新的噪声导致融合之后的文本信息丢失的问题，本文在各层之间引入级联融合注意力 (cascading fusion attention, CFA) 模块，以强化相邻层级特征之间的空间关联性，提高文本检测精度。为了高效融合 CFA 产生的 4 种特征图，解决特征融合时局部信息缺失的问题，先采用双线性插值方法统一尺寸，再采用两级正交融合注意力 (dual-

orthogonal fusion attention, D-OFA) 模块在水平和垂直方向分别提取不同方向的特征信息, 具体来说, 第 1 级分别融合 2 个低级特征和 2 个高级特征生成空间信息图和语义信息图, 第 2 级则融合空间信息图和语义信息图得到最终的掩码图。最后, 经过可微分二值化处理得到预测的结果。

本文的贡献如下:

1) 提出一种由基于 MiT-B2 的编码器、基于增强特征金字塔的解码器和可微分二值化模块组成的文本检测模型, 有效解决了文本边缘定位不准确、小尺度文本漏检和背景像素误检的问题。

2) 设计了级联融合注意力模块和两级正交融合注意力模块, 以减少上采样过程中的信息丢失和突出文本空间特征。

3) 在公开数据集 ICDAR2015、ShopSign1265 和 MTWI 上与 8 种模型进行了对比, 取得了最好的结果, F 值分别为 87.8%、59.1% 和 74.8%。

1　相关工作

1.1　基于 Transformer 的网络模型

视觉 Transformer (vision Transformer,ViT)[20] 首次将在自然语言处理领域表现良好的 Transformer 应用到计算机视觉领域。ViT 将每幅图像视为一系列的图像块, 然后将它们输入多个 Transformer 层以进行分类。ViT 主要由 3 部分组成: 位置编码器用于将位置信息嵌入到补丁块作为输入; 编码器用于提取样本的语义信息和空间信息; 解码器用来输出分类的结果。虽然 ViT 在大规模数据集上取得了巨大成功, 但由于 ViT 将图像切块并展平, 破坏了内部的空间信息, 导致训练过程收敛慢。文献 [21] 用条件位置编码代替了 ViT 中的预定义位置嵌入, 使 Transformer 能够处理任意大小的输入图像而无需插值。文献 [22] 设计了一种基于结构嵌套的 Transformer 架构, 通过内外 2 个 Transformer 联合, 提取图像局部和全局的特征, 提高了模型的识别效果。Swin Transformer[23] 充分利用窗口的设计, 将卷积神经网络的局部性引入 Transformer, 以加强窗口内图像块之间的信息交互, 减少计算量。文献 [24] 采用金字塔结构结合不同尺度的特征图, 从计算量和复杂度方面考虑, 对多头注意力机制作了一定的改进, 通过简单地堆叠多个独立的 Transformer 编码块, 在目标检测和语义分割方面比 ResNet 有相当大的提升。

以上模型均在 Transformer 基础上进行了相应的提升和改进, 但仍有不足。首先, ViT 的局限性在于处理图片时无法保留空间信息, 而且模型的深度不够, 不能像 CNN 一样提取深度特征, 只能生成单一的特征图; 其次, Swin Transformer 和 TNT (Transformer in Transformer)[22] 均对编码器作出了改进, 有利于处理图像的局部特征与全局特征的关系, 但处理分辨率高的图片时, 计算量仍很大。针对以上不足, Segformer[25] 作出了相应的解决: 1) 利用 4 个 Transformer 块, 可以得到 4 种尺度的特征图, 并在编码器中去掉了位置嵌入, 可以适应任意的测试分辨率, 避免了测试图像与训练图像尺寸不同而导致模型性能下降的问题。2) 采用高效的自注意力机制, 引入放缩系数, 通过与全连接层结合, 在不损失图像信息的情况下, 降低算法复杂度。

1.2　可微分二值化

基于分割的文本检测方法可以更直观地描述任意形状的文本区域, 是因为在后处理部分对分割操作得到的结果进行了二值化操作。二值化操作是指对概率图的逐个像素通过二值化方法转换成 0 或 1 的二值映射, 0 为背景, 1 为文本区域, 以便于将文本区域与背景分开。

传统的二值化通过设置固定阈值进行文本区域与背景区域的划分, 如果像素值大于阈值则为 1, 小于阈值则为 0。该固定阈值是根据经验人工设定, 阈值是否合适直接影响到最终结果的好坏, 而且该方法不适用于具有复杂背景的场景图像。

为了解决固定阈值的局限性, Liao 等[16] 提出一种可微分二值化算法, 可以将阈值设置成可学习的参数并与网络模型一起训练, 从而自适应地确定不同位置的阈值。一般来说, 文本中心区域阈值大, 边缘区域阈值小, 设置自适应阈值可以提高对任意形状文本检测的精度。可微分二值化不仅简化了后处理过程, 而且还能在复杂场景中分离出文本区域, 提高文本检测的性能。

2　本文模型

图 1 是本文提出模型的总体架构, 提出模型由 MiT-B2 组成的编码器、EFPN 组成的解码器和基于可微分二值化算法的后处理模块 3 部分构成。对于编码器, 输入为原始的文本图像, 经过 MiT-B2 处理后, 生成不同通道、不同尺度的 4 张特征图 $\boldsymbol{M}_1$、$\boldsymbol{M}_2$、$\boldsymbol{M}_3$、$\boldsymbol{M}_4$; 对于解码器, 首先将编码器的输出经过 1×1 的卷积将通道数统一后, 经过级联融合注意力模型结合相邻尺度的特征图生成 $\boldsymbol{F}_1$、$\boldsymbol{F}_2$、$\boldsymbol{F}_3$、$\boldsymbol{F}_4$, 接着将其分别上采样至原图 1/4 大小得到 $\boldsymbol{D}_1$、$\boldsymbol{D}_2$、$\boldsymbol{D}_3$、$\boldsymbol{D}_4$, 最后, 使用两级正交融合注意力模块融合 4 张特征图并输出解码器的

结果 $\boldsymbol{C}_1$；对于后处理模块部分，首先要根据解码器的输出分别预测文本中心和区域边界位置得到概率图和阈值图，接着采用可微分二值化算法处理得到近似二值图，实现将文本区域将文本区域和背景划分开，最后将预测结果展示到原图中即得到最终的文本检测结果。

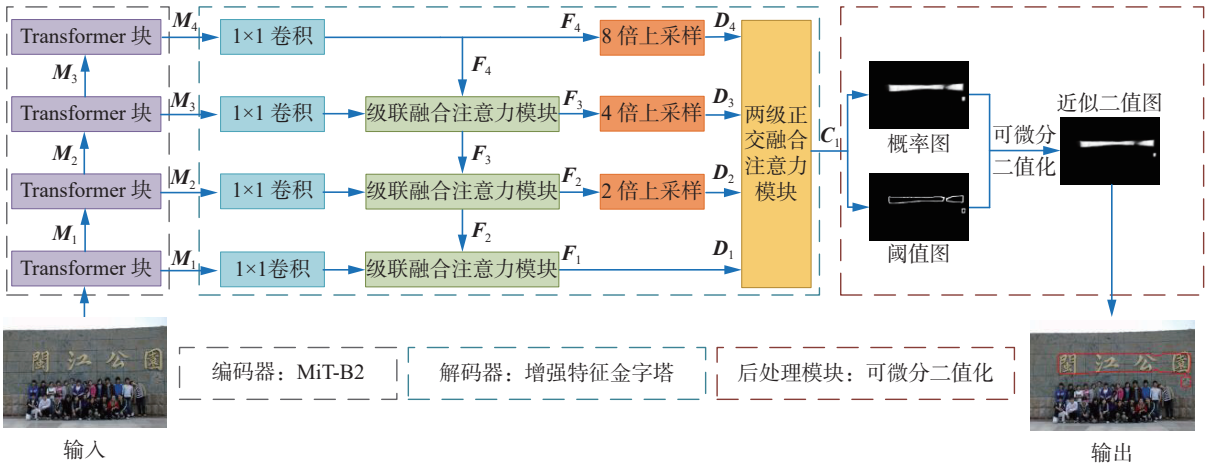


图 1　本文模型总体架构

Fig. 1　Overall architecture of our model

2.1　基于 MiT-B2 的编码器

与 ViT 只能产生单一分辨率的特征图不同，MiT-B2 在编码器的 4 个阶段生成了不同分辨率的特征图 $\boldsymbol{M}_1$、$\boldsymbol{M}_2$、$\boldsymbol{M}_3$、$\boldsymbol{M}_4$，分别为原图尺度的 1/4、1/8、1/16、1/32。利用低分辨率的特征图，确定文本中心区域的位置；利用高分辨率的特征图，得到文本区域的边缘信息。如图 2 所示，1 个 Transformer 块是由 N 个高效自注意力 (efficient self-attention) 模块和混合前馈网络 (mix-feed-forward network, Mix-FFN) 连接而成，之后使用 1 个重叠合并 (overlapped patch merging) 模块进行重叠的图像块合并，以保持这些图像块周围的局部连续性。

