同濟大學

TONGJI UNIVERSITY

计算机视觉课程期中论文

学　　院　计算机科学与技术学院

专　　业　计算机科学与技术

学生姓名　李闯

学　　号　2253214

指导教师　赵才荣

日　　期　2025 年 4 月 19 日

目 录

装 订 线

1 引言

1.1 研究背景

场景文字识别与检测（Scene Text Detection and Recognition, STD/STR）作为计算机视觉领域的重要研究方向，在过去十年中获得了学术界和工业界的持续关注。这项技术旨在从自然场景图像中自动定位并识别文本内容，其应用场景已渗透到现代社会的多个关键领域。在智能交通系统中，车牌识别技术实现了车辆身份的自动化管理；在移动支付领域，票据和证件识别简化了金融业务流程；在工业自动化方面，产品标签识别提升了生产线效率；而在智能终端设备上，实时翻译和场景理解等功能都依赖于精准的文本检测与识别技术。

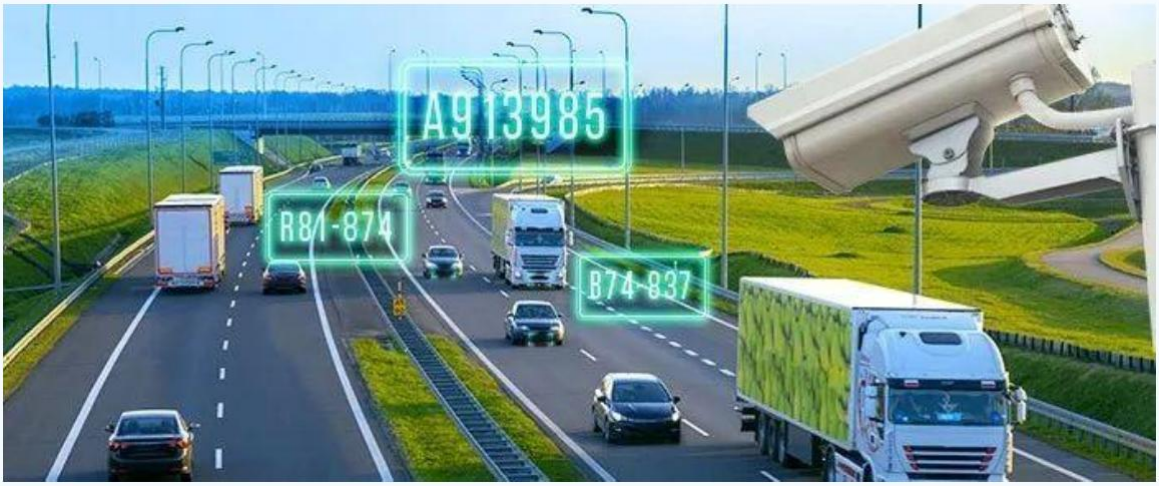

图 1 车牌识别(示例)

与传统的文档 OCR（光学字符识别）技术相比，场景文本识别面临着更为复杂的挑战（如表 1 所示）。首先，自然场景中的文本往往存在于复杂的背景干扰中，如街景中的文字可能被树木、行人或其他物体部分遮挡。其次，文本的字体、大小和方向变化极大，从规整的印刷体到手写体，从几像素的小文字到占据图像主要区域的大文字，都给检测算法带来困难。再者，任意形状文本（如弯曲排列的广告文字、弧形分布的商标文字）的定位和识别需要特殊的处理机制。此外，光照条件不均、运动模糊、低分辨率等成像质量问题进一步增加了技术实现的难度。

表 1 场景文本识别与文档 OCR 的技术挑战对比

挑战	场景文本识别	传统文档 OCR
背景复杂度	高：自然场景包含复杂纹理、遮挡物、动态干扰（如行人、车辆）	低：通常为纯色背景（如白纸），结构规整

文本多样性	极大：任意字体、大小（10px~1000px）、颜色、语言混排	有限：标准印刷字体，字号相对统一（通常 12pt~72pt）
文本形状	任意方向（0°~360°）、弯曲、扭曲、透视变形	水平或垂直排列，基本无形变
成像质量	低光照、运动模糊、低分辨率、阴影、反光	高对比度、均匀光照、300dpi 以上分辨率
标注成本	高：需标注多边形顶点（4~20 点）、字符级标注	低：矩形框标注+转录文本即可

