# OCR方法调研

## 一 OCR研究方向

1.两步走实现OCR (主流)

Step1文本检测：文字在哪里。

Step2文本识别：对定位好的文字区域识别，这个文字是什么。好的方法都是基于神经网络的实现。因为是有监督学习，数据质量影响模型的性能。

2.端到端的方法(效果不理想)

## 二 文本检测

### 基于回归

* 基于自顶向下的通用目标检测

• EdgeBox+CNN(Jaderberg, IJCV 2016)——[Reading Text in the Wild with Convolutional Neural Networks](https://www.robots.ox.ac.uk/~vgg/publications/2016/Jaderberg16/jaderberg16.pdf)

1)将文字的识别 ，转变成为分类问题。

2)步骤分为两部分，基于 region proposal mechanism 的检测文字部分，以及基于 CNN 的文字识别部分。在第一部分的区域候选框，首先保证较高的召回率，尽量把所有可能的文字区域先检测出来；然后经过一个 filtering 阶段，保证文字检测的精度。在第二部分的单词识别，不同于传统的单个字母的识别，其采用 CNN 去一次性的识别出单词。

• FCRN (Gupta, CVPR 2016)——[Synthetic data for text localisation in natural images](https://arxiv.org/pdf/1604.06646.pdf)

1)将word人工的嵌入到自然图片中，人工生成带有文本的图片（synthText）。

2)提出一种FRCN（ Fully-Convolutional Regression Network）的网络来检测文本：Step 1.生成目标候选框Step 2.提取每个候选的CNN特征图 Step 3.使用 class specific SVMs 过滤 proposals

• TextBox (Liao AAAI 2017; TIP 2018)——[TextBoxes：A Fast Text Detector with a Single Deep Neural Network](https://arxiv.org/abs/1611.06779)

1)改进版的SSD用来解决文字检测问题

2)端到端识别的pipeline:Step 1图像输入到修改版SSD网络中 + 非极大值抑制（NMS）→ 输出候选检测结果Step 2候选检测结果 + CRNN进行单词识别 → 新的检测结果 + 识别结果

• TextBox++ (Liao, TIP 2018)——[TextBoxes++: A Single-Shot Oriented Scene Text](https://arxiv.org/abs/1801.02765)

1. 扩展在TextBoxes只可以检测水平文本，TextBoxes++可以检测多角度的文本
2. 该方法也是基于SSD结构且是针对词来检测的

• EAST (Zhou, CVPR 2017)——[EAST: An Efficient and Accurate Scene Text Detector](https://arxiv.org/abs/1704.03155)

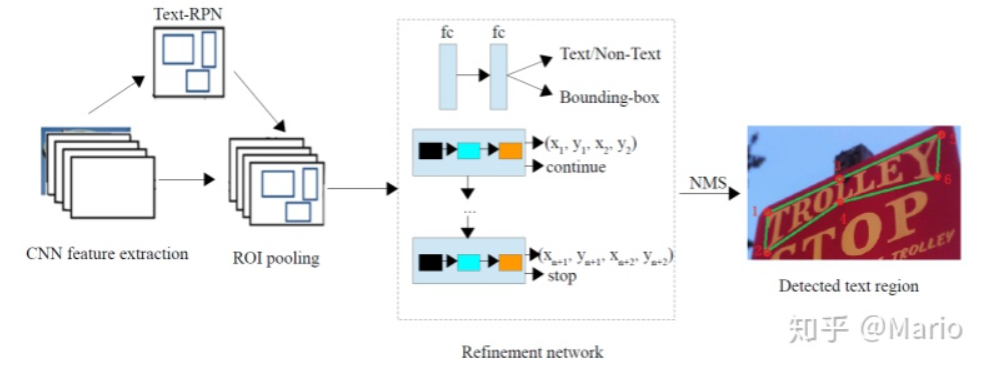
1)利用多层卷积神经网络(ResNet/MobileNet)提取图像特征，再利用该特征分别进行两个任务，像素点的分类，对应像素点的框的回归。最后将两个任务结果结合起来，并用非极大值抑制NMS来得到最终检测结果。

• DMPNet, (Liu, CVPR 2017)——[Deep Matching Prior Network: Toward Tighter Multi-oriented Text Detection](http://arxiv.org/abs/1703.01425)

一般情况下，对文本的检测都是使用矩形框，但是这会带来背景噪声，产生不必要的重叠甚至信息的丢失。为此，论文提出一种基于CNN的方法，使用更紧密的四边形来对文字进行检测。首先，论文使用四边形滑动窗口来对文本进行粗略的检测，并提出一种共享的蒙特卡洛方法来对多边形面积进行快速而精确的计算，然后基于一种序贯协议确定四个点的顺序，并回归最终的四边形预测结果。

• ASSTD (X. Wang, et.al, PR2019)——[Arbitrary Shape Scene Text Detection with Adaptive Text Region Representation.](http://ieeexplore.ieee.org/document/8953907)

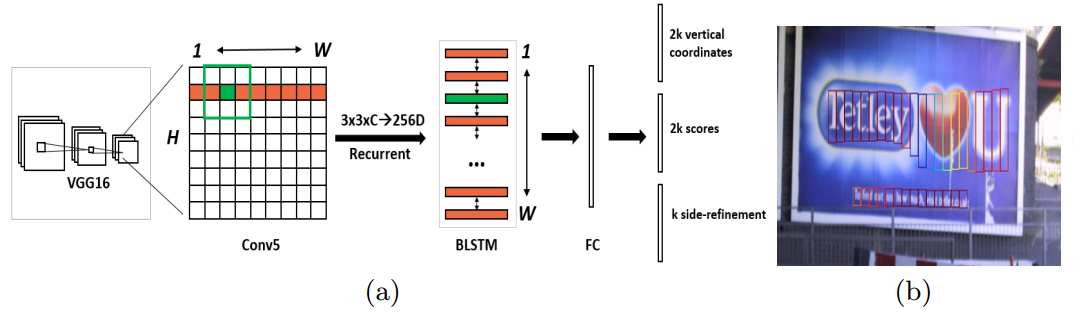
1. 主要解决的问题是任意形状的文本检测



1. 给定输入图像，首先利用一个文本区域候选网络Text-RPN（backbone是SE-VGG16）来获取文本候选区域（text proposal），同时利用CNN获得输入图像的feature map，通过使用ROI pooling结合feature map得到改进网络（Refinement network）的输入。
2. 上步中得到的text proposal被改进网络（Refinement network）验证并修正。该网络存在三个分支，分别是文字/非文字分类、边界框修正和基于RNN的自适应文本区域表示。在RNN中，每个时间步长预测一对边界点，直到stop标签被预测到。通过这种方式，可以用自适应数量的边界点表示任意形状的文本区域。
3. 通过NMS得到最终的结果。

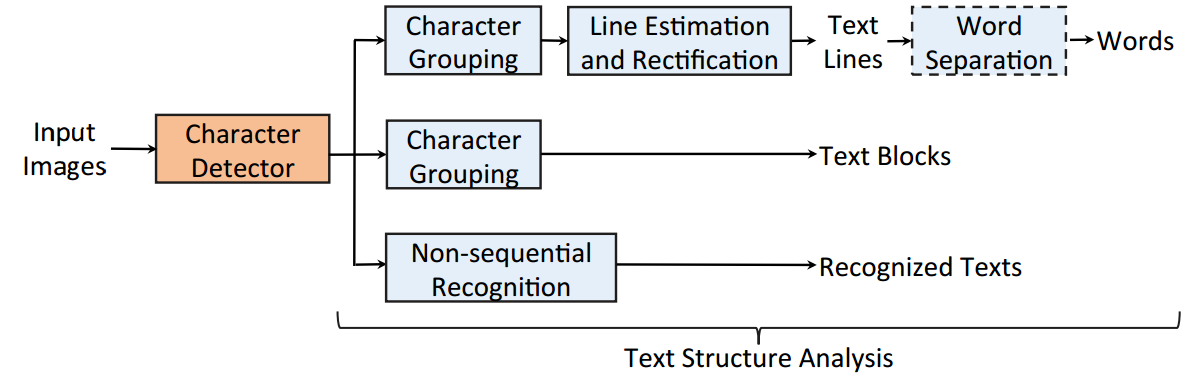
* 基于成分分析的自底向上的检测

•CTPN (Tian, ECCV 2016)——[Detecting Text in Natural Image with Connectionist Text Proposal Network](https://arxiv.org/abs/1609.03605)



