Fourier Contour Embedding for Arbitrary-Shaped Text Detection 复现

吴坤忠

摘要

任意形状文本检测的主要挑战之一是设计一个良好的文本实例表示，使网络能够学习不同的文本几何差异。现有的方法大多是使用图像掩膜或轮廓点序列在笛卡尔坐标系或极坐标系下对图像空间域的文本实例进行建模。然而，掩码表示可能需要经历繁杂的后处理，而点序列表示可能对具有高度弯曲形状的文本的建模能力有限。为了解决这些问题，本文复现的论文提出在傅里叶域中对文本实例进行建模，并提出了一种新颖的傅里叶轮廓嵌入 (FCE) 方法，可以将任意形状的文本轮廓表示为一个傅里叶向量。提出的模型 FCENet 的结构由骨干网络、特征金字塔网络 (FPN) 和简单的傅立叶反变换 (IFT) 和非最大抑制 (NMS) 后处理构成。与以前的方法不同的是，FCENet 首先预测文本实例的紧凑傅里叶签名向量，然后在测试过程中通过 IFT 和 NMS 重建文本轮廓。大量的实验表明，FCE 对高度弯曲的场景文本轮廓的拟合具有很好的准确性和鲁棒性，实验也验证了 FCENet 对任意形状文本检测的有效性和良好的泛化性。此外，实验结果表明，FCENet 在 CTW1500 和 Total-Text 上优于目前最先进的 (SOTA) 方法，特别是在具有挑战性的高曲线文本子集上

关键词：弯曲文本检测；文本实例表示；傅里叶变换

1 引言

得益于目标检测和实例分割的发展，文本检测在近些年也取得了一些重大进展。目前文本检测方法大致可分为基于分割的方法和基于回归的方法。在近年来，场景文本检测取得了瞩目的进步并被广泛应用到自动驾驶和场景分析等领域。随着文本检测算法的迭代，场景文本检测的关注点已经从原来的水平方向文本和多方向文本转到更具挑战性的任意形状文本上。

为了使文本检测算法在任意形状文本上达到更好的效果，亟需一种任意形状文本实例表示方法来提高算法的性能，良好的文本实例表示应当满足求解简单、表示方法参数量低、灵活度高的特点。目前现有的任意形状文本检测器大多在图像的空间域中表示文本实例，基于空间域的表示方法大体上可以分为两种，即像素掩膜表示和轮廓点序列表示。其中，像素掩膜表示方法可能需要繁杂和耗时的后处理过程，同时对训练样本量的需求往往也会更大；而轮廓点序列表示方法对高度弯曲文本的表达能力有限。由于傅里叶系数表示在理论上可以拟合任意的封闭曲线，并且文本轮廓更多集中在低频分量上，所以通过在傅立叶域对不规则场景文字实例进行表征能很好地解决上述问题，并且具有简单、紧凑、对复杂轮廓表达能力好的特点。

2 相关工作

2.1 基于分割的方法

基于分割的文本检测方法主要是从语义分割中获得灵感的，语义分割使用逐像素掩码隐式编码文本实例 [1,9,12,15,16,18,19,21,22,24]。这些方法大多遵循组件分组范式，首先检测场景文本实例的组件，然后聚合这些组件以获得最终的掩码输出。对于基于像素的方法，首先使用实例/语义分割框架获得像素级评分图，然后对文本像素进行分组，得到输出的文本掩码 [16,19,21,22]。为了进一步提高性能，一些方法会在转换后的空间上进行预测，然后重建最终的映射。例如，Tian 等人的论文中 [16] 将每个文本实例假设为一个聚类，并通过像素聚类预测嵌入映图;TextField [22] 通过链接带有深方向场的相邻像素生成候选文本部分。

还有一些基于片段的方法，它们的指导思想主要是先检测包含部分单词或文本行（片段) [9,10,12,14,15,18] 或字符 [1,24] 的片段，然后将片段分组为整个单词/文本行。PSENet [18] 用相应的核对每个文本实例进行检测，并采用递进尺度算法逐步扩展预定义核，得到最终检测结果。SegLink++ [14] 使用最小生成树的实例感知组件分组实现密集和任意形状的场景文本检测。CRAFT [1] 获得字符级检测，并估计字符之间的亲和力，从而实现最终的检测。