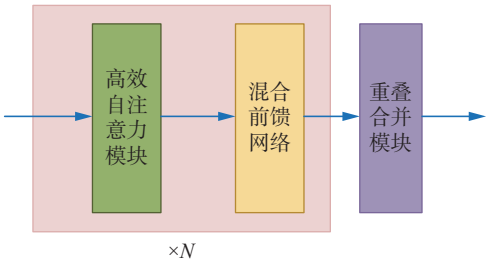


图 2　Transformer 块结构

Fig. 2　Structure of a Transformer block

高效自注意力模块　在原始的多头自注意中，每个头 $\boldsymbol{Q}$(query)、$\boldsymbol{K}$(key)、$\boldsymbol{V}$(value) 具有相同的维度 $N\times C$，其中 $N=H\times W$ 为图片的尺寸，C 为通道数，自注意机制定义为

$$A(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{head}}}}\right)\boldsymbol{V}$$

式中：d_{head} 表示每个头的通道数，Softmax(·) 表示 Softmax 激活函数。该过程的计算复杂度为 $O(N^2)$，在处理高分辨率的图片时，会产生相当大的计算量，为此，MiT-B2 引入一个缩放系数 a，在保证通道数相同的情况下，通过 a 倍下采样操作，将输入 $\boldsymbol{K}$ 的维度降低了 a 倍，从而使自注意机制的复杂度从 $O(N^2)$ 降低到 $O(N^2/a)$，具体过程可表示为

$$\hat{\boldsymbol{K}}=\mathrm{LayerNorm}(\mathrm{Conv}_{a\times a}(\boldsymbol{K}))$$

式中：$\boldsymbol{K}$ 是原始多头自注意力的输入，$\hat{\boldsymbol{K}}$是经过缩放变化后的输出，LayerNorm(·) 表示层标准化，$\mathrm{Conv}_{a\times a}(\cdot)$ 表示卷积核大小为 a、步幅为 a、填充为 0 的卷积操作。

混合前馈网络　由于 ViT 的位置编码分辨率是固定的，当测试图片的分辨率与训练图片的分辨率不同时，需要进行位置编码操作，这往往会导致模型精度下降。为此，MiT-B2 去掉了位置编码，并设计了 Mix-FFN，Mix-FFN 将 FFN 与 3×3 的卷积和多层感知机 (multilayer perceptron, MLP) 相结合。具体过程为：将相邻高效自注意力模块的结果作为输入，首先经过 1 个 MLP 进行降维处理，然后经过 3×3 的卷积操作进行特征提取，最后经过 GELU 激活和 MLP 处理恢复到输入维度，再与原输入特征相加得到输出结果，公式为

$$\boldsymbol{x}_{\mathrm{out}}=\mathrm{MLP}(\mathrm{GELU}(\mathrm{Conv}_{3\times3}(\mathrm{MLP}(\boldsymbol{x}_{\mathrm{in}}))))+\boldsymbol{x}_{\mathrm{in}}$$

式中：MLP(·) 表示多层感知机，GELU(·) 表示 GELU 激活函数，$\mathrm{Conv}_{3\times3}(\cdot)$ 表示卷积核大小为 3×3、步幅为 2、填充为 1 的卷积操作，$\boldsymbol{x}_{\mathrm{in}}$ 表示混合前馈网络的输入，$\boldsymbol{x}_{\mathrm{out}}$ 表示混合前馈网络的输出。

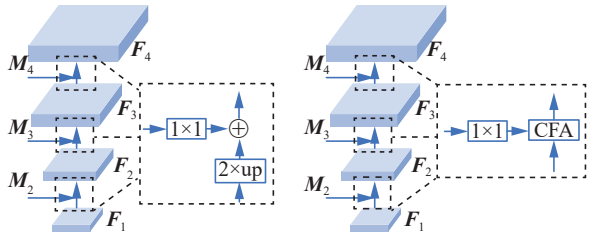
重叠合并模块　如果直接按照 ViT 中的融合

策略进行下采样操作, 将会损失块与块之间空间一致性。在 MiT-B2 中采用 7×7 和 3×3 的 2 种卷积, 进行下采样的同时也能学习到相邻图像块之间的位置关系。经过此模块的融合, 可以得到不同分辨率的特征图。

2.2　基于增强特征金字塔的编码器

2.2.1　级联融合注意力模块

由于编码器生成的不同层次特征图表达能力不同, 浅层特征主要反映空间细节信息, 深层特征主要反映整体语义信息。为了兼顾细节信息和整体特征, FPN[14] 通过上采样结构将深层特征与浅层特征逐级合并, 形成了由深到浅的层级结构。具体来说, 如图 3(a) 所示, 首先将低分辨率的特征图进行 2 倍上采样, 然后使用卷积核为 1×1 的卷积统一通道, 最后按元素对应位置相加, 得到新的特征图。虽然 FPN 有效处理了多尺度特征融合的问题, 但是采用最近邻插值后直接相加的融合方式有以下 2 点不足: 一方面由于 2 个特征图语义信息差距较大, 直接相加使得深层的语义信息不一定能有效传播; 另一方面, 最近邻插值后, 语义信息加强的同时会引入新的噪声, 导致细节信息丢失。



(a) 原始上采样结构　　(b) 改进后上采样结构

图 3　原始上采样结构与改进后上采样结构

Fig. 3　Original upsampling structure and modified upsampling structure

为了缓解上采样过程引入噪声带来的负面影响, 充分利用不同层级特征的细节信息和语义信息同时匹配 Segformer 的 4 级结构, 本文在特征金字塔的层与层之间引入一种级联融合注意力模块, 将低尺度特征图与对应的高尺度特征图进行融合, 如图 3(b) 所示。具体来说, 在特征金字塔的层与层之间添加了 3 个 CFA 模块, 分别使用 1×1 的卷积将 $\boldsymbol{M}_i(i=1, 2, 3)$ 的通道数统一为 256, 然后与 $\boldsymbol{F}_{i+1}(i=1, 2, 3)$ 经过 CFA 融合生成 $\boldsymbol{F}_i(i=2, 3, 4)$。CFA 可以强调全局特征, 同时也可以保留局部的小文本及文本边缘信息, 提高模型在极端尺度变化下的检测能力。

CFA 是由级联平均 (cascade average pooling, CAP) 模块、级联最大 (cascade max pooling, CMP) 模块和 Ghost 模块 (ghost module, GM)3 部分组成, 如图 4 所示。

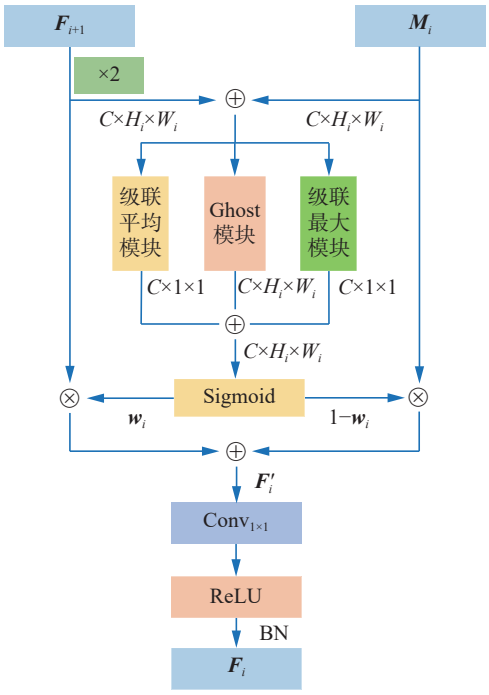


图 4　CFA 模块结构

Fig. 4　Structure of CFA module

CFA 的输入分别为 2 个高低尺度的特征图 $\boldsymbol{M}_i$ 和 $\boldsymbol{F}_{i+1}$, 输出为融合后的特征图 $\boldsymbol{F}_i$。CFA 处理过程如下:

1) 首先对 $\boldsymbol{F}_{i+1}$ 进行 2 倍上采样处理, 与 $\boldsymbol{M}_i$ 统一尺度。随后两者逐像素相加得到 $\boldsymbol{X}_i(i=1, 2, 3)$, 将 $\boldsymbol{X}_i$ 分别通过 CAP 和 CMP 提取全局特征, 同时通过 GM 提取局部特征, 以得到各通道的权重。最后将 3 部分的结果相加, 得到与特征图大小相同的融合权重 $\boldsymbol{w}_i$, 定义为

$$\boldsymbol{w}_i=\sigma(\mathrm{CAP}(\boldsymbol{X}_i)+\mathrm{CMP}(\boldsymbol{X}_i)+\mathrm{GM}(\boldsymbol{X}_i))$$

$$\boldsymbol{X}_i=\boldsymbol{M}_i+\mathrm{Up}_2(\boldsymbol{F}_{i+1}), i=1,2,3$$

式中: $\sigma(\cdot)$ 表示的是 Sigmoid 激活函数, CAP$(\cdot)$ 表示级联平均模块, CMP$(\cdot)$ 表示级联最大模块, GM$(\cdot)$ 表示 Ghost 卷积[26] 模块, $\boldsymbol{X}_i(i=1, 2, 3)$ 表示 3 个模块的共同输入, $\boldsymbol{w}_i(i=1, 2, 3)$ 表示的是经过 CAP、CMP 和 GM 处理得到的融合权重, $\mathrm{Up}_2(\cdot)$ 表示 2 倍上采样处理, $\boldsymbol{M}_i$ 和 $\boldsymbol{F}_{i+1}$ 为 CFA 的 2 个输入。