这篇文章的方法最大亮点在于把RNN引入检测问题（以前一般做识别）。文本检测，先用CNN得到深度特征，然后用固定宽度的anchor来检测text proposal（文本线的一部分），并把同一行anchor对应的特征串成序列，输入到RNN中，最后用全连接层来分类或回归，并将正确的text proposal进行合并成文本线。这种把RNN和CNN无缝结合的方法提高了检测精度。

• WordSup (Hu, ICCV 2017)——[Exploiting Word Annotations for Character based Text Detection](https://openaccess.thecvf.com/content_ICCV_2017/papers/Hu_WordSup_Exploiting_Word_ICCV_2017_paper.pdf)



采用单词、文本行的标注信息进行监督学习来辅助字符检测；在ICDAR2013数据集上F值90+，后来的方法能超过这篇文章的寥寥无几。

利用Faster RCNN或SSD等目标检测框架检测单个字符，并利用单词、文本行的标注信息进行监督学习方法概述：来辅助字符检测器的训练（字符模型和mask进行交替训练），最后采用图模型来进行文本行合并、LSTM空格检测来进行单词切割等后处理。

文章出发点：各式各样的文字（中英、曲线文本、公式）等排列布局不同，但最小单元都是字符。所以，检测字符具有通用性。但目前有字符的标注数据集太少了，大部分都是以单词、文本行的方式标注的。故这篇文章想要利用这些单词、文本行的标注来帮助字符检测。主要是利用了已标注信息进行监督训练。

•SegLink (Shi, CVPR 2017)——[Detecting Oriented Text in Natural Images by Linking Segments](https://arxiv.org/abs/1703.06520)

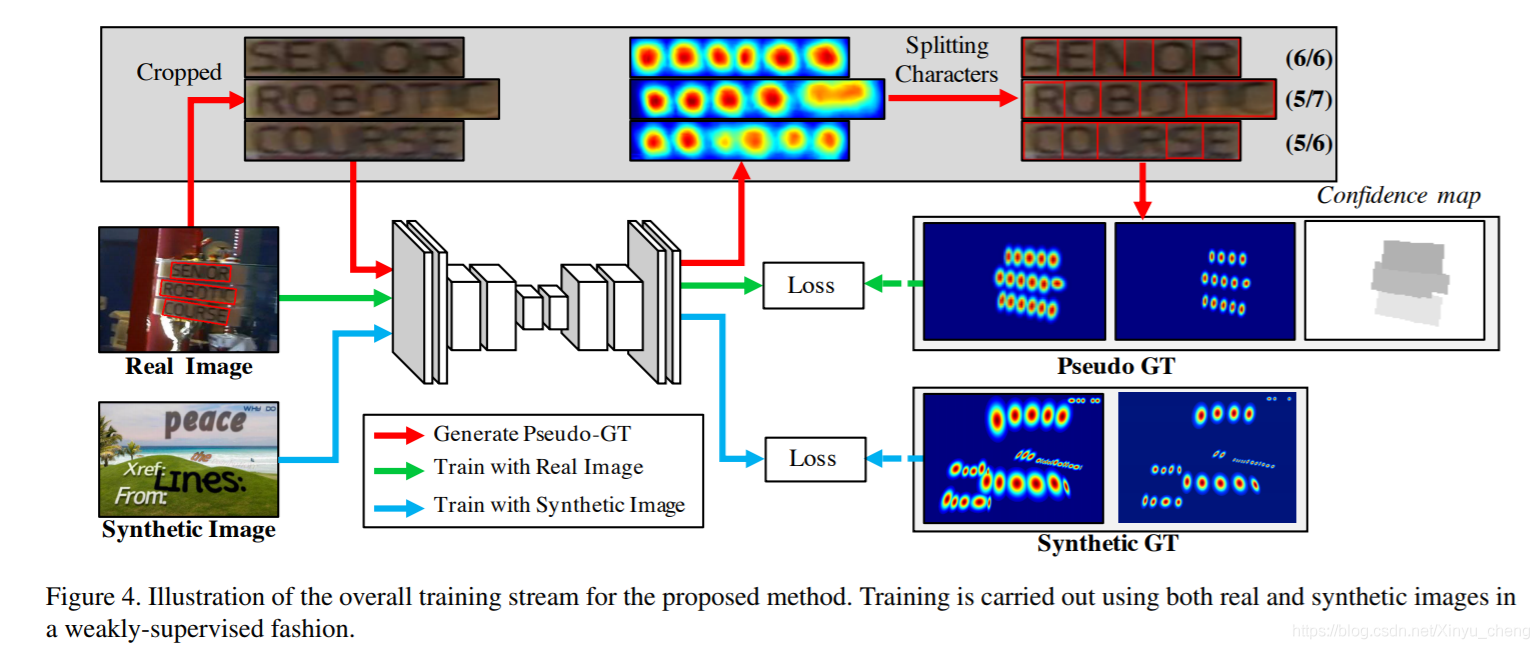
本文针对文字检测的特定环境改进SSD算法，提出Seglink多方向文本检测方法。核心是将文本检测转换成两个局部元素的检测：segment和link。segment 是一个有方向的box，覆盖文本内容的一部分，而link则连接了两个相邻的segments，表达了这两个segment是否属于同一个文本。本文算法通过在多尺度上进行segment和link的检测，最终按照links的表达将相关的segment合并成最终的bounding box。经实验证明，该算法提高了检测正确率及训练效率

# • SegLink++ (Tang, PR 2019)——[Detecting Dense and Arbitrary-shaped Scene Text by Instance-aware Component Grouping](https://www.sciencedirect.com/science/article/abs/pii/S0031320319302511?via%3Dihub)



该论文首先提出了一种文本块单元之间的吸引关系和排斥关系的表征，然后设计了一种最小生成树算法进行单元组合得到最终的文本检测框。另外，该论文还设计了Instance-aware Loss损失函数，把文本行实例的信息引入到文本块单元的训练过程，实现了后处理过程和网络端到端的训练。

•CRAFT (Baek, CVPR 2019)——[Character Region Awareness for Text Detection](https://arxiv.org/pdf/1904.01941.pdf)



1）针对基于bounding box的检测方法在检测弯曲文本上效果不行，提出了一种新的用于检测弯曲文字的方法。

2）所提出的方法可以做到字符级检测，即构建了一个专用于字符检测的网络(VGG16骨干)，而不是单词。人造数据和真实数据集。真实图片没有字符级别的标注时，自己检测合成产生标签再进行训练。如上图所示，对真实场景中的数据集和人造合成的数据集有不同的训练方式。有单词标签时，通过网络得到区域分数预测，并根据区域得分划分字符。

• RelationNet (Ma, ICDAR 2019);

•DirectReg (He, ICCV 2017, TIP 2018);

•LOMO (Zhang, CVPR 2019)——[Look More Than Once： An Accurate Detector for Text of Arbitrary Shapes](https://arxiv.org/abs/1904.06535)