有些方法在变换后的空间中训练预测器，并根据预测的特征重构输出掩模。例如，Tian 等人 [16] 通过将像素嵌入到相同文本的像素倾向于在相同聚类中的空间中构建了判别表示，反之亦然;Xu 等 [22] 提出 TextField 学习一个方向字段来分离相邻的文本实例。

2.2 基于回归的方法

基于回归的方法是对基于分割的方法的补充，基于分割的方法显式地用文本区域的轮廓 (点序列) 编码文本实例。它们旨在采用文本实例的直接形状建模来处理复杂的几何方差 [4,17,23,25–27]，并且通常更简单，更容易训练。然而，点序列对复杂文本实例的约束表示能力可能会限制网络的性能。

为了解决这一问题，论文精心设计了许多模块，以进一步提高点序列表示的灵活性。LOMO [25] 引入了迭代细化模块 (IRM) 和形状表达模块 (SEM) 来逐步细化直接回归的文本定位。Zhang 等 [26] 利用 cnn 对从文本实例中分割出来的一系列小矩形组件的几何属性 (如高度、宽度和方向) 进行回归，并引入一个图卷积网络 (Graph Convolutional Network, GCN) 来推断不同文本组件之间的联系。TextRay [17] 在极坐标系统中制定文本轮廓，并提出了一种单次无锚框架来学习几何参数。论文 [7] 提出了利用 Bezier 曲线对曲线文本进行参数化，并利用 BezierAlign 方法实现了场景文本定位的 SOTA 性能。

最近的研究表明，有效的轮廓建模对于不规则文本实例检测 [17,25,26] 和下游识别 [7] 至关重要。因此，设计一种灵活而简单的任意形状文本检测表示具有重要意义。

2.3 显性与隐性文本形状表示

从文本形状表示的角度来看，目前的文本模型大致可以分为两类。即，通过逐像素掩模 [1,5,9,18–20] 或通过转换特征重建的掩模 [16,22] 隐式地建模文本形状的方法，以及使用笛卡尔系统 [2,26] 或极系统 [17] 中的轮廓点序列显式地建模文本形状的方法。

然而，逐像素掩模可能会导致网络本质上的高计算复杂性 (例如，复杂的后处理)，并且需要大量的训练数据，而在轮廓上采样的点序列可能具有有限的表示能力 [10, 17, 25, 26]。

为了解决这个问题，Liu 等 [7] 引入了 Bezier 曲线来参数化曲线文本，但是在某些情况下，Bezier 曲线的控制点设置可能会限制其表示能力。在本文中，通过在傅里叶域中表述文本实例，能以鲁棒和简单的方式拟合任何封闭连续轮廓。在下一节中，我们将探讨 FCE 在任意形状文本检测中的潜力。

3 本文方法

3.1 本文方法概述

在本节中，将首先介绍论文中提出的傅立叶轮廓嵌入 (FCE) 方法，该方法可以将任意形状的文本轮廓近似为一个紧凑的傅立叶签名向量。然后介绍论文提出的基于 FCE 的 FCENet 模型。

为了将文本实例轮廓从点序列转换为傅立叶特征向量，论文作者首先提出了傅立叶轮廓嵌入 (FCE) 方法，具体是设计了一种重采样方案，在每个文本轮廓上获得固定数量的密集点，然后通过傅里叶变换 (FT) 将轮廓在空间域中采样的点序列嵌入到傅里叶域中。同时为了保持得到的傅里叶特征向量的唯一性，规定以文本轮廓与经过文本中心点的水平线的最右侧交点作为采样起点，将采样方向固定为顺时针方向，并保持沿文本轮廓的采样间隔不变。