2) 接着使用第一步得到的融合权重 $\boldsymbol{w}_i$ 以元素级方式对 $\boldsymbol{M}_i$ 和 $\boldsymbol{F}_{i+1}$ 进行动态选择, 输出得到加权特征图 $\boldsymbol{F}'_i(i=1, 2, 3)$, 公式为

$$\boldsymbol{F}'_i=\mathrm{Up}_2(\boldsymbol{F}_{i+1})\times\boldsymbol{w}_i+\boldsymbol{M}_i\times(1-\boldsymbol{w}_i), i=1,2,3$$

式中: 融合权重 $\boldsymbol{w}_i$ 由 0 和 1 之间的实数组成, 通过与 $1-\boldsymbol{w}_i$ 结合使用, 使得网络能够在 $\boldsymbol{M}_i$ 和 $\boldsymbol{F}_{i+1}$ 之间进行加权平均。

3) 最后经过 1×1 的卷积处理得到融合特征

图 $\boldsymbol{F}_i$, 公式为

$$\boldsymbol{F}_i = \beta(\delta(\mathrm{Conv}_{1\times1}(\boldsymbol{F}'_i))), i = 1,2,3$$

式中: $\delta(\cdot)$ 表示 ReLU 激活函数, $\beta(\cdot)$ 表示批处理归一化 (batch normalization, BN), $\mathrm{Conv}_{1\times1}(\cdot)$ 表示进行 1×1 的卷积操作。

CAP 模块　CAP 模块的主要思想是利用全局平均池化以及通道间的交互作用获取文本图像的全局特征, 结构如图 5 所示。

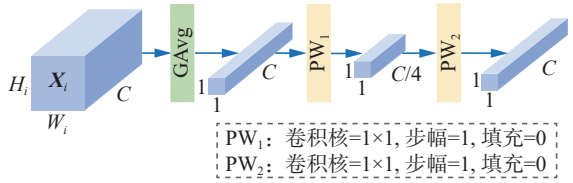


图 5　CAP 模块结构

Fig. 5　Structure of CAP module

CAP 处理过程如下: 输入 $\boldsymbol{X}_i \in \mathbf{R}^{C\times H_i\times W_i}$ 的图像先进行全局平均池化操作得到 C×1×1 的标量。为了尽可能保持模型的轻量级, 再采用卷积核大小为 1×1 的逐点卷积 (point wise convolution, PW-Conv) 将通道数压缩到原来的 1/4, 然后经过 ReLU 激活, 最后通过 1 个逐点卷积恢复到原来的维度, 定义为

$$\mathrm{CAP}(\boldsymbol{X}_i) = \beta(\mathrm{PW}_2(\delta(\beta(\mathrm{PW}_1(\mathrm{GAvg}(\boldsymbol{X}_i))))))$$

$$\mathrm{GAvg}(\boldsymbol{X}_i) = \frac{1}{H_i \times W_i}\sum_{m=1}^{H_i}\sum_{n=1}^{W_i}\boldsymbol{X}_{i[:,m,n]}$$

式中: GAvg(·) 为全局平均池化操作, $\mathrm{PW}_1(\cdot)$ 为逐点卷积组成的降维层, 通道数由 C 变为 C/4, $\mathrm{PW}_2(\cdot)$ 为逐点卷积组成的升维层, 通道数由 C/4 变为 C, 4 为通道衰减率, $\boldsymbol{X}_i$ 表示 CAP(·) 输入, H_i、W_i 分别表示特征图的长和宽, $\boldsymbol{X}_{i[:,m,n]}$ 表示特征图 $\boldsymbol{X}_i$ 每个通道上 (m, n) 位置处的像素值。

CMP 模块　CMP 与 CAP 结构类似, 用全局最大池化操作替换 CAP 中的全局平均池化, 以保留各通道最重要的信息, 其余结构不变, 定义为

$$\mathrm{CMP}(\boldsymbol{X}_i) = \beta(\mathrm{PW}_2(\delta(\beta(\mathrm{PW}_1(\mathrm{GMax}(\boldsymbol{X}_i))))))$$

$$\mathrm{GMax}(\boldsymbol{X}_i) = \max\{\boldsymbol{X}_{i[:,m,n]}\}$$

式中: GMax(·) 表示全局最大池化操作, $\max\{\boldsymbol{X}_{i[:,m,n]}\}$, $m \in [1,H_i]$, $n \in [1,W_i]$ 表示在 $\boldsymbol{X}_i$ 每个通道的 $H_i \times W_i$ 个像素中取最大值。

GM　仅在全局范围内聚合上下文信息会在一定程度上削弱小尺度文本的特征, 为了缓解这一问题, 本文引入 GM, 如图 6 所示, 采用 Ghost 卷积通过少量的标准卷积和线性运算保留了足够的局部特征。GM 与 CMP 和 CAP 并列连接, 保证了融合权重 $\boldsymbol{w}_i$ 在空间维度和通道维度与输入特征 $\boldsymbol{X}_i$ 的一致性。

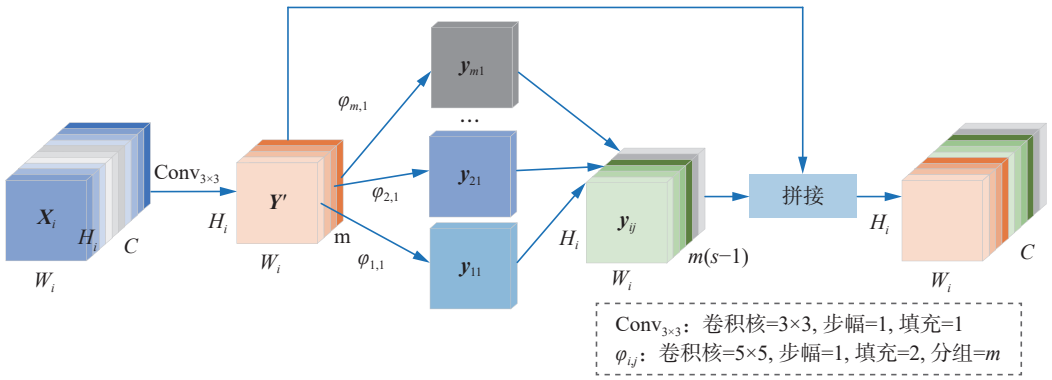


图 6　GM 结构

Fig. 6　Structure of GM

GM 处理过程如下:

1) 首先对于输入的特征图 $\boldsymbol{X}_i \in \mathbf{R}^{C\times H_i\times W_i}$ 进行卷积核大小为 3×3、步幅为 1、填充为 1 的卷积操作得到 $\boldsymbol{Y}' \in \mathbf{R}^{m\times H_i\times W_i}$, m=C/s, C 为输入通道数, s 为表示通道压缩的超参数, 默认设置为 2, 定义为

$$\boldsymbol{Y}' = \mathrm{Conv}_{3\times3}(\boldsymbol{X}_i)$$

2) 接着对 $\boldsymbol{Y}'$按通道进行线性变化, 得到特征图 $\boldsymbol{y}_{ij}$, $i \in [1, m]$, $j \in [1, s-1]$, 特征图总数为 $m(s-1)$, 公式为

$$\boldsymbol{y}_{ij} = \varphi_{i,j}(\boldsymbol{y}'_i), i \in [1,m], j \in [1, s-1] \qquad (1)$$

式中: $\boldsymbol{y}'_i$表示 $\boldsymbol{Y}'$第 i 个通道的特征图, $\varphi_{i,j}(\cdot)$ 表示对第 i 个特征图采用卷积核大小为 5×5、步幅为 1、填充为 2 的卷积进行第 j 次线性变换。线性操作的组数与特征图输入通道数相同, 均为 m, 每组输出 s−1 个 $H_i \times W_i$ 的特征图。

3) 最后将 $\boldsymbol{Y}'$与 $\boldsymbol{y}_{ij}$ 按通道拼接的结果作为 GM 模块的输出, 定义为

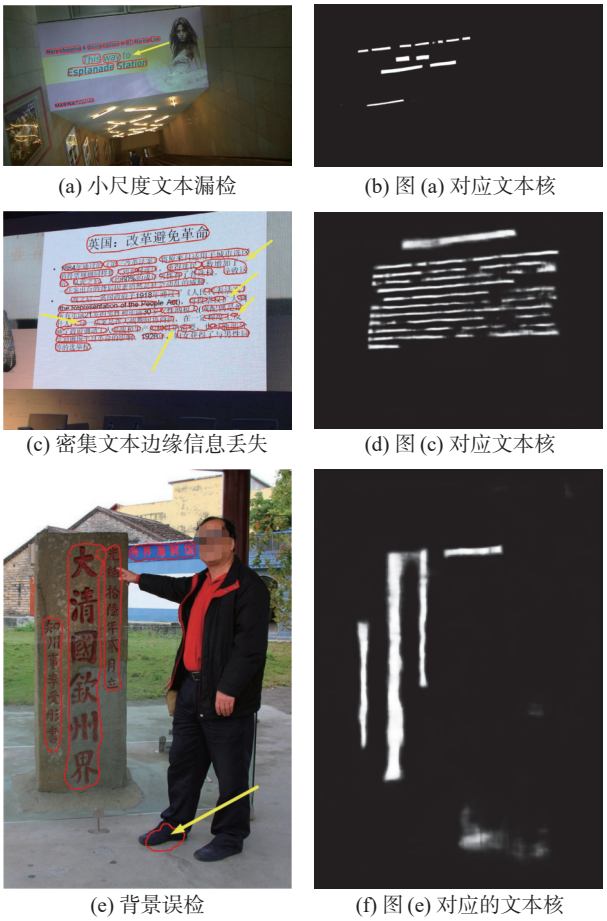
$$\mathrm{GM}(\boldsymbol{X}_i) = \mathrm{Concat}(\boldsymbol{Y}', [\boldsymbol{y}_{11}\ \boldsymbol{y}_{12}\ \cdots\ \boldsymbol{y}_{m(s-1)}])$$