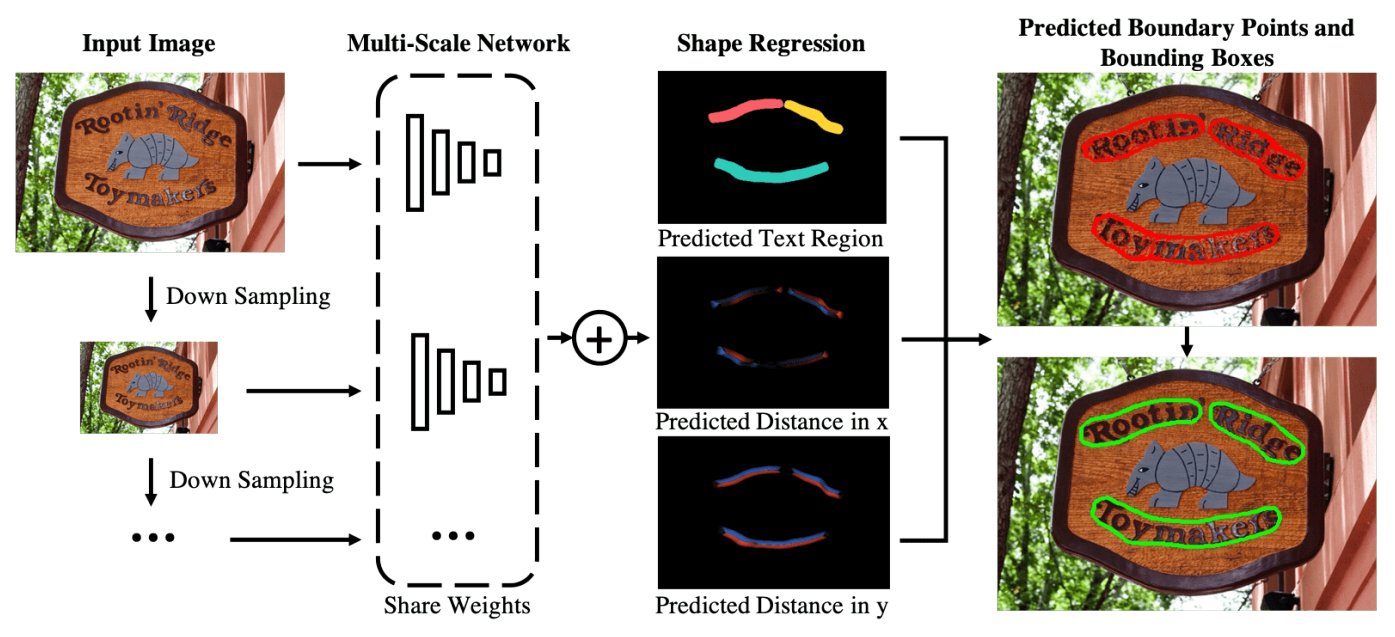
1. LOMO可以逐步地调整文本（或者换句话说，不止看一次）。LOMO由直接回归器（DR）、迭代细化模块（IRM）和形状表示模块（SEM）组成。首先，DR分支生成四边形的文本proposals。接下来，IRM基于提取的初步proposals的特征块，通过迭代细化逐步感知整个长文本。最后，通过考虑文本实例的几何属性，包括文本区域、文本中心线和边界偏移，引入SEM来得到更加精确的不规则的多边形的文本表示
2. 论文主要贡献：

提出了一个迭代细化模块(**IRM**)，它改善了**长文本**检测的性能。

引入实例级形状表示模块(**SEM**)，用于解决**任意形状的场景文本**的检测问题。

加入**迭代细化模块**和**形状表示模块**的LOMO可以以端到端的方式训练，并在几个包括不同形式的文本实例（**有向的、长的、多语言的和弯曲的**）的基准数据集上实现了最佳的性能。

•MSR (Xue, IJCAI 2019)——[Multi-Scale Shape Regression for Scene Text Detection](https://www.ijcai.org/Proceedings/2019/139)

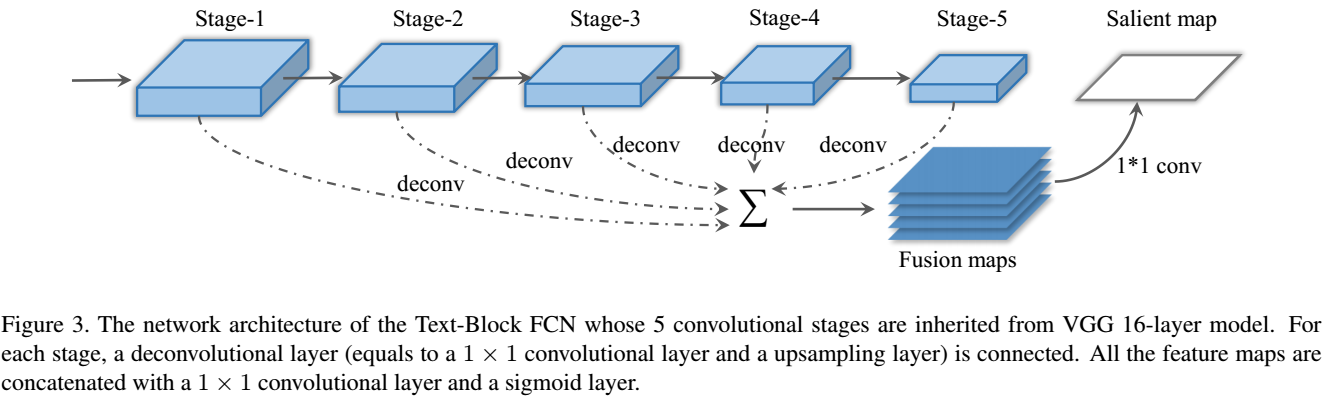


1. 作者认为目前所存的文本检测方法存在定位不精确和检测不可靠的弊端，因此他们提出了多尺度(Multi-scale)-外形回归(Shape Regression)模型MSR.与EAST一脉相承，MSR也属于场景文本检测中的回归方法，不过它回归的是文本中心区域像素点距离最近的边框点的XY距离，而EAST回归的是文本中心区域像素点距离上下左右四边界的XY距离，因此MSR的回归方式可有效地避免在长文本中卷积层的感受野受限的缺点。此外，MSR的另一贡献点为提出了Multi-scale NetWork,既可以预测不同层级的细节，又可以提取图片中不同尺度的文本特征.
2. 亮点：（1）multi-scale网络中利用FPN的up-sampling把多个不同scale得到的结果进行融合；（2）boundary-point regression部分直接预测点与最近的boundary point的dx和dy，思路清晰易实现。

• GCN+Relation Reason (Zhang, CVPR 2020);

### 基于分割

• FCN based (Zhang CVPR 2016)——[Multi-Oriented Text Detection with Fully Convolutional Networks](https://openaccess.thecvf.com/content_cvpr_2016/papers/Zhang_Multi-Oriented_Text_Detection_CVPR_2016_paper.pdf)



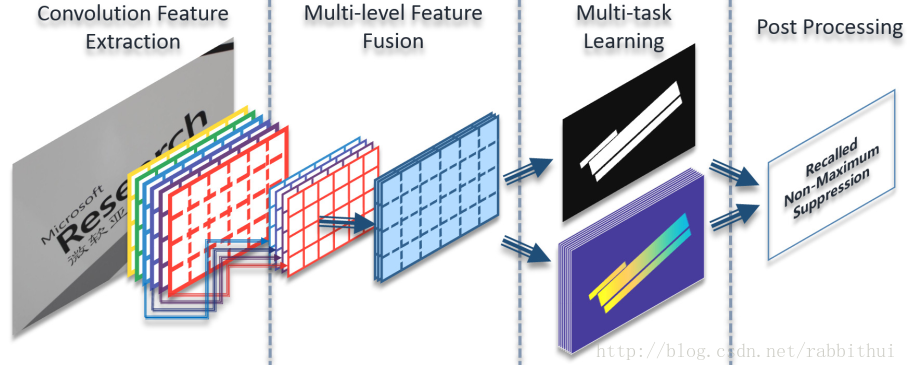
方法：

1. **文本块检测:** 先利用**text-block FCN**得到salient map，再对salient map进行连通分量分析得到text block；
2. **文本线形成:** 对text-block利用MSER提取候选字符区域，利用候选字符区域估计整个block的方向，再结合候选字符的bounding box生成每个每条文本行；
3. **文本线过滤:** 利用**centroid FCN**得到每条文本行中的字符的质心，利用质心过滤非文本行；