之后，基于 FCE 作者进一步构建了用于任意形状文本检测的 FCENet。它由 ResNet50 的主干与可变形卷积网络 (DCN) [28]、特征金字塔网络 (FPN) [6] 和傅立叶预测头组成。预测头有两个独立的分支，即分类分支和回归分支。前者用于预测文本区域掩码和文本中心区域掩码，后者用于在傅里叶域中预测文本傅里叶特征向量，并将其送入傅里叶反变换 (IFT) 中重构文本轮廓点序列。由于 FCE 的重采样方案，作者提出的模型在回归分支中的损失在不同的数据集上是兼容的，尽 CTW1500 [8] 和 Total-Text [2] 等数据集对于每个文本实例具有不同数量的轮廓点。

3.2 傅里叶轮廓嵌入

对于任意封闭轮廓曲线，作者使用一个封闭曲线的参数方程完成任意文本封闭轮廓从实数域到复数域的嵌入，函数的表示如下：

$$f(t) = x(t) + iy(t) \tag{1}$$

其中 i 表示虚单位。$(x(t)\ y(t))$ 表示特定时刻 t 的空间坐标，实变量 $t \in (0, 1)$，因为 f 是闭合轮廓，所以 $f(t) = f(t+1)$，$f(t)$ 可以通过傅里叶反变换 (IFT) 重新表述为:

$$f(t) = f(t, c) = \sum_{k=-\infty}^{+\infty} c_k e^{2\pi i k t} \tag{2}$$

其中 k 为频率 $k \in Z$，c_k 为表征频率 k 初始状态的复值傅立叶系数。Eq.(2)中的每个分量 $c_k e^{2\pi i k t}$ 表示一个固定频率 k 的圆周运动，具有给定的初始手部方向矢量 c_k，因此轮廓可以看作是不同频率圆周运动的组合，如图 1中粉色圆圈所示。

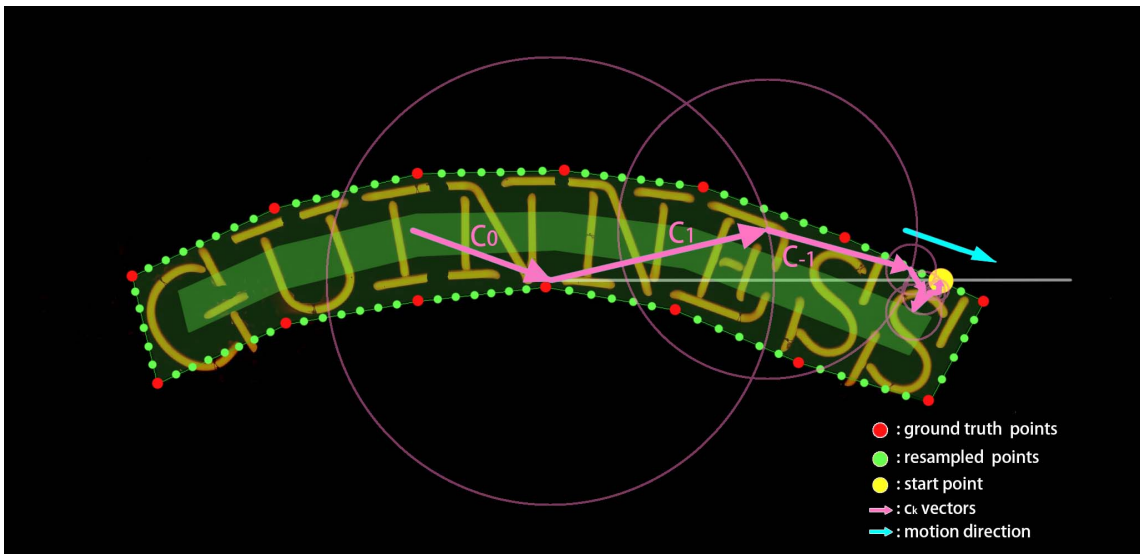

图 1. FCE 示意图

从 Eq.(2)可以看出，低频分量负责粗略的文本轮廓，而高频分量负责轮廓的细节。根据论文作者的经验发现，只保留较少的频率分量，同时丢弃其他频率，可以获得令人满意的文本轮廓近似，(在作者的实验设置中 $k = 5$)。