式中: Concat(·) 表示按通道拼接操作, 所得特征图通道数为 C=m+$m(s$−1), 尺寸为 $H_i \times W_i$。

2.2.2　两级正交融合注意力模块

FPN 的上采样结构结合了深层特征与浅层特

征, 生成了 4 张不同尺度的特征图 $\boldsymbol{F}_i(i=1, 2, 3, 4)$。为了融合不同层次特征, FPN 首先对 $\boldsymbol{F}_i(i=1, 2, 3, 4)$ 分别进行上采样处理, 得到统一尺度的 4 层特征 $\boldsymbol{D}_i(i=1, 2, 3, 4)$, 随后直接将 $\boldsymbol{D}_i(i=1, 2, 3, 4)$ 按通道拼接, 再使用卷积进行特征融合并输出到指定维度。该融合方式存在 2 点不足: 一方面忽视了不同空间信息直接的关联, 不利于小尺度文本 (如图 7(a)) 和密集文本 (如图 7(c)) 的检测; 另一方面背景信息的迭代累积会产生大量的冗余信息, 不利于将背景与文本区域划分 (如图 7(e))。



(a) 小尺度文本漏检　(b) 图 (a) 对应文本核

(c) 密集文本边缘信息丢失　(d) 图 (c) 对应文本核

(e) 背景误检　(f) 图 (e) 对应的文本核

图 7　文本检测中存在的问题

Fig. 7　Problems in text detection

为了解决上述问题, 本文提出两级正交融合注意力模块, 结构如图 8 所示。一方面, 使用非对称卷积对文本区域进行水平和垂直建模; 另一方面, 只对同层次的相邻特征图进行融合, 达到不同层次的特征负责不同尺度的文本的目标。D-OFA 分为 2 层, 第 1 层的输入为 4 层特征 $\boldsymbol{D}_i(i=1, 2, 3, 4)$, 随后将 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$ 与 $\boldsymbol{D}_3$ 和 $\boldsymbol{D}_4$ 分为 2 组, 分别通过正交融合注意力模块生成 $\boldsymbol{E}_1$ 和 $\boldsymbol{E}_2$, 第 2 层则将 $\boldsymbol{E}_1$ 和 $\boldsymbol{E}_2$ 融合得到 $\boldsymbol{C}_1$, 具体过程如公式

$$\boldsymbol{E}_1 = \mathrm{OFA}(\boldsymbol{D}_1, \boldsymbol{D}_2)$$

$$\boldsymbol{E}_2 = \mathrm{OFA}(\boldsymbol{D}_3, \boldsymbol{D}_4)$$

$$\text{D-OFA}(\boldsymbol{D}_1, \boldsymbol{D}_2, \boldsymbol{D}_3, \boldsymbol{D}_4) = \boldsymbol{C}_1 = \mathrm{OFA}(\boldsymbol{E}_1, \boldsymbol{E}_2)$$

式中: OFA(·) 表示正交融合注意力模块, D-OFA(·) 表示两级正交融合注意力模块, $\boldsymbol{E}_1$ 和 $\boldsymbol{E}_2$ 为第 1 层的 2 个输出, 同时作为第 2 层的输入通过 OFA 模块得到 $\boldsymbol{C}_1$。虽然 $\boldsymbol{D}_i(i=1, 2, 3, 4)$ 的通道数与维度均相同, 但是 $\boldsymbol{D}_1$ 和 $\boldsymbol{D}_2$ 侧重文本的边缘信息, $\boldsymbol{D}_3$ 和 $\boldsymbol{D}_4$ 侧重文本所在的区域信息, 两两结合可以保留不同层次的特征。

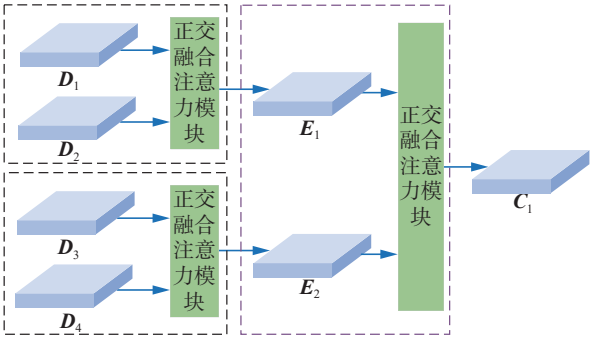


图 8　D-OFA 模块结构

Fig. 8　Structure of D-OFA module

OFA 模块　OFA 的主要思想为使用非对称卷积对文本区域进行水平和垂直处理, 以获取不同方向的空间信息, 并添加残差结构防止梯度消失或退化, 其结构如图 9 所示。

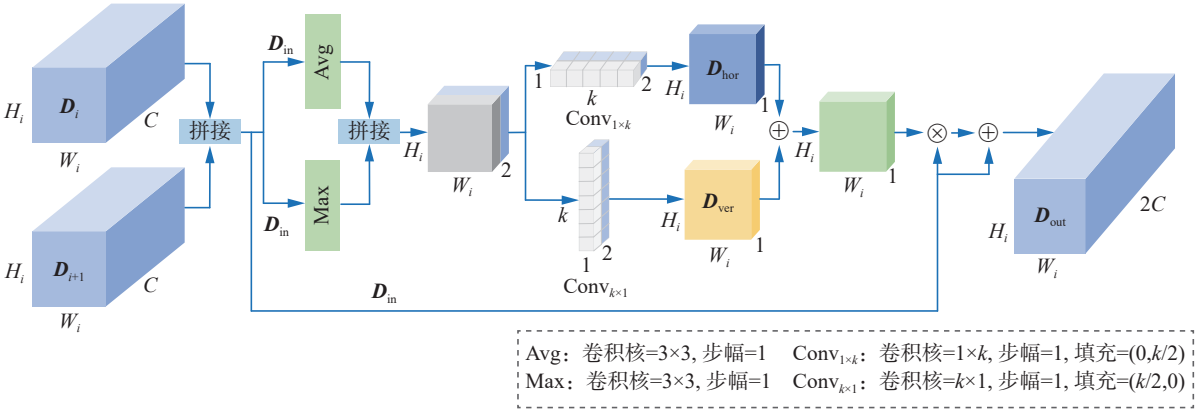


图 9　OFA 模块结构

Fig. 9　Structure of OFA module

OFA 处理具体过程如下：

1) 对 2 个输入 $\boldsymbol{D}_i \in \mathbf{R}^{C\times H_i\times W_i}$ 和 $\boldsymbol{D}_{i+1} \in \mathbf{R}^{C\times H_i\times W_i}$ (i=1,3), 按通道进行拼接得到 $\boldsymbol{D}_{\text{in}} \in \mathbf{R}^{2C\times H_i\times W_i}$, 定义为

$$\boldsymbol{D}_{\text{in}} = \text{Concat}(\boldsymbol{D}_i, \boldsymbol{D}_{i+1}), i = 1,3$$

2) 经过最大池化和平均池化处理后, 可以得到 2 个中间特征图, 将这 2 个特征图按通道拼接后分别采用水平卷积和垂直卷积处理, 并用 Sigmoid 激活后得到 $\boldsymbol{D}_{\text{hor}} \in \mathbf{R}^{1\times H_i\times W_i}$ 和 $\boldsymbol{D}_{\text{ver}} \in \mathbf{R}^{1\times H_i\times W_i}$, 公式为

$$\boldsymbol{D}_{\text{hor}} = \sigma(\text{Conv}_{1\times k}(\text{Concat}(\text{Avg}(\boldsymbol{D}_{\text{in}}) \,\|\, \text{Max}(\boldsymbol{D}_{\text{in}})))) \quad (2)$$

$$\boldsymbol{D}_{\text{ver}} = \sigma(\text{Conv}_{k\times 1}(\text{Concat}(\text{Avg}(\boldsymbol{D}_{\text{in}}) \,\|\, \text{Max}(\boldsymbol{D}_{\text{in}})))) \quad (3)$$

式中: $\text{Conv}_{1\times k}(\cdot)$ 表示输入通道为 2、输出通道为 1、卷积大小为 $1\times k$ 的卷积操作, $\text{Conv}_{k\times 1}(\cdot)$ 表示输入通道为 2、输出通道为 1、卷积大小为 $k\times 1$ 的卷积操作, k 为超参数, 默认设置为 3, 调参过程见 3.5.1 节; Avg(·) 和 Max(·) 分别表示平均池化和最大池化操作, 采用 2 种池化方式, 防止了单一池化方式造成的信息丢失。$\boldsymbol{D}_{\text{hor}}$ 表示水平方向的特征, $\boldsymbol{D}_{\text{ver}}$ 表示垂直方向的特征, 两者尺度相同, 通道数均为 1。