贡献：

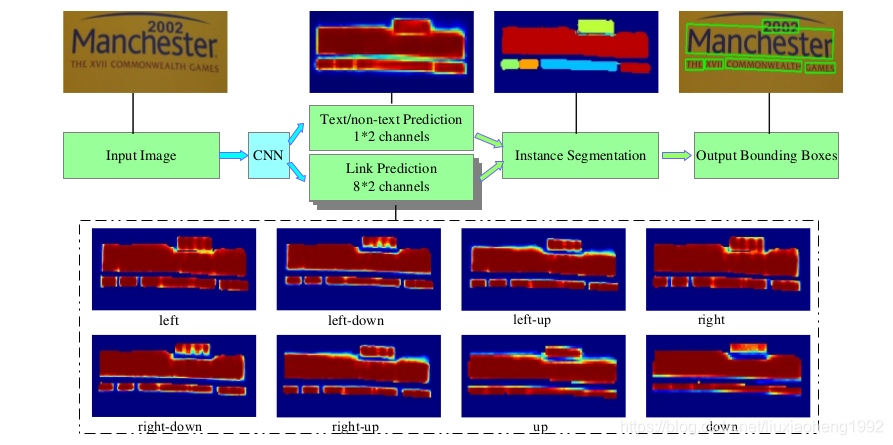
1. 用FCN生成文字/非文字的salient map；
2. 利用局部（Component based )和全局（context of the text block ) 的信息生成文本线；
3. 用FCN得到字符的质心；

• DDR (He, ICCV 2017)——[Deep Direct Regression for Multi-Oriented Scene Text Detection](https://arxiv.org/pdf/1703.08289.pdf)



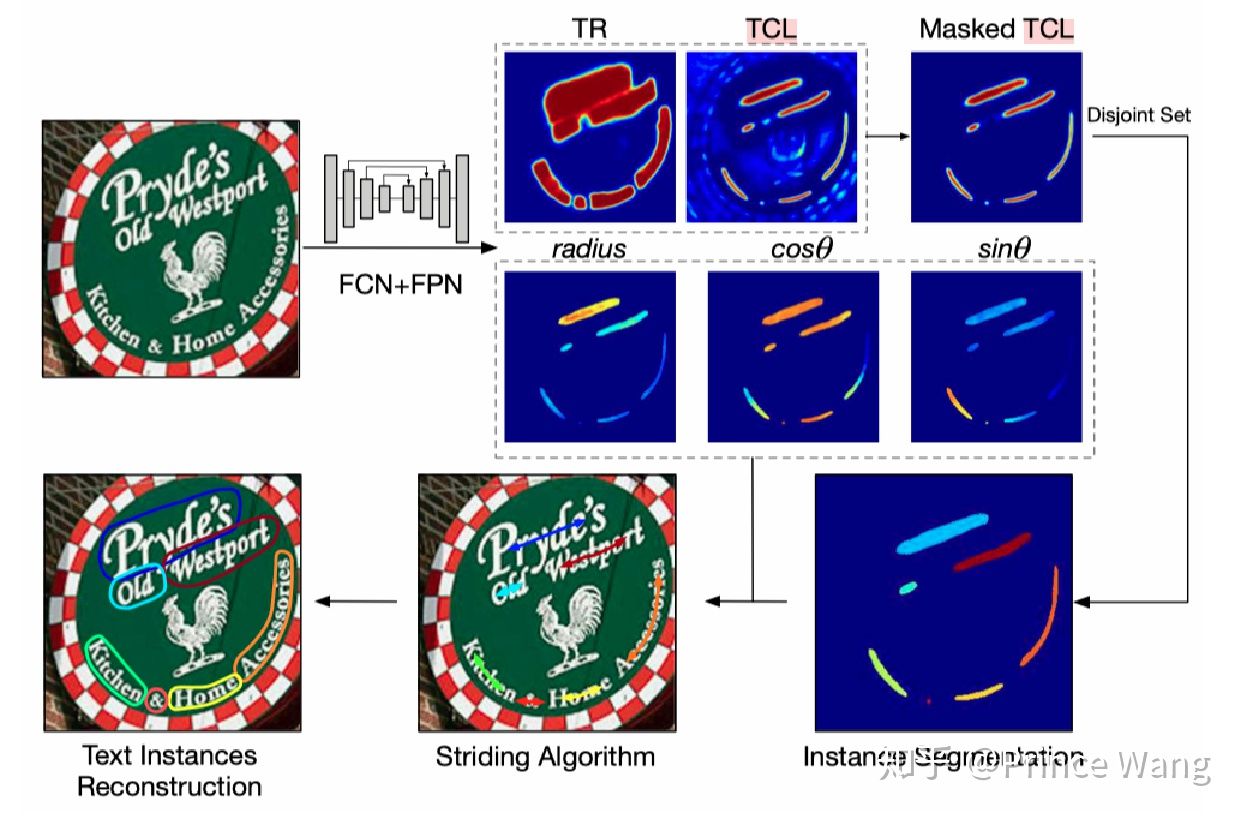
1. 特征提取：在多个卷积层进行特征提取，增加了最大感受野的范围。确保回归任务可以检测出更长的文本，同时提高边界预测的准确度。
2. 特征融合：从多个尺度的feature map上提取特征，再进行融合。为了减少计算量，文章仅将融合特征map上采样至原图像尺寸的1/4，具体见下图所示（在原论文图基础上做了尺寸的标注）。
3. 多任务学习：主要有两个任务的学习，分类任务和回归任务。分类任务的输出MclsMcls是一个大小为S4×S4S4×S4的2维张量，用来预测是否为文本，张量内的元素分数越高是文本的可能性就越大；回归任务的输出MlocMloc是一个大小为S4×S4×8S4×S4×8的3维张量，MlocMloc中位置(w,h,c)(w,h,c)的值L(w,h,c)L(w,h,c)表示原始图像中点(4w,4h)(4w,4h)到文本bounding box的4个顶点的offsets。

• PixelLink (Deng, AAAI 2018)——[Detecting Scene Text via Instance Segmentation](https://arxiv.org/abs/1801.01315)



1. 使用Instance Segmentation的思路（利用了二分类+每个像素点与周围8个顶点的link分类
2. 边界框不需要通过回归进行预测，直接通过positive pixel组成的Connected Components (CC)通过类似minAreaRect的方法OpenCV实现
3. 利用图像处理的方法进行后处理，不需要NMS

• TextSnake (Long, ECCV 2018)——[A Flexible Representation for Detecting Text of Arbitrary Shapes](https://arxiv.org/abs/1807.01544)



主要提出了一种能够灵活表示任意弯曲形状文字的数据结构——TextSnake，主要思想是使用多个不同大小，带有方向的圆盘(disk)对标注文字进行覆盖，并使用FCN来预测圆盘的中心坐标，大小和方向进而预测出场景中的文字，本文的主要贡献有以下几点：

1. 提出一种基于多圆盘覆盖来表示文字的方法——TextSnake
2. 基于TextSnake表示，提出一种检测场景文本的方法
3. 在两个新提出的数据集（Total-Text and SCUT-CTW1500）上，获得了state-of-the-art的效果，这两个新的数据集都包含曲线文本，波浪形文本等非常规标注