由于在实际应用中无法得到文本轮廓函数 f 的真实解析形式，所以作者采用离散化将连续函数 f 离散成 N 个点，如 $f(\frac{n}{N})$，其中 $n \in [1, ..., N]$。在这种情况下，Eq.(2)中的 c_k 可以通过傅里叶变换计算为:

$$c_k = \frac{1}{N} \sum_{n=1}^{N} c_k e^{-2\pi i k \frac{n}{N}} \tag{3}$$

式中 $c_k = u_k + iv_k$，其中 u_k 为复数的实部，v_k 为复数的像部。特别地，当 $k = 0$ 时，$c_0 = u_0 + iv_0 = \frac{1}{N}\sum_n f(\frac{n}{N})$ 为轮廓的中心位置。对于任意文本轮廓 f，论文提出的傅立叶轮廓嵌入 (FCE) 方法可以将其在傅立叶域中表示为紧凑的 2(2K + 1) 维向量 $[u_{-K}, v_{-K} \cdots u_0, v_0 \cdots u_K, v_K]$，称为傅立叶签名向量。

3.3 FCENET

作者提出的 FCENet 采用自上而下的方案。如图 2所示，它包括带可变形卷积 [28] 的 ResNet50 [3] 特征提取网络作为主干，以特征金字塔网络 FPN [6] 作为 neck 层用来提取多尺度特征，以及傅里叶预测头。作者对 FPN 的特征图 P3、P4 和 P5 进行预测，头部网络有两个分支，分别负责分类和回归。每个分支由三个 3 × 3 卷积层和一个 1×1 卷积层组成，每个卷积层后面都有一个 ReLU 非线性激活层。

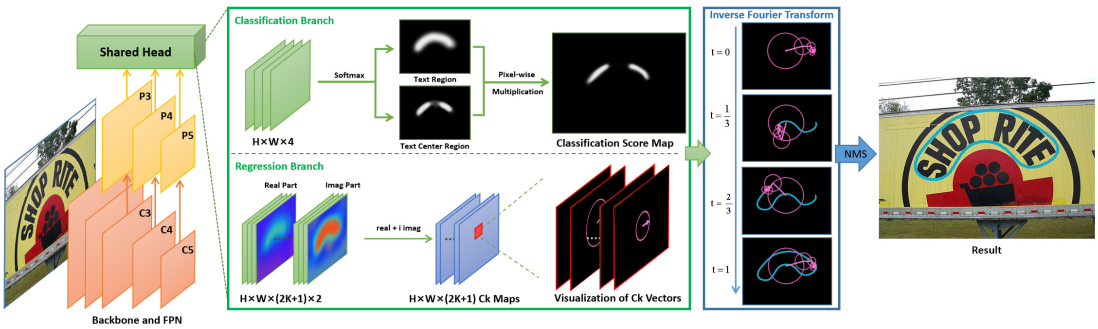

图 2. FCENET 模型结构图

在分类分支中，论文预测文本区域 (TR, Text Regions) 的每一像素的掩膜。论文发现文本中心区域 (TCR, Text Center Region) 的预测能进一步提升性能。论文认为这就是因为 TCR 可以有效地过滤掉文本边界周围的低质量预测。在回归分支中，模型对文本中的每个像素对文本的傅里叶特征向量进行回归。为了处理不同规模的文本实例，P3、P4 和 P5 的特征分别负责小、中、大文本实例。最后通过 IFT 和 NMS 将检测结果从傅里叶域重构到空间域。

在生成 Ground-Truth 时，对于分类任务，论文使用了 TextSnake 中的方式来获得 TCR 的掩膜，通过文本收缩蒙版，收缩因子为 0.3(如图 1中绿色蒙版)。对于回归任务，通过使用收缩因子为 0.3 0.30.3 来收缩文本区域，如 Fig.2 中绿色掩膜区域。对于回归任务，论文通过提出的 FCE 方法计算了真实文本轮廓的傅里叶特征向量 $\bar{c}$。对于一个文本实例掩模中的所有像素，模型都预测一个文本轮廓。同一文本实例中的不同像素共享相同的傅里叶签名向量，但 c_0 除外。