3) 将 $\boldsymbol{D}_{\text{hor}}$ 和 $\boldsymbol{D}_{\text{ver}}$ 相加得到融合权重, 将 $\boldsymbol{D}_{\text{in}}$ 与融合权重相乘后的结果再与 $\boldsymbol{D}_{\text{in}}$ 相加即为输出 $\boldsymbol{D}_{\text{out}} \in \mathbf{R}^{2C\times H_i\times W_i}$, 定义为

$$\boldsymbol{D}_{\text{out}} = \boldsymbol{D}_{\text{in}} \times (\boldsymbol{D}_{\text{hov}} + \boldsymbol{D}_{\text{ver}}) + \boldsymbol{D}_{\text{in}}$$

3　实验结果及分析

3.1　数据集

ICDAR (international conference on document analysis and recognition) 2015 数据集[27] 主要由广告牌、商标文字等自然场景图片构成, 包括不同方向、不同尺度的英文字符, 共包含 1 500 张图片, 其中训练集 1 000 张, 测试集 500 张。

MTWI (ICPR 2018 contest on robust reading for multitype web images) 数据集[28] 为基于网络图像的中英混合数据集, 主要由合成图像、产品描述、网络广告构成, 涵盖几十种字体, 包含密集的小文本或多语言文本, 共包含 20 000 张图片, 其中训练集 10 000 张, 测试集 10 000 张。

ShopSign1265 数据集[29] 是河南大学张重生教授团队收集的中文街景数据集, 主要包括广告牌和店铺招牌, 采用四边形框对文本进行标注, 共包含 1 265 张图片, 其中训练集 1 012 张, 测试集 253 张。

3.2　实验设置

本次实验的训练和测试都是在 Ubuntu18.04 系统下进行的, 显卡型号是 NVIDIA GeForce RTX 3090, 显存大小 24 GB, CUDA 为 10.2 版本, 训练采用随机梯度下降法。

实验初始动量设置为 0.9, 权重衰减设置为 0.000 1, 初始学习率设置为 0.007, 采用指数变换策略动态调整学习率。为了提高训练效率, 本文首先对所有文本图像进行随机缩放, 在保持原有长宽比的情况下, 短边缩放至 640 ~ 800 像素, 长边不超过 1 600 像素, 接着在 (−10°, 10°) 的角度内进行随机旋转, 随后使用固定大小的矩形框进行任意位置裁剪并随机翻转。经过数据增强后的文本图像尺度各异, 角度各不相同, 能很好地提高模型的鲁棒性。

3.3　评价指标

为了与其他文本检测算法比较, 本文使用召回率 R、准确率 P 和 F 值 (F-measure, F) 3 个指标衡量自然场景下文本检测算法性能。计算公式为

$$R = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}}$$

$$P = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FP}}}$$

$$F = 2 \times \frac{P \times R}{P + R}$$

式中: N_{TP} 表示真正类 (true positive, TP), N_{FP} 表示假正类 (false positive, FP), N_{FN} 表示假负类 (false negative, FN), 三者为根据实际分类与预测分类将预测图像划分的结果, 真正类表示将正类预测为正类的结果, 假正类表示将负类预测为正类的结果, 假负类表示将正类预测为负类的结果。本实验中, 文本区域设置为正类, 用 1 表示, 背景区域设置为负类, 用 0 表示。

3.4　对比实验

为了验证本文方法的有效性, 在 ICDAR2015 数据集、ShopSign1265 数据集和 MTWI 数据集上将本文提出方法与 TextSnake[30]、PSENet[12]、PAN[15]、ContourNet[31]、DRRGN[32]、FCENet[33]、DBNet++[34] 和 MS-ROCANet[35] 进行对比。在本文模型中, MiT-B1 和 MiT-B2 分别为 Segformer-B1[25] 和 Segformer-B2[25] 的解码器部分, 二者的区别在于 Transformer 块的层数不同。

3.4.1　各数据集定量分析

表 1 为各模型在 ICDAR2015、ShopSign1265 和 MTWI 3 个数据集上的定量比较。从表 1 中可以看出, 在 ICDAR2015 数据集上, 本文所提以 MiT-B1 为编码器的方法速度达到了 18.6 f/s, 召回率为 84.7%, 准确率为 89.1%, F 值为 86.8%; 以 MiT-B2 为编码器的方法速度为 12.8 f/s, 召回率为 85.2%, 准确率为 90.5%, F 值为 87.8%, 在几种

方法中达到了最高, 分别比 TextSnake、PSENet、PAN、ContourNet、DRRGN、FCENet、DBNet++和 MS-ROCANet 高出了 5.2 百分点、2.1 百分点、4.9 百分点、0.9 百分点、1.2 百分点、1.6 百分点、0.5 百分点和 1.4 百分点, 说明本文提出的方法在检测精度方面达到了最优。在速度方面, 以 MiT-B2 为编码器的方法与 DBNet++检测速度类似, 均在 11 f/s 左右, 略低于以 MiT-B1 为编码器方法的 18.6 f/s。虽然 PAN 达到了最快的 26.1 f/s, 但是在召回率、准确率和 F 值方面与以 MiT-B2 为编码器的方法还是有一定差距, 原因在于 PAN 选取了 ResNet18[9] 作为主干网络, 在编码器生成多尺度特征图过程中, 连续使用卷积进行下采样操作导致特征信息丢失较多, 而且后续过程无法恢复。

表 1　不同模型在 ICDAR2015 数据集、ShopSign1265 数据集、MTWI 数据集上的定量比较

Table 1　Comparison of results of different models on ICDAR2015 dataset, ShopSign1265 dataset and MTWI dataset

模型	ICDAR2015数据集				ShopSign1265数据集				MTWI数据集			
	召回率/%	准确率/%	F值/%	速度/(f/s)	召回率/%	准确率/%	F值/%	速度/(f/s)	召回率/%	准确率/%	F值/%	速度/(f/s)
TextSnake	80.4	84.9	82.6	1.1	48.7	52.1	50.3	0.3	56.9	66.7	61.4	6.3
PSENet	84.5	86.9	85.7	1.6	48.5	67.1	56.3	1.1	52.7	82.3	64.2	9.2
PAN	81.9	84.0	82.9	**26.1**	48.0	63.6	54.7	**18.3**	65.9	83.5	73.7	**62.5**
ContourNet	**86.1**	87.6	86.9	3.5	47.2	69.3	56.1	1.7	61.8	74.8	67.7	18.6
DRRGN	84.7	88.5	86.6	3.5	53.2	56.4	54.8	1.3	65.4	75.2	69.9	15.3
FCENet	82.6	90.1	86.2	—	53.4	56.5	54.9	—	63.2	81.0	71.0	—
DBNet++	83.9	90.9	87.3	10.0	44.1	**74.3**	55.3	4.5	60.9	**86.5**	71.5	43.2
MS-ROCANet	83.2	89.8	86.4	—	49.8	63.3	55.8	—	65.9	83.9	73.8	—
本文(MiT-B1)	84.7	89.1	86.8	18.6	52.7	63.6	57.6	8.9	66.0	82.8	73.4	46.7
本文(MiT-B2)	85.2	**90.5**	**87.8**	12.8	**54.8**	64.2	**59.1**	5.7	**67.9**	83.2	**74.8**	38.5

在 Shopsign1265 数据集的检测结果中可以看出, 以 MiT-B2 为编码器的方法的召回率最高达到了 54.8%, 以 MiT-B1 为编码器的方法召回率为 52.7%, 位于第 4 位。PSENet、ContourNet 和 DBNet++的准确率分别比以 MiT-B2 为编码器的方法高出 2.9 百分点、5.1 百分点和 10.1 百分点, 但是 F 值却分别低了 2.8 百分点、3.0 百分点和 3.8 百分点。以 MiT-B2 为编码器的方法与 TextSnake 相比召回率、准确率和 F 值分别提升了 6.1 百分点、12.1 百分点和 8.8 百分点。与 PAN 相比, 召回率、准确率和 F 值分别提升了 6.8 百分点、0.6 百分点和 4.4 百分点。与 DRRGN 相比, 召回率、准确率和 F 值分别提升了 1.6 百分点、7.8 百分点和 4.3 百分点, DRRGN 虽然可以检测任意形状的文本区域, 但是对于小尺度文本特征不敏感, 容易出现漏检。将以 MiT-B2 为编码器的方法与 FCENet 相比, 召回率、准确率和 F 值分别提升了 1.4 百分点、7.7 百分点和 4.2 百分点。相比于最近提出的 MS-ROCANet, 以 MiT-B2 为编码器的方法的召回率、准确率和 F 值分别提升了 5.0 百分点、0.9 百分点和 3.3 百分点。在速度方面, 以 PAN 为编码器的方法达到了最快的 18.3 f/s, 由于图像分辨率较高, 其他方法速度差距不明显。

在 MTWI 数据集上, 以 MiT-B2 为编码器的方法的召回率和 F 值分别为 67.9% 和 74.8%, 在各个对比方法中达到了最高, 准确率低于 PAN、DBNet++和 MS-ROCANet。在速度方面, PAN 的速度最快, 达到了 62.5 f/s, 本文提出的以 MiT-B1 为编码器的方法速度为 46.7 f/s, 位于第 2, 而以 MiT-B2 为编码器的方法与 DBNet++的检测速度类似, 均在 40 f/s 左右。