• Border Learning (Wu ICCV 2017)——[Self-organized Text Detection with Minimal Post-processing via Border Learning](https://openaccess.thecvf.com/content_ICCV_2017/papers/Wu_Self-Organized_Text_Detection_ICCV_2017_paper.pdf)

针对文字检测中提取文字的行级候选区需要大量后处理、复杂且耗时的问题，提出了一种基于边界学习的场景文字检测方法（只需要极少后处理、可以区分任意形状的文本区）。

文章的主要贡献有：

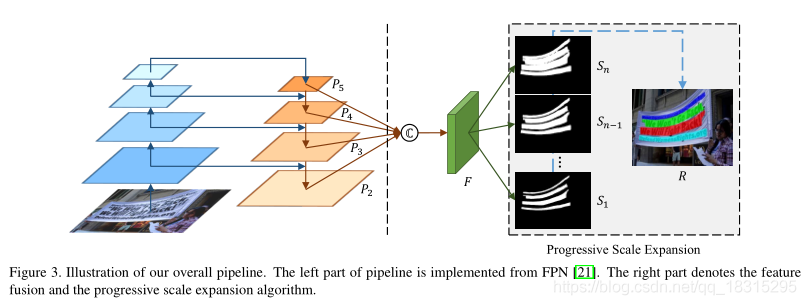
1. 分析了将文字检测作为二分类问题（文字/非文字）的局限性；
2. 首次将边界类别引入到文本检测问题中，验证了文本边界在很大程度上简化了解码过程；
3. 建立了一个新的用于文本检测的ppt数据集（10692幅图像，非文本、边界和文本标注）；
4. 提出了一个轻量级FCN(0.28M参数，支持多分辨率分析)来有效地学习文本图像中的边界。

• ShapeAware (Tian, CVPR 2019)——[Learning Shape-Aware Embedding for Scene Text Detection](https://openaccess.thecvf.com/content_CVPR_2019/papers/Tian_Learning_Shape-Aware_Embedding_for_Scene_Text_Detection_CVPR_2019_paper.pdf)



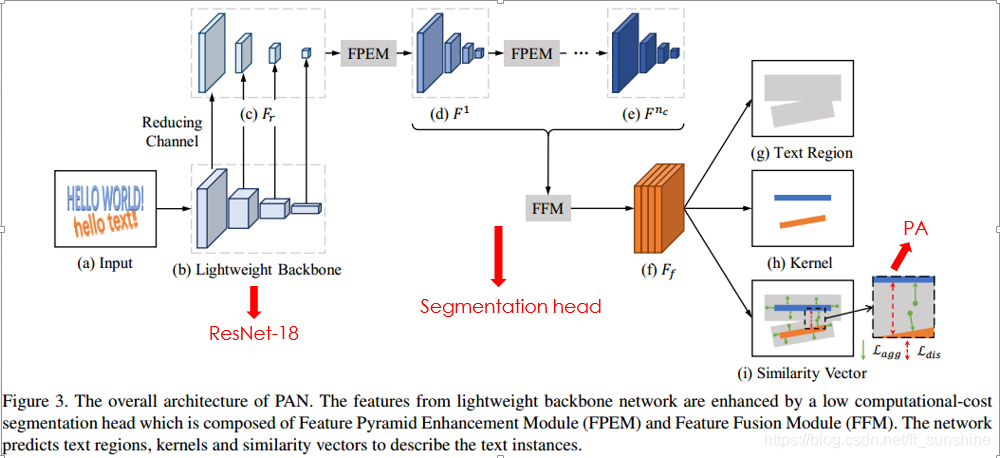
1. 提出了Shape-Aware Loss，可以很好的分割相邻的文本实例并且可以检测很大很长的文本实例（将图像像素映射到嵌入特征空间中，在该空间中，属于同一文本实例的像素会更加接近彼此，反之不同文本实例的像素将会远离彼此）。
2. 提出了一种新的文本检测流程，可以检测任意形状的文本实例。
3. 在三个有代表性的场景文本数据集（ICDAR15,，MSRA-TD500，and CTW1500）上表现优异。

• PSENet (Wang, CVPR 2019)——[Shape Robust Text Detection with Progressive Scale Expansion Network](https://arxiv.org/abs/1806.02559)



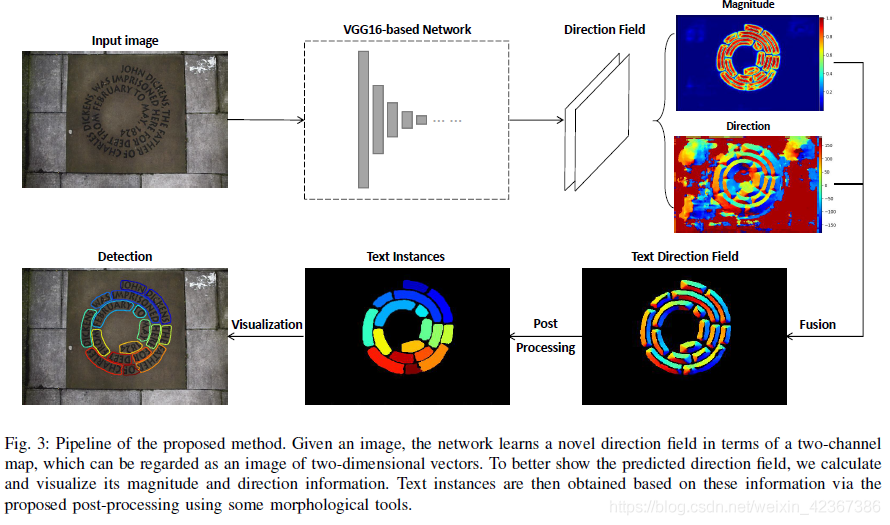
文章主要思想：PSENet是一种基于语义分割的方法用于检测任意方向的文本，采用了渐进式尺度扩展的方法用来区分邻近的文本块。首先，能够对任意形状的文本进行定位；其次，提出了一种渐进的尺度扩展算法，该算法可以成功地识别相邻文本块。即本文算法可以精确地检测任意形状的文本，并且准确地将文本实例紧密地分开。

• PAN (Wang, ICCV 2019)——[Efficient and Accurate Arbitrary-Shaped Text Detection with Pixel Aggregation Network](https://openaccess.thecvf.com/content_ICCV_2019/papers/Wang_Efficient_and_Accurate_Arbitrary-Shaped_Text_Detection_With_Pixel_Aggregation_Network_ICCV_2019_paper.pdf)



通过引入一个由特征金字塔模块与特征融合模块组成的轻量级分割 head，它在提升特征提取能力的同时可以降低计算量；此外，研究人员还提出了像素聚合算法，来预测文字的核与周围像素间的相似向量。以上两个优势使得 PAN 成为了一个高效且精准的任意形状文字检测器。在 Total-Text 和 CTW1500 数据集上实验显示，与当前已提出的其它最先进模型相比，PAN 在速度与精度上都更加优秀。

• TextField (Xu, TIP 2019)——[Learning A Deep Direction Field for Irregular Scene Text Detection](https://arxiv.org/pdf/1812.01393.pdf)



亮点：提出的TextField方法非常新颖，用点到最近boundary点的向量来区分不同instance，旨在分割文本并分离相邻的文本实例

在本文中，采用Instance-segmentation思路，提出了一种名为TextField的新型文本检测器，用于检测不规则场景文本。

具体来说，我们用一个VGG+FPN网络学习一个方向场TextField，该方向场由二维矢量的图像表示。它对二进制文本mask和用于分隔相邻文本实例的方向信息进行编码。