3.4 损失函数定义

基于 FCE 的网络的优化目标函数如下:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_{reg} \tag{4}$$

其中 $\mathcal{L}_{cls}$ 和 $\mathcal{L}_{reg}$ 分别是分类分支和回归分支的损失。λ 是平衡 $\mathcal{L}_{cls}$ 和 $\mathcal{L}_{reg}$ 的参数。作者在实验中固定 $\lambda = 1$。$\mathcal{L}_{reg}$ 由两部分组成:

$$\mathcal{L}_{cls} = \mathcal{L}_{tr} + \lambda \mathcal{L}_{tcr} \tag{5}$$

其中 $\mathcal{L}_{tr}$ 和 $\mathcal{L}_{tcr}$ 分别为文本区域 (TR) 和文本中心区域 (TCR) 的交叉熵损失。为了解决样本不平衡问题，$\mathcal{L}_{tr}$ 采用 OHEM [13]，负样本与正样本之比为 3:1。

对于 $\mathcal{L}_{reg}$，论文不是对预测的傅里叶特征向量与相应的真实值进行最小化距离优化，而是对重建后的空间域文本轮廓进行最小化优化，更能反应文本检测的质量，公式如下:

$$\mathcal{L}_{reg} = \frac{1}{N'} \sum_{i \in \tau} \sum_{n=1}^{N'} w_i l_1(F^{-1}(\frac{n}{N'}, \bar{c}_i), F^{-1}(\frac{n}{N'}, \hat{c}_i)) \tag{6}$$

其中 l_1 为用于回归的 smooth-l_1 损失 [11]，$F^{-1(\cdot)}$ 为 Eq.(2)中的 IFT。τ 为文本区域像素索引集。$\bar{c}_i$ 和 $\hat{c}_i$ 是真实文本标签的傅里叶签名向量和像素 i 的预测向量。如果像素 i 在对应的文本中心区域，则 $w_i = 1$，如果不在，则为 0.5。N' 是文本轮廓上的采样数。如果 N' 太小 (通常 $N' < 30$)，可能会导致过拟合。因此，作者在实验中固定 $N' = 50$。

4 复现细节

4.1 与已有开源代码对比

本文的复现过程使用了源代码，借助源代码复现了论文中的实验，论文中有两个实验的数据集是在开源数据集上改造的，作者并没有公开这两个改造后的数据集以及用于制造数据集的代码，本文根据原论文的描述和指导，自己编写了制造这两个数据集的代码，并使用改造的数据集进行了原论文的实验，实验结果与原论文相符。

4.2 实验环境搭建

软件环境：运行实验的操作系统为 ubantu 18.04，实验代码使用 python 编写，所以基于 python 搭建相关的软件环境，首先实验使用的是 python3.8，其次使用的深度学习框架为 pytorch，版本为 1.10.0。由于训练使用了 NVIDIA 的 GPU，所以还需安装对应的 cuda，版本为 10.2。同时代码实现主要借助 OpenMMLab 的算法库，使用了 MMOCR 1.0.1，mmdet 3.1.0。

实验平台：实验是在搭载了 4 张 NVIDIA Tesla V100 显卡的服务器上进行的。

5 实验结果分析

主要复现了论文中的三个实验，第一个实验是一个性能对比实验，首先在 CTW1500，Total-Text 和 ICDAR2015 三个数据集上分别训练并测试模型的性能，如表1所示是我复现得到的实验数据，其中包含了两个模型，上面的模型和下面的模型区别就在于它们的 backbone 里有没有使用可变形卷积，上面是没有包含的，下面的是有包含的。可变形卷积是在传统的 CNN 中引入了可变行卷积核，传统的卷积操作是在固定的网格上进行的，可变形卷积通过添加一个卷积核的偏移量，使得卷积核可以在输入上产生局部的形变。它在处理具有形变结构的图像或目标时表现得更出色，所以下面这个包含可变形卷积的模型具有更好的性能。复现的结果表明 FCENet 在 CTW1500 和任意形状文本的 Total-Text 数据集上获得了最佳的精度 (P) 和 F-measure (F) 性能，并取得了不错的召回率 (R) 性能，复现得到的数据基本与原论文的数据是一致的。