3.4.2　ICDAR2015 数据集定性分析

图 10 为 ICDAR2015 数据集上的可视化结果。在图中的右侧提示牌区域, 大部分方法存在较明显的文本漏检现象, TextSnake 的结果出现了 4 处, 分别是第 1 行右侧、第 2 行右侧和第 4、5 行, 而且没有将相邻的文本实例划分开; ContourNet 的结果比 TextSnake 的结果稍好, 能较准确地划分出相邻文本间的间隔, 但仍出现 4 处漏检; PSENet、PAN、DRRGN 和 FCENet 均出现了 3 处, 漏检的区域大致相同, 为第 1 行左右两侧字符、第 2 行最右侧字符和第 3 行最左侧字符; DBNet++表现较好, 仅出现 1 处漏检。从图 10(g)和 (h) 可以看出, DBNet++和 MS-ROCANet 均出现了文本边缘定位不准确的情况, DBNet++在第 1 行和第 2 行的最右侧字符文本检测不完整, MS-ROCANet 则只检测到了文本行, 没有根据字符之

间的间隔进一步划分。本文提出的以 MiT-B1 为编码器的方法和以 MiT-B2 为编码器的方法均可以准确定位文本区域，结果边界清晰完整，没有漏检的情况。

图 10　ICDAR2015 数据集的检测结果

Fig. 10　Results of the ICDAR2015 dataset

3.4.3　ShopSign1265 数据集定性分析

图 11 为光照变化较明显的场景下的可视化结果，对于上方的大尺度文本，由于图像中只包含第 2 个字符的一部分，且真值图中并未标注，所以不作为参考。如图 11(b)、(d) 所示，PSENet 和 ContourNet 在左上光照较强的部分出现了较多的字符漏检现象，而且检测出的文本之间的间隔不明显，存在重叠的现象。从图 11(a)、(c)、(e) 和图 11(g) 可以看出，虽然 TextSnake、PAN、DRRGN 和 DBNet++均能检测到大部分字符，但是细节信息学习不充分，字符之间、文本行之间的界限划分不明显。FCENet 和以 MiT-B1 为编码器的方法可以对相邻文本字符进行较准确的划分，但是对于光照强烈的模糊区域不容易将文本与背景分开，均出现了 2 处漏检的情况，如图 11(f)、(i) 所示。而以 MiT-B2 为编码器的方法字符区域边缘定位准确，对不同尺度的字符有良好的感知能力。

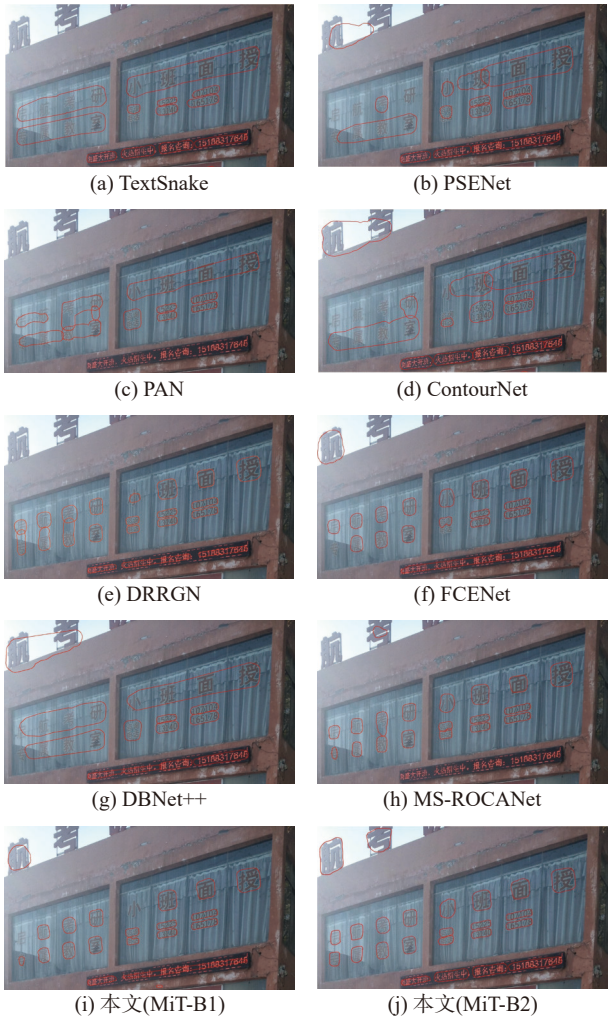


图 11　ShopSign1265 数据集的检测结果

Fig. 11　Results of the ShopSign1265 dataset

3.4.4　MTWI 数据集定性分析

MTWI 数据集上的可视化结果如图 12 所示，主要对比的区域有左侧的艺术字和右侧镜头上的小尺度文本。对于左侧艺术字区域的检测情况如下：如图 12(a) 和图 12(g) 所示，TextSnake 没有检测出艺术字，DBNet++仅检测出 1 个文本实例，对不规则的文本实例鲁棒性较差；如图 12(b)、(c) 和图 12(e) 所示，PSENet、PAN 和 DRRGN 会将几个邻近的文本预测为一个整体，检测能力有待提升。对于小尺度文本，如图 12(d)、(f) 和 (g) 所示，ContourNet、FCENet 和 DBNet++检测效果比较好，能准确检测出大部分文本实例，但是位于镜头中间左侧的文本存在漏检情况。如图 12(i) 和 (j) 所示，以 MiT-B1 为编码器的方法和以 MiT-B2 为编码器的方法明显优于其他方法，证明了本文提出方法有充分的学习能力和良好的鲁棒性。

(a) TextSnake　(b) PSENet　(c) PAN　(d) ContourNet　(e) DRRGN　(f) FCENet　(g) DBNet++　(h) MS-ROCANet　(i) 本文(MiT-B1)　(j) 本文(MiT-B2)

图 12　MTWI 数据集的检测结果

Fig. 12　Results of the MTWI dataset

3.4.5　模型复杂度分析

为了研究本文模型的空间复杂度和时间复杂度, 表 2 从总参数量和浮点运算次数 (gigabit floating-point operations per second, GFLOPS) 2 方面对 TextSnake、PSENet、PAN、DRRGN、DBNet++、以 MiT-B1 为编码器的方法和以 MiT-B2 为编码器的方法进行了比较。由表 2 可以看出, 以 MiT-B2 为编码器的方法参数量为 21.49 MB, 运算次数为 5.59×10^9, 相比于检测速度相似的 DBNet++分别降低了 17.12% 和 26.45%。TextSnake、PSENet 和 DRRGN 的计算量分别超出了以 MiT-B2 为编码器的方法的 220.21%、156.71% 和 481.40%, 这是由于以上 3 种方法后处理过程比较复杂, 导致模型推理速度比较慢; PAN 方法虽然参数量和计算量在几种方法中达到了最优, 但是检测效果与以 MiT-B2 为编码器的方法有一定差距。

表 2　不同模型的参数量和计算量比较

Table 2　Comparison of computational and parametric quantities of different models

模型	参数量/MB	GFLOPS/10^9
TextSnake	19.12	17.90
PSENet	28.63	14.35
PAN	**11.61**	**3.52**
DRRGN	40.80	32.50
DBNet++	25.93	7.60
本文(MiT-B1)	14.44	4.05
本文(MiT-B2)	21.49	5.59

3.5　消融实验

3.5.1　不同超参数对模型的影响

如式 (1) ~ (3) 所示, 本文所提模型有 2 个超参数, 分别是 CFA 模块的 GM 部分用于决定标准卷积输出通道数 $m=C/s$ 中的 s, 以及 D-OFA 模块中决定 2 个正交卷积核大小的 k。

首先固定 k=3, 将 s 调整为{2, 3, 4, 5}中的任意一个, 结果如表 3 所示。随着 s 的增大, 经过普通卷积形成的特征通道数减少, 线性操作形成的特征通道数会变多, 虽然减少了 CFA 模块整体的参数量, 但是模型的精度也逐渐降低。s=2 比 s=5 的召回率、准确率和 F 值分别提高了 6.0 百分点、1.7 百分点和 4.1 百分点, 参数量增加了 143.1 KB。

表 3　CFA 中的 s 取不同值时 ICDAR2015 数据集的结果

Table 3　Experimental results of ICDAR2015 dataset with different values of s in CFA

s	召回率/%	准确率/%	F值/%	参数量/KB
2	**85.2**	**90.5**	**87.8**	336.18
3	83.0	89.9	86.3	270.51
4	81.1	90.2	85.4	220.32
5	79.2	88.8	83.7	**193.08**

接着，将 CFA 中的 s 设置为 2，在{1,3,5,7}范围内调整 D-OFA 中的卷积核大小 k，结果如表 4 所示。

表 4　D-OFA 中的 k 取不同值时 ICDAR2015 数据集的结果
Table 4　Experimental results of ICDAR2015 dataset with different values of k in D-OFA　%

组号	$Conv_{1\times k}$	$Conv_{k\times 1}$	召回率	准确率	F值
1	1	1	83.5	90.1	86.7
	3	3	**85.2**	90.5	**87.8**
	5	5	83.1	89.5	86.2
	7	7	83.3	89.0	86.1
2	1	3	84.8	88.6	86.7
	5	3	83.9	**91.5**	87.5
	7	3	82.9	88.2	85.5
3	3	1	84.1	90.8	87.3
	3	5	84.4	90.3	87.2
	3	7	83.1	89.9	86.4