基于网络学习得到的方向场的两张score map图，我们应用基于形态学的后处理来实现最终检测。

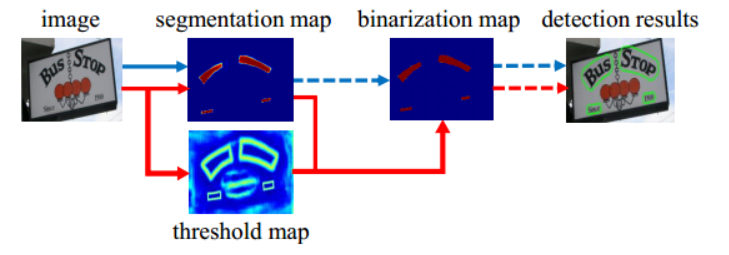
• SAST (Wang, MM 2019)——[A Single-Shot Arbitrarily-Shaped Text Detector based on Context Attended Multi-Task Learning](https://arxiv.org/pdf/1908.05498.pdf)

本文提出了一种新的基于分割的文本检测器SAST。它利用基于全卷积网络（FCN）的上下文多任务学习框架来学习文本区域的各种几何特征，从而构造文本区域的多边形表示。考虑到文本的连续性特征，通过引入Context Attention Block 捕捉像素的长范围相关性，一次来获得更加可靠的分割结果。在后处理过程中，提出一个点到边的对齐方法，来将像素聚类称为文本实力，这样就通过一次采样图片，把高级别的特征和低级别的特征结合在一起。此外，利用所提出的几何性质可以更有效地提取任意形状文本的多边形表示。

• MCN (Z Liu, IJCV 2020);

• LOMT (Wang, CVPR 2020)

• DB (Liao, AAAI 2020)——[Real-time Scene Text Detection with Differentiable Binarization](https://arxiv.org/abs/1911.08947)

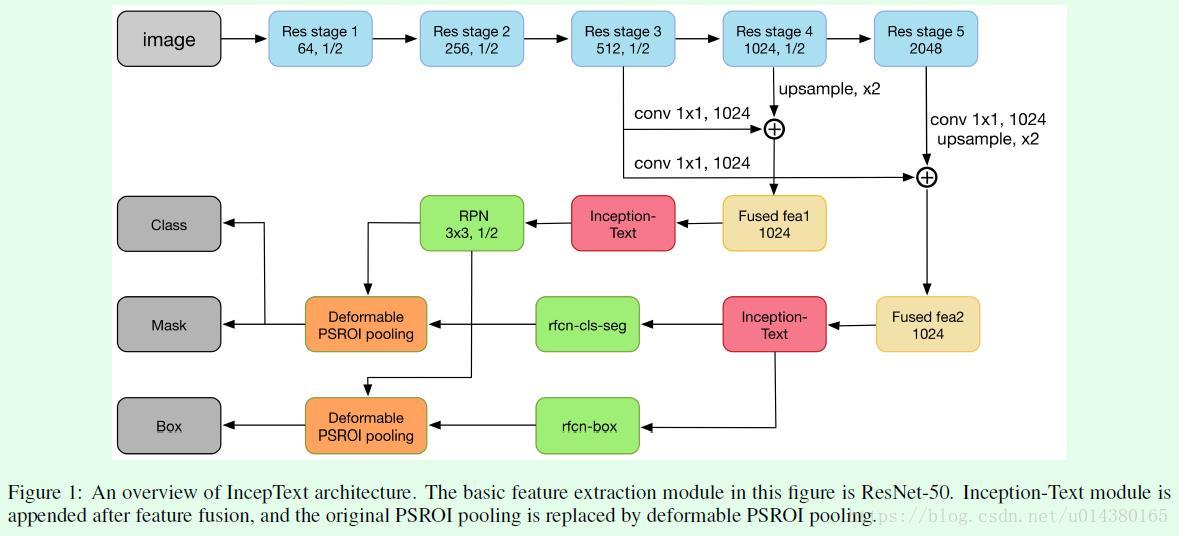


相较于传统方法中得到分割结果之后采用一个固定的阈值得到二值化的分割图。

DB在阈值选取上，通过网络去预测图片每个位置处的阈值，而不是采用一个固定的值，这样就可以很好将背景与前景分离出来。但这会给训练带来梯度不可微的情况，对此提出Differentiable Binarization来解决不可微的问题。

### 回归和分割融合

• IncepText (Yang, IJCAI 2018)——[A New Inception-Text Module with Deformable PSROI Pooling for Multi-Oriented Scene Text Detection](https://arxiv.org/abs/1805.01167)



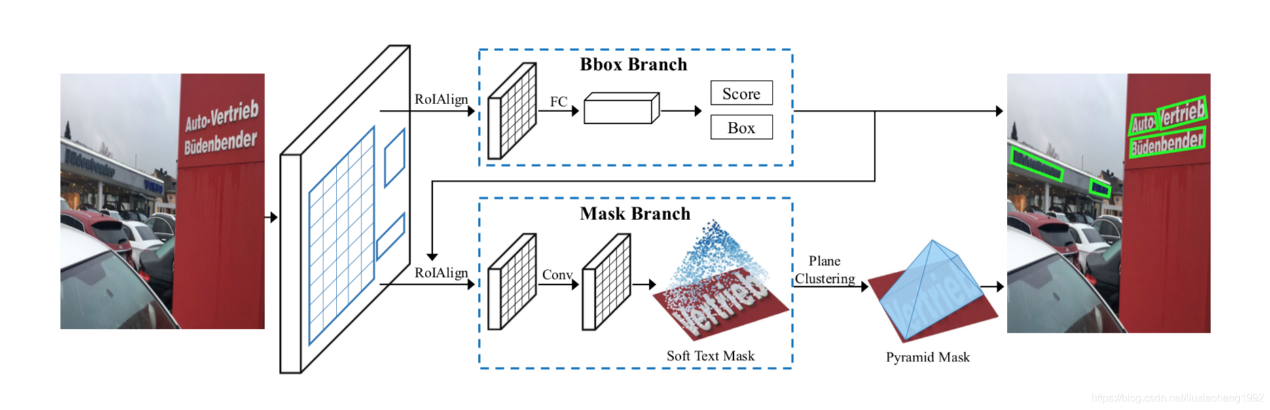
本研究的主要贡献如下：

1. 我们提出了一个用于多方向场景文本检测的新的 Inception-Text 模块。实验表明，该模块可以大幅提高准确率，且计算成本很低。
2. 我们提出使用可变形的 PSROI 池化模块来处理多方向的文本。在对可变形 PSROI 池化和量化评估学习到的偏移部分进行的量化研究表明，它可以有效处理任意方向的场景文本。
3. 我们在三个基准数据集 ICDAR2015，RCTW-17 和 MSRA-TD500 上评估了新方法，结果表明新的方法在没有任何额外数据的情况下，在几个基准测试中都取得了业内最优的表现。

• FTSN (Dai, ICPR 2018)——[Fused Text Segmentation Networks for Multi-oriented Scene Text Detection](https://arxiv.org/abs/1709.03272v1)

本文我们介绍了一种新的端对端框架，基于instance-aware视角的分割来解决多朝向场景文本检测。我们提出fused text segmentation 网络，该网络在特征提取阶段结合了多层次特征，因为文本instance与通用目标检测相比依赖于更细的特征表达。该网络将检测和分割文本instance联合一起做，利用基于语义分割和region proposal的目标检测任务两者的优点。不包含任何其他的流程，我们的方法成为多朝向场景文本检测（ICDAR2015随机场景文本和MSRA-TD500）的最佳，hmean值分别达到84.1%和82%。这说明了该方法的有效性。

• PMTD (arXiv 2019)——[Pyramid Mask Text Detector](https://arxiv.org/abs/1903.11800)