Methods	CTW1500			Total-Text			ICDAR1500		
	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)
FCENET'	74.1	77.3	75.7	75.5	82.7	78.9	80.5	89.3	84.7
FCENET	84.7	85.4	85.0	83.2	88.6	85.8	84.7	85.4	85.0

表 1. 在 CTW1500、Total-Text 和 ICDAR2015 上的性能测试，FCENet' 表示使用不带 DCN 的 ResNet50 作为主干的 FCENet

复现的第二个实验是对模型泛化能力的验证实验，方法是通过将 CTW1500 数据集切割成原来的 50% 和 25%，然后使用切割后的数据集对 FCENET' 与 ABCNet 模型进行训

练，比较不同数据集对两者的性能影响情况。如表2所示是我的实验结果，得到的结果表明与 FCENET’ 模型比较，ABCNet 在训练数据减少的时候，性能下降比较多，而 FCENET’ 的性能则不会下降太多，能保持在一个比较高的水平。需要说明的是，由于官方没有给出他们做这个实验时候的数据集，所以这个实验的数据集是自己根据论文的说明编写代码将原来的大数据集随机切分出来的，因此测出来的数据跟原论文的数据可能有一定的出入，但总体趋势和现象以及最后得出的结论是一致的。

Data	Methods	R(%)	P(%)	F(%)
100%	ABCNet	73.5	75.5	74.5
	FCENET’	74.1	77.3	75.7
50%	ABCNet	66.7	71.9	69.2
	FCENET’	69.1	75.3	72.1
25%	ABCNet	65.2	67.6	66.4
	FCENET’	68.9	71.8	70.3

表 2. 不同训练数据下模型在 CTW1500 泛化能力比较

复现的第三个实验是在包含高度弯曲文本的数据集下的测试实验，该实验是为了验证论文的模型在面对高弯曲文本的时候的优秀性能表现。为此论文的实验方法是从 CTW1500 数据集中将其中非弯曲的文本数据去除，只留下高度弯曲的文本，从而构造出一个包含高度弯曲文本的数据子集，然后在这个数据子集下对模型的性能进行验证。这里同样由于官方没有给出他们做该实验时候的数据集，所以本文所作复现实验的数据集是自己按照论文中的说明自己写程序去提取的，构造出来的数据子集跟原论文作者的肯定是不完全一样的，所以数据上也会有所差距。如图 3是我复现的可视化结果图，定性表现了 FCENET 的优越性能。如表3所示是复现的实验结果，总共测试两个模型，结果定量的表明了相比于 ABCNet，论文提出的 FCENET 在检测高弯曲文本时具有更好的性能，体现出了该模型在高弯曲文本检测上的优越性，该现象与结论也是跟论文中的一致。

图 3. CTW1500 高度弯曲线文本子集的定性比较

Methods	R(%)	P(%)	F(%)
ABCNet	66.8	86.3	75.3
FCENET	83.0	84.3	83.6

表 3. CTW1500 高度弯曲线文本子集的定量比较

6 总结与展望

论文主要研究用于任意形状文本检测的显式形状建模，提出了傅里叶轮廓嵌入方法，该方法可以精确地逼近任何封闭形状。基于该方法进一步提出了 FCENet，它首先在傅里叶域中预测文本实例的傅里叶特征向量，然后通过傅里叶反变换在图像空间域中重构文本轮廓点序列。FCENet 可以端到端优化，无需任何复杂的后处理即可实现。本文复现了原论文对 FCENet 进行的广泛的评估。实验结果验证了 FCE 的表示能力，特别是对高曲线文本的表示能力，以及 FCENet 在小样本训练时的良好泛化。此外，FCENet 在 CTW1500 上达到了 SOTA 性能，在 ICDAR2015 和 Total-Text 上也达到了接近 SOTA 的结果。

参考文献

[1] Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. Character region awareness for text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9365–9374, 2019.