由于 D-OFA 模型的参数量均在 20B 左右，没有单独列出。第 1 组实验验证 $Conv_{k\times 1}$ 和 $Conv_{1\times k}$ 中的 k 取相同值的结果，可以看到 k=3 的情况下 F 值最高，这是因为 k=1 时的卷积核大小为 1×1，不能引入更多空间信息，而当 k=5 或 k=7 时，步幅设置为 1 会导致特征冗余，引入更多的计算量。第 2 组和第 3 组实验则是 $Conv_{k\times 1}$ 和 $Conv_{1\times k}$ 中的 k 取不同值的结果，结果最好的 F 值为 87.5%，相比 k 均取 3 时的召回率和 F 值降低了 1.3 百分点和 0.3 百分点，这是由于水平卷积和垂直卷积所提取的特征存在差异，融合后会降低模型的泛化能力。

3.5.2　采用不同编码器的结果对比

为了证明本文基于 Transformer 编码器的有效性，将 MiT-B2[25] 与 ResNet[13]、ShuffleNetV2[36]、MobileNetV2[37]、ResNeSt[38]、ConvNeXt[39] 共 5 种卷积神经网络进行对比。本文采用的 ResNet50 主体包含 4 个阶段，每个阶段都包含不同数量的残差块，通过残差连接解决了网络层数加深性能下降的问题。

表 5 为采用不同编码器在 ICDAR2015 数据集上实验的结果。从表 5 可以看出，以 MiT-B2 为编码器的方法在召回率、准确率和 F 值 3 个指标上均达到了最优。以 MiT-B2 为编码器的方法的 F 值比以 MobileNetV2 为编码器的方法的 F 值高出 7.6 百分点，表中 6 种方法在准确率方面基本持平，准确率最高的以 MiT-B2 为编码器的方法与最低的以 ConvNeXt 为编码器的方法相差 5.0 百分点，但是召回率方面相差较大，以 MiT-B2 为编码器的方法比以 MobileNetV2 为编码器的方法高出了 11.5 百分点。在召回率方面，以 ResNet 为编码器的方法与以 MiT-B2 为编码器的方法相差 6.8 百分点。以 ConvNeXt 为编码器的方法使用 7×7 的卷积增加感受野，召回率达到了 84.7%。以 MiT-B2 为编码器的方法使用 Transformer 增强全局信息的交互能力，生成特征图的语义信息和空间信息更加丰富，所以召回率相比第 2 名以 ConvNeXt 为编码器的方法提高了 0.5 百分点，F 值也提升了 2.7 百分点。

表 5　不同编码器在 ICDAR2015 数据集的实验结果
Table 5　Using different encoder in the ICDAR2015 dataset　%

编码器	召回率	准确率	F值
ResNet	78.4	89.3	83.5
ShuffleNetV2	77.6	88.2	82.6
MobileNetV2	73.7	87.9	80.2
ResNeSt	81.4	89.1	85.1
ConvNeXt	84.7	85.5	85.1
MiT-B2	**85.2**	**90.5**	**87.8**

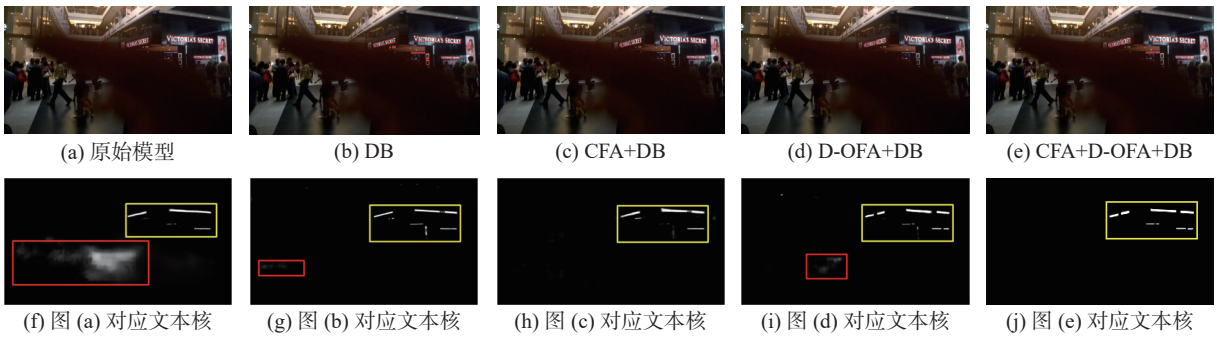
3.5.3　采用不同注意力模块的结果对比

为了验证本文所提出的 CFA 和 D-OFA 的有效性以及 DB 算法在模型中的作用，在 ICDAR2015 数据集上进行消融实验。表 6 为采用不同模块在 ICDAR2015 数据集上的结果。图 13 为在 ICDAR-2015 数据集上的可视化结果。图 13(a)~(e) 为 5 种模型的检测结果，图 13(f)~(j) 为分别对应的文本核区域，白色为检测出的文本区域，黑色为背景，颜色越白代表成为最终文本框的概率越大。

表 6　不同模块在 ICDAR2015 数据集的实验结果
Table 6　Using different module in the ICDAR2015 dataset　%

模型	召回率	准确率	F值
MiT-B2+FPN	79.0	87.3	82.9
MiT-B2+FPN+DB	77.6	88.6	84.6
MiT-B2+FPN+CFA+DB	82.8	89.7	86.1
MiT-B2+FPN+D-OFA+DB	83.5	**90.5**	86.9
MiT-B2+FPN+CFA+D-OFA+DB	**85.2**	**90.5**	**87.8**

DB 模块的作用　通过表 6 第 1 行和第 2 行可以看出，在使用 DB 算法进行后处理的情况下，准确率和 F 值分别提高了 1.3 百分点和 1.7 百分点，但是召回率降低了 1.4 百分点。图 13(b) 为 MiT-B2 +FPN+DB 的结果，相比图 13(a)，引入 DB 模块减少了大规模误检，但由于上采样过程会导致信息丢失，仍出现了小尺度误检的情况。



(a) 原始模型　(b) DB　(c) CFA+DB　(d) D-OFA+DB　(e) CFA+D-OFA+DB

(f) 图 (a) 对应文本核　(g) 图 (b) 对应文本核　(h) 图 (c) 对应文本核　(i) 图 (d) 对应文本核　(j) 图 (e) 对应文本核

图 13　不同模块在 ICDAR2015 数据集的实验结果

Fig. 13　Results of using different module in the ICDAR2015 dataset

CFA 和 DB 模块的作用　通过表 6 第 2 行和第 3 行比较, 在只引入 CFA 时, 召回率、准确率和 F 值分别提高了 5.2 百分点、1.1 百分点和 1.5 百分点, 证明了 CFA 模块的有效性。图 13(c) 为 MiT-B2+FPN+CFA+DB 的结果, 通过 CFA 获得了更多的细节信息, 可以大大减少误检率, 但由于背景信息的干扰会出现漏检, 如图 13(h) 黄色区域中间部分, 将一处横向文本区域误认为背景。

D-OFA 和 DB 模块的作用　通过第 2、4 行比较, 引入 D-OFA 模块后, 召回率、准确率和 F 值分别提升了 5.9 百分点、1.9 百分点和 2.3 百分点, 证明了 D-OFA 模块的有效性。图 13(d) 为 MiT-B2+FPN+D-OFA+DB 的结果, 加入 D-OFA 后, 增强了不同方向之间的空间信息交互能力, 在一定程度上减少了漏检的情况, 进一步细化了文本区域的边界。在图 13(i) 的黄色框中可以看出 4 处文本均能正确检测, 而且相比于图 13(f) ~ (h), 不存在文本之间粘连的情况, 但由于上采样过程中的信息丢失, 在红色框区域出现了背景误检的情况。

CFA、D-OFA 和 DB 模块的作用　在同时引入 CFA、D-OFA 以及 DB 算法的情况下, 召回率提升了 6.2 百分点、准确率提升了 3.2 百分点、F 值提升了 4.9 百分点。图 13(e) 为 MiT-B2+FPN+CFA+D-OFA+DB 的结果, 同时加入 CFA、D-OFA 和 DB, 文本区域边界定位准确, 没有出现误检、漏检的情况, 如图 13(j) 所示, 文本核明显突出而且没有无关像素的影响。

通过以上实验可以证明, DB 通过设置可学习的阈值, 提升了文本区域和背景划分的准确度; CFA 可以级联相邻尺度特征图强化全局信息, 减少误检率; D-OFA 通过对水平和垂直方向的空间特征融合, 有效抑制了背景像素的干扰, 突出文本区域特征, 从而说明了本文提出模型的各模块都是必要的, 也说明了本文提出模型具有有效性。

4　结束语

针对自然场景文本检测的复杂背景像素误报、文本漏检和边缘定位不准确的问题, 提出了一种基于 Segformer 的端到端文本检测模型。通过 CFA, 有效增强了小尺度文本的全局感知能力; 在特征融合过程中引入 D-OFA, 细化了文本之间的位置关系和轮廓信息, 减少了背景信息冗余。所提方法在 ICDAR2015、ShopSign1265 和 MTWI3 个数据集上的 F 值均达到了最优, 在参数量和计算量方面也有一定下降。实验结果表明, 所提方法有效解决了文本边缘定位不准确、小尺度文本漏检和背景像素误检的问题。在未来的工作中, 将针对如何实现自然场景文本的实时检测展开研究。一方面, 本文所提方法的网络结构较为复杂, 在提升文本检测精度的同时会使高分辨率的图像推理时间增加, 达不到实时检测的要求。另一方面, 可微分二值化算法将阈值设置为可学习的参数, 但是对于重叠文本区域检测效果较差, 需要对后处理技术进一步改进。

参考文献:

[1]　朱志颖. 基于深度学习的街景文本检测与识别研究 [D]. 南京: 南京邮电大学, 2023.
ZHU Zhiying. Research on street view text detection and recognition based on deep learning[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2023.