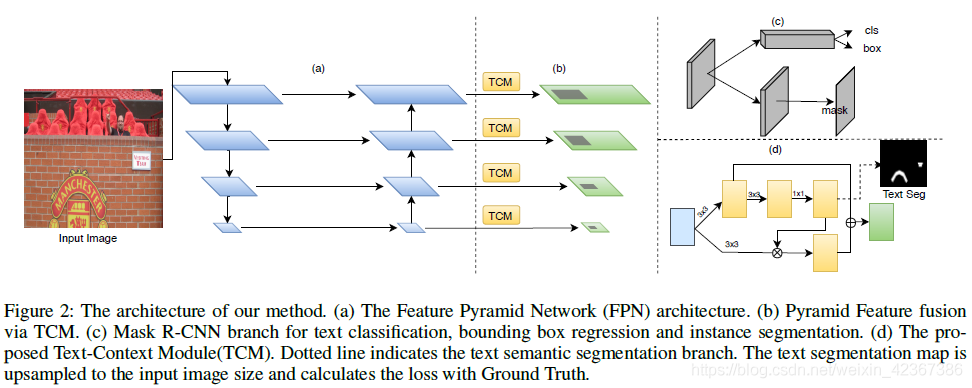
迭代过程为：

1. 将网络输出预测图，将大于0.1的像素看成是正样本，这些正样本组成一个集合称为P。将P的中心点作为金字塔的顶点，并设其值为1。预测出的boundingbox的四个顶点作为金字塔底的四个点
2. 经过上面步骤就可以得到初始的四个金字塔平面，然后将预测出来的点分配给离点最近的平面
3. 所以点分配完成后采用最小二乘法更新平面的位置
4. 回到第一步进行迭代

迭代的结束条件为最小二乘返回一个很小的值（文章设为1e-4）或者到达最大的迭代次数（文章设为10），这样就得到最终的文本框

• SBD (Liu, IJCAI 2019)

• SPCNet (Xie, AAAI 2019)——[Scene Text Detection with Supervised Pyramid Context Network](https://arxiv.org/abs/1811.08605)



SPCNet是一个曲文检测算法，采用Instance-segmentation的思路，在MaskR-CNN的基础上，提出并添加了一个全局文本分割分支，完成曲文检测。

创新点：

1. 针对误检的问题，提出了**TCM模块**和**Re-Score机制**；
2. SPCNet可以灵活地检测各种形状的文本。

• DirectReg+Pos.-Sensitie Segment (P Cheng, IEEE TCSVT 2020)

• ContourNet (Wang, CVPR 2020)——[Taking a Further Step toward Accurate Arbitrary-shaped Scene Text Detection](https://openaccess.thecvf.com/content_CVPR_2020/papers/Wang_ContourNet_Taking_a_Further_Step_Toward_Accurate_Arbitrary-Shaped_Scene_Text_CVPR_2020_paper.pdf)

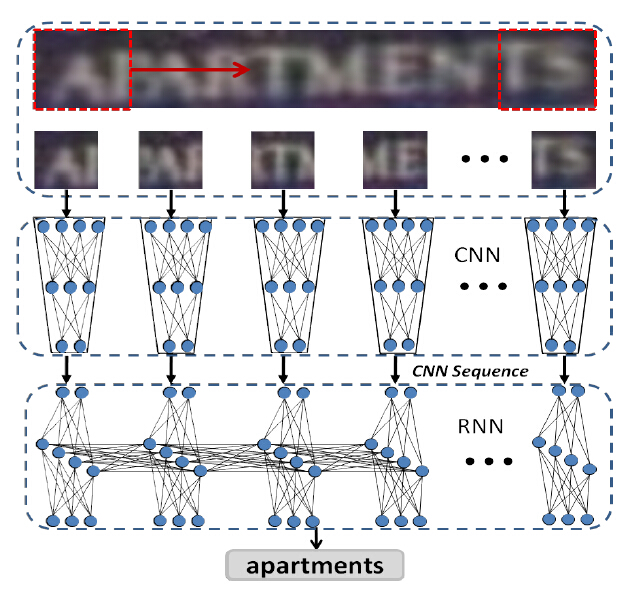
场景文字图片中，文本尺度变化大，背景纹理复杂等困难导致了现有的场景文本检测算法普遍存在两个问题，伪召回（False Positives）多，以及对尺度变化剧烈的文本检测不准确。针对伪召回问题，现有的方法大多基于文本的上下文信息，全局与局部关系，以及纹理信息建模。本文作者观察发现，伪召回中的纹理特征具有强烈的无方向性，即类似物理上的各向同性，于是将其解耦为水平方向和竖直方向的轮廓检测，通过其联合响应来抑制伪召回。针对尺度变换剧烈问题，现有的方法多着眼于融合多尺度特征来预测边界框，作者则另辟蹊径关注文本自身形状信息的建模以及利用尺度不敏感的指标来优化网络的检测。

## 三 文本识别

### 基于CTC

[CTC详解](https://zhuanlan.zhihu.com/p/42719047)

• P. He et al., AAAI 2016 (DTRN: CNN+RNN+CTC)——[Reading Scene Text in Deep Convolutional Sequences](https://arxiv.org/pdf/1506.04395v2.pdf)



方法：

提出了一个结合CNN与RNN的端到端DTRN系统：CNN将输入的单词图片编码为顺序序列，RNN将CNN序列译码（识别）为单词串

主要贡献：

1. 建立了一个统一deep recurrent system，同时利用CNN和RNN的优点来解决自然场景文本识别的问题；
2. 这是将卷积序列与系列标记模型相结合解决这个问题的第一次尝试；
3. 提出的DTRN在几个benchmark中，都获得了最佳的结果。

• B. Shi et al. , TPAMI 2017 (CRNN: CNN+RNN+CTC)——[End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition](https://arxiv.org/abs/1507.05717)

• F Yin, et al., arXiv 2017 (CNN+CTC); Y. Liu et al., ECCV 2018 （GAN+CTC)

• Borisyuk F et al., the 24th ACM SIGKDD International Conference. ACM, 2018.[Rosetta: Large scale system for text detection and recognition in images](https://arxiv.org/abs/1910.05085)

1. 文本检测方法基于 Faster-RCNN 模型 [24]，负责检测图像中包含文本的区域。
2. 文本识别方法使用全卷积字符识别模型，处理检测到的区域，并识别这些区域所包含的文本内容。

• H Li, W Wang, PR 2020

### 2.基于注意力机制

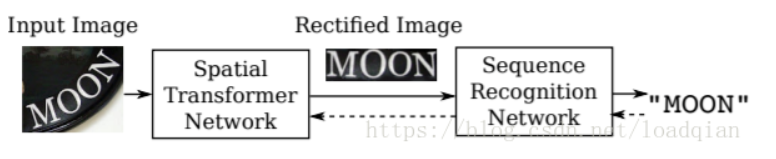
• C Lee et al., CVPR 2016

•Wei Liu et al. BMVC2016——[STAR-Net: A SpaTial Attention Residue Network for Scene Text Recognition](http://www.bmva.org/bmvc/2016/papers/paper043/index.html)

1)面向不规则文字的空间注意力机制

• B. Shi et al., CVPR 2016——[Robust Scene Text Recognition with Automatic Rectification](https://arxiv.org/abs/1603.03915v1)

1. 一个对于不规则的文字具有鲁棒性的识别模型。RARE是一个深度神经网络，包括一个空间变换网络Spatial Transformer Network (STN)和一个序列识别网络Sequence Recognition Network (SRN)，两个网络同时用BP算法进行训练。

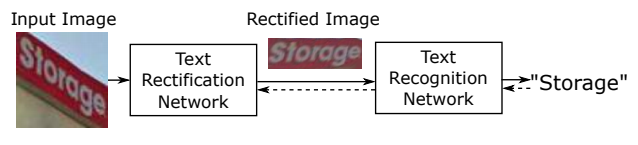


1. 一张图像先通过Thin-Plate-Spline (TPS)变换成一个正规的、更易读的图像，此变换可以矫正不同类型的不规则文本，包括透射变换和弯曲的文本。TPS变换由一组基准点（fiducial points）表示，坐标通过卷积神经网络回归得到。然后再放入SRN中进行识别。SRN使用序列识别的attention-based方法，包含一个编码器和一个解码器。编码器产生一个特征表示序列，即序列的特征向量；解码器根据输入序列循环地产生一个字符序列。这个系统是一个端到端的文本识别系统，在训练过程中也不需要额外标记字符串的关键点、字符位置等。