[2] Chee Kheng Ch'ng and Chee Seng Chan. Total-text: A comprehensive dataset for scene text detection and recognition. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, pages 935–942. IEEE, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018.

[5] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11474–11481, 2020.

[6] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[7] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[8] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognition*, 90:337–345, 2019.

[9] Shangbang Long, Jiaqiang Ruan, Wenjie Zhang, Xin He, Wenhao Wu, and Cong Yao. Textsnake: A flexible representation for detecting text of arbitrary shapes. In *Proceedings of the European conference on computer vision (ECCV)*, pages 20–36, 2018.

[10] Chixiang Ma, Lei Sun, Zhuoyao Zhong, and Qiang Huo. Relatext: Exploiting visual relationships for arbitrary-shaped scene text detection with graph convolutional networks. *Pattern Recognition*, 111:107684, 2021.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[12] Baoguang Shi, Xiang Bai, and Serge Belongie. Detecting oriented text in natural images by linking segments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2550–2558, 2017.

[13] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[14] Jun Tang, Zhibo Yang, Yongpan Wang, Qi Zheng, Yongchao Xu, and Xiang Bai. Seglink++: Detecting dense and arbitrary-shaped scene text by instance-aware component grouping. *Pattern recognition*, 96:106954, 2019.

[15] Zhi Tian, Weilin Huang, Tong He, Pan He, and Yu Qiao. Detecting text in natural image with connectionist text proposal network. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 56–72. Springer, 2016.

[16] Zhuotao Tian, Michelle Shu, Pengyuan Lyu, Ruiyu Li, Chao Zhou, Xiaoyong Shen, and Jiaya Jia. Learning shape-aware embedding for scene text detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4234–4243, 2019.

[17] Fangfang Wang, Yifeng Chen, Fei Wu, and Xi Li. Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection. In *Proceedings of the 28th ACM international conference on multimedia*, pages 111–119, 2020.

[18] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9336–9345, 2019.

[19] Wenhai Wang, Enze Xie, Xiaoge Song, Yuhang Zang, Wenjia Wang, Tong Lu, Gang Yu, and Chunhua Shen. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8440–8449, 2019.

[20] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11753–11762, 2020.

[21] Enze Xie, Yuhang Zang, Shuai Shao, Gang Yu, Cong Yao, and Guangyao Li. Scene text detection with supervised pyramid context network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9038–9045, 2019.

[22] Yongchao Xu, Yukang Wang, Wei Zhou, Yongpan Wang, Zhibo Yang, and Xiang Bai. Textfield: Learning a deep direction field for irregular scene text detection. *IEEE Transactions on Image Processing*, 28(11):5566–5579, 2019.

[23] Chuhui Xue, Shijian Lu, and Wei Zhang. Msr: multi-scale shape regression for scene text detection. *arXiv preprint arXiv:1901.02596*, 2019.

[24] Cong Yao, Xiang Bai, Nong Sang, Xinyu Zhou, Shuchang Zhou, and Zhimin Cao. Scene text detection via holistic, multi-channel prediction. *arXiv preprint arXiv:1606.09002*, 2016.

[25] Chengquan Zhang, Borong Liang, Zuming Huang, Mengyi En, Junyu Han, Errui Ding, and Xinghao Ding. Look more than once: An accurate detector for text of arbitrary shapes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10552–10561, 2019.

[26] Shi-Xue Zhang, Xiaobin Zhu, Jie-Bo Hou, Chang Liu, Chun Yang, Hongfa Wang, and Xu-Cheng Yin. Deep relational reasoning graph network for arbitrary shape text detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9699–9708, 2020.

[27] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017.

[28] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019.