[2]　周燕, 韦勤彬, 廖俊玮, 等. 自然场景文本检测与端到端识别: 深度学习方法 [J]. 计算机科学与探索, 2023, 17(3): 577−594.
ZHOU Yan, WEI Qinbin, LIAO Junwei, et al. Natural scene text detection and end-to-end recognition: deep learning methods[J]. Journal of frontiers of computer science and technology, 2023, 17(3): 577−594.

[3]　李祥鹏, 闵卫东, 韩清, 等. 基于深度学习的车牌定位和识别方法 [J]. 计算机辅助设计与图形学学报, 2019, 31(6): 979−987.

LI Xiangpeng, MIN Weidong, HAN Qing, et al. License plate location and recognition based on deep learning[J]. Journal of computer-aided design & computer graphics, 2019, 31(6): 979−987.

[4] 刘光辉, 张钰敏, 孟月波, 等. 双分支跨级特征融合的自然场景文本检测 [J]. 智能系统学报, 2023, 18(5): 1079−1089.
LIU Guanghui, ZHANG Yumin, MENG Yuebo, et al. Natural scene text detection based on double-branch cross-level feature fusion[J]. CAAI transactions on intelligent systems, 2023, 18(5): 1079−1089.

[5] 王润民, 桑农, 丁丁, 等. 自然场景图像中的文本检测综述 [J]. 自动化学报, 2018, 44(12): 2113−2141.
WANG Runmin, SANG Nong, DING Ding, et al. Text detection in natural scene image: a survey[J]. Acta automatica sinica, 2018, 44(12): 2113−2141.

[6] LIU Wei, ANGUELOV D, ERHAN D, et al. SSD: single shot MultiBox detector[C]//European conference on computer vision. Cham: Springer, 2016: 21−37.

[7] REN Shaoqing, HE Kaiming, 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.

[8] JIANG Yingying, ZHU Xiangyu, WANG Xiaobing, et al. R2CNN: rotational region CNN for orientation robust scene text detection[EB/OL]. (2017−06−29)[2023−01−11]. https://arxiv.org/abs/1706.09579.

[9] LIAO Minghui, SHI Baoguang, BAI Xiang, 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.

[10] LIAO Minghui, SHI Baoguang, BAI Xiang. TextBoxes++: a single-shot oriented scene text detector[J]. IEEE transactions on image processing, 2018, 27(8): 3676−3690.

[11] HE Tong, HUANG Weilin, QIAO Yu, et al. Accurate text localization in natural image with cascaded convolutional text network[EB/OL]. (2016−03−31)[2023−01−11]. https://arxiv.org/abs/1603.09423.

[12] LI Yi, WU Zhe, ZHAO Shuang, et al. PSENet: psoriasis severity evaluation network[C]//Proceedings of the AAAI conference on artificial intelligence. Palo Alto: AAAI, 2020: 800−807.

[13] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016: 770−778.

[14] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu: IEEE, 2017: 936−944.

[15] WANG Wenhai, XIE Enze, SONG Xiaoge, et al. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network[C]//2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019: 8439−8448.

[16] LIAO Minghui, WAN Zhaoyi, YAO Cong, et al. Real-time scene text detection with differentiable binarizationk[C]//Proceedings of the AAAI conference on artificial intelligence. Palo Alto: AAAI, 2020: 11474−11481.

[17] 邵海琳, 季怡, 刘纯平, 等. 基于增强特征金字塔网络的场景文本检测算法 [J]. 计算机科学, 2022, 49(2): 248−255.
SHAO Hailin, JI Yi, LIU Chunping, et al. Scene text detection algorithm based on enhanced feature pyramid network[J]. Computer science, 2022, 49(2): 248−255.

[18] 雷小唐, 胡靖. 文本中心像素重建实现任意形状的文本检测 [J]. 计算机工程与应用, 2023, 59(8): 148−156.
LEI Xiaotang, HU Jing. Text center pixel reconstruction to achieve efficient arbitrary shape text detection[J]. Computer engineering and applications, 2023, 59(8): 148−156.

[19] 梁浩然, 叶凌晨, 梁荣华, 等. 注意力监督策略下的自然场景文本检测算法 [J]. 计算机辅助设计与图形学学报, 2022, 34(7): 1011−1019.
LIANG Haoran, YE Lingchen, LIANG Ronghua, et al. Text detection algorithm for natural scenes under attention supervision strategy[J]. Journal of computer-aided design & computer graphics, 2022, 34(7): 1011−1019.

[20] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16 × 16 words: transformers for image recognition at scale [EB/OL]. (2020−10−22) [2023−01−11]. https://arxiv.org/abs/2010.11929.

[21] CHU Xiangxiang, TIAN Zhi, ZHANG Bo, et al. Conditional positional encodings for vision transformers [EB/OL]. (2021−02−22) [2023−01−11]. https://arxiv.org/abs/2102.10882.

[22] HAN Kai, XIAO An, WU Enhua, et al. Transformer in transformer[J]. Advances in neural information processing systems, 2021, 34: 15908−15919.

[23] LIU Ze, LIN Yutong, CAO Yue, et al. Swin transformer: hierarchical vision transformer using shifted windows [C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 2021: 9992−10002.

[24] WANG Wenhai, XIE Enze, LI Xiang, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]//2021 IEEE/CVF International Conference on Computer Vision. Montreal: IEEE,

2021: 548−558.

[25] XIE Enze, WANG Wenhai, YU Zhiding, et al. SegFormer: simple and efficient design for semantic segmentation with transformers[J]. Advances in neural information processing systems, 2021, 34: 12077−12090.

[26] HAN Kai, WANG Yunhe, TIAN Qi, et al. GhostNet: more features from cheap operations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 1577−1586.

[27] KARATZAS D, GOMEZ-BIGORDA L, NICOLAOU A, et al. ICDAR 2015 competition on Robust Reading[C]// 2015 13th International Conference on Document Analysis and Recognition. Tunis: IEEE, 2015: 1156−1160.

[28] HE Mengchao, LIU Yuliang, YANG Zhibo, et al. ICPR2018 contest on robust reading for multi-type web images[C]//2018 24th International Conference on Pattern Recognition. Beijing: IEEE, 2018: 7−12.

[29] ZHANG Chongsheng, PENG Guowen, TAO Yuefeng, et al. ShopSign: a diverse scene text dataset of Chinese shop signs in street views[EB/OL]. (2019−03−25)[2023−01−11]. https://arxiv.org/abs/1903.10412.

[30] LONG Shangbang, RUAN Jiaqiang, ZHANG Wenjie, et al. TextSnake: a flexible representation for detecting text of arbitrary shapes[C]//European conference on computer vision. Cham: Springer, 2018: 19−35.

[31] WANG Yuxin, XIE Hongtao, ZHA Zhengjun, et al. ContourNet: taking a further step toward accurate arbitrary-shaped scene text detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 11750−11759.

[32] ZHANG Shixue, ZHU Xiaobin, HOU Jiebo, et al. Deep relational reasoning graph network for arbitrary shape text detection[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2020: 9696−9705.

[33] ZHU Yiqin, CHEN Jianyong, LIANG Lingyu, et al. Fourier contour embedding for arbitrary-shaped text detection [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville: IEEE, 2021: 3122−3130.

[34] LIAO Minghui, ZOU Zhisheng, WAN Zhaoyi, et al. Real-time scene text detection with differentiable binarization and adaptive scale fusion[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(1): 919−931.

[35] LIU Jinpeng, WU Song, HE Dehong, et al. MS-ROC-ANet: multi-scale residual orthogonal-channel attention network for scene text detection[C]//2022 IEEE International Conference on Acoustics, Speech and Signal Processing. Singapore: IEEE, 2022: 2200−2204.

[36] MA Ningning, ZHANG Xiangyu, ZHENG Haitao, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design[C]//European conference on computer vision. Cham: Springer, 2018: 122−138.

[37] SANDLER M, HOWARD A, ZHU Menglong, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 4510−4520.

[38] ZHANG Hang, WU Chongruo, ZHANG Zhongyue, et al. ResNeSt: split-attention networks[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022: 2735−2745.

[39] LIU Zhuang, MAO Hanzi, WU Chaoyuan, et al. A ConvNet for the 2020s[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022: 11966−11976.

作者简介：

张铭泉，副教授，主要研究方向为计算机组成、机器学习、模式识别。发表学术论文 20 余篇。E-mail: mqzhang@ncepu.edu.cn。

张泽恩，硕士研究生，主要研究方向为深度学习和文本检测。E-mail: zze15832206526@163.com。

曹锦纲，讲师，主要研究方向为图像处理和模式识别。发表学术论文 10 余篇。E-mail: caojg168@126.com。