• X. Yang et al. IJCAI 2017;

• Bai et al. , CVPR 2018; Liu et al., AAI 2018

• Shi et al., TPAMI 2018 （ASTER)—— [An Attentional Scene Text Recognizer with Flexible Rectification](https://ieeexplore.ieee.org/abstract/document/8395027)

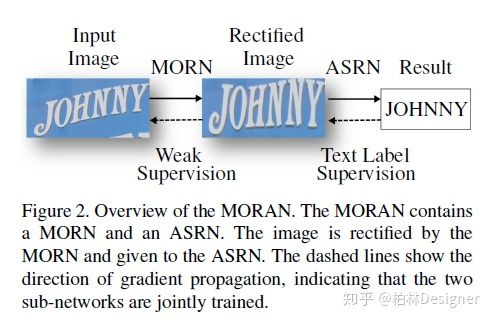


1. 针对不规则文字，先矫正成正常线性排列的文字，再识别；
2. 整合矫正网络和识别网络成为一个端到端网络来训练；
3. 矫正网络使用STN，识别网络用经典的sequence to sequence + attention

亮点：

1. 效果太好了，针对普通文字和不规则文字，尤其是不规则文字
2. 端到端训练 + 不需要人工标注（无需矫正网络的控制点）

• Luo et al, PR 2019 （MORAN)——[A Multi-Object Rectified Attention Network for scene text recognition](https://www.sciencedirect.com/science/article/abs/pii/S0031320319300263)



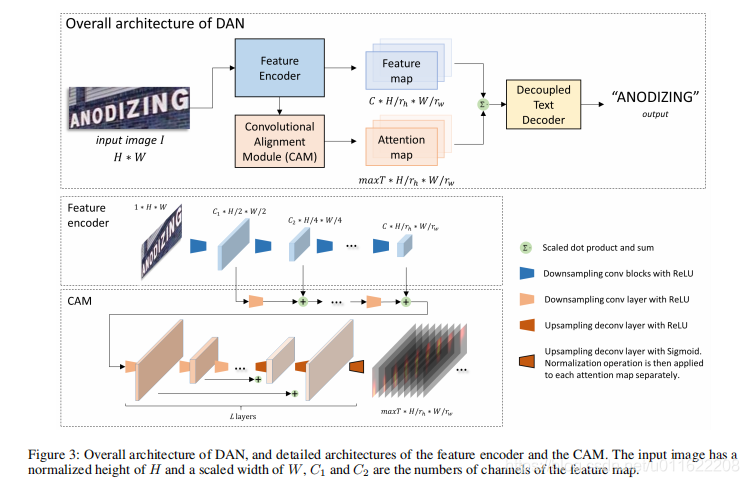
这篇论文提出了自然场景下弯曲文字识别的一种网络MORAN。MORAN 由两个部分组成：一个是弯曲矫正网络MORN，一个是识别网络ASRN。

几个创新点和论文贡献：

1. propose the MORAN framework to recognize irregular scene text.
2. Trained in a weak supervision way, the subnetwork MORN is flexible. It is free of geometric constraints and can rectify images with complicated distortion.
3. propose a fractional pickup method for the training of the attention-based decoder in the ASRN. To address noise perturbations, we expand the visual field of the MORAN, which further improves the sensitivity of the attentionbased decoder.

• Li et.al, AAAI 2019,

• T Wang et al., AAAI 2020 (DAN)——[Decoupled Attention Network for Text Recognition](https://arxiv.org/abs/1912.10205)



文字识别方法在不规则文本识别上，多采用attention的方式，但是采用自回归模型的attention容易存在attention对齐的问题。文章提出了一个 a decoupled attention network (DAN)网络。其主要由：1.特征采集器;2.一个卷积对齐模块，根据编码器的输出进行特征对齐；3.一种解耦的文本解码器，通过联合使用特征图和注意图进行最终预测

• Deli Yu et al. 2020——SRN:[Towards Accurate Scene Text Recognition with Semantic Reasoning Networks](https://arxiv.org/abs/2003.12294)

1. 语义推理网络(SRN):骨干网ResNet、FPN提取视觉特征

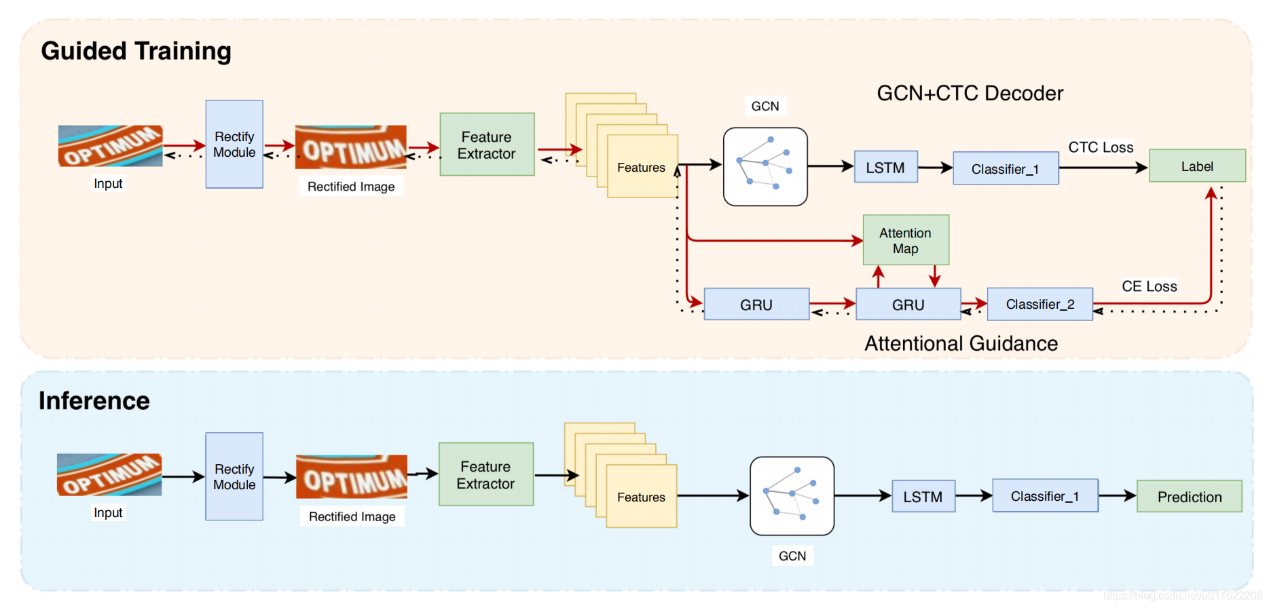
+并行视觉注意模块(PVAM) 使用self attention提取每个时间步的视觉特征

+全局语义推理(GSRM) 编码全局信息

+视语义融合解码器(VSFD)融合视觉信息和语义信息，获得最终结果

### 3.CTC和注意力机制融合

• W Hu et al., (GTC), AAAI 2020——[Guided Training of CTC towards Efficient and Accurate Scene Text Recognition](https://www.aaai.org/ojs/index.php/AAAI/article/view/6735)



解决的问题：

(1) 字符位置的不确定导致CNN学习的特征差强人意

(2) 未考虑特征的局部关联导致解码错误

• R Litmon et al., (SCATTER), CVPR 2020——[Selective Context Attentional Scene Text Recognizer](https://openaccess.thecvf.com/content_CVPR_2020/papers/Litman_SCATTER_Selective_Context_Attentional_Scene_Text_Recognizer_CVPR_2020_paper.pdf)