装

订

线

1.2 研究现状

下表总结了部分主流的场景文本检测（STD）与识别（STR）模型及其关键信息：

表 2 主流 STD/STR 模型相关信息

类别	模型名称	发表时间	期刊/会议
场景文字检测	**DBNet**	2019	AAAI
	DBNet++	2022	CVPR
	EAST	2017	CVPR
	FCENet	2021	CVPR
	SegLink	2017	CVPR
	CTPN	2016	ECCV
场景文字识别	CRNN	2015	TPAMI
	RARE	2016	CVPR
	FAN	2017	ICCV
	ABINet	2021	CVPR

1.3 目前所作的工作

目前已完成的工作主要包括文献调研与实验验证两方面：在文献调研方面，系统梳理了场景文本检测与识别领域的经典算法发展脉络，重点研究了 DBNet/DBNet++的可微分二值化机制及其多尺度文本检测应用，以及 ABINet 的视觉-语言模态双向交互机制；在实验验证方面，已成功复现 DBNet++模型并在 ICDAR2015 数据集上取得 88%的准确率。

2 理论部分

2.1 DBNet 模型

2.1.1 DBNet 模型概述

DBNet（Differentiable Binarization Network）是一种基于分割的文本检测方法，通过引入**可微分二值化（Diiferentiable Binarization, DB）模块**，解决了传统分割方法中二值化阈值需手工设定与训练过程解耦的问题。其核心思想是将文本检测任务建模为**概率图（Probability Map）**和**阈值图（Threshold Map）**的联合预测，并通过端到端学习实现高精度任意形状文本检测。

2.1.2 网络结构

DBNet 的结构可分为三部分（如图 1 所示）：

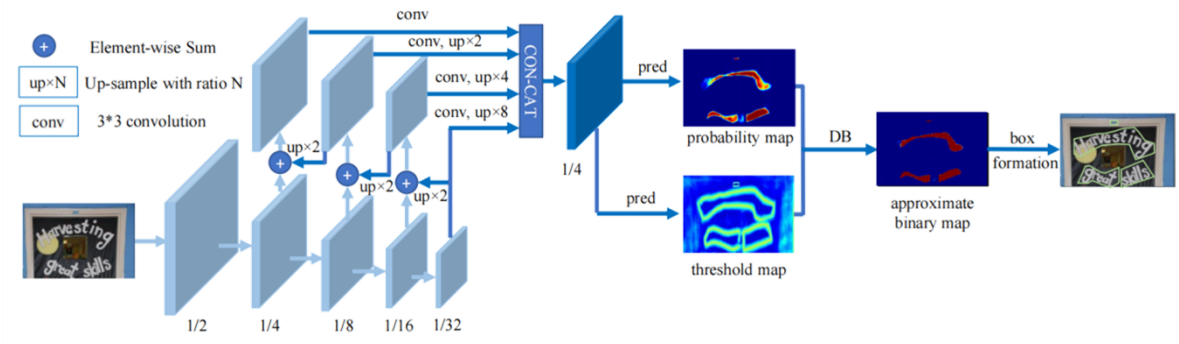

图 2 DBNet 网络结构图

1. Backbone:：通常采用 FPN（Feature Pyramid Network）或 ResNet 提取多尺度特征。

2. DB 模块：

(1) 概率图（Probability Map）：每个像素点属于文本区域的概率P ∈[0, 1]。

(2) 阈值图（Threshold Map）：动态预测局部阈值T ∈[0, 1]，适应不同文本区域的明暗变化。阈值图的标签生成基于文本多边形标注，具体步骤如下：

① 生成文本区域掩码

根据输入的文本标注多边形（如四边形顶点坐标）生成二进制掩码G，其中文本区域为1，背景为 0。

② 收缩文本区域

使用 Vatti 裁剪算法对原始文本多边形向内收缩，得到收缩后的掩码G_s，收缩偏移量D的计算公式：

$$D = \frac{A(1-\mathrm{r}^2)}{\mathrm{L}} \tag{1}$$

其中A, L, r分别表示原始多边形面积、原始多边形周长、收缩比例（论文默认为 0.4）。

③ 构建带状区域

定义带状区域为原始文本区域G与收缩区域G_s之间的部分，对于带状区域内的像素根据其到文本边界的距离赋予高斯权重（值域为[0.3, 0.7]）,对于其余位置（文本内部和背景）则将相应的阈值设置为 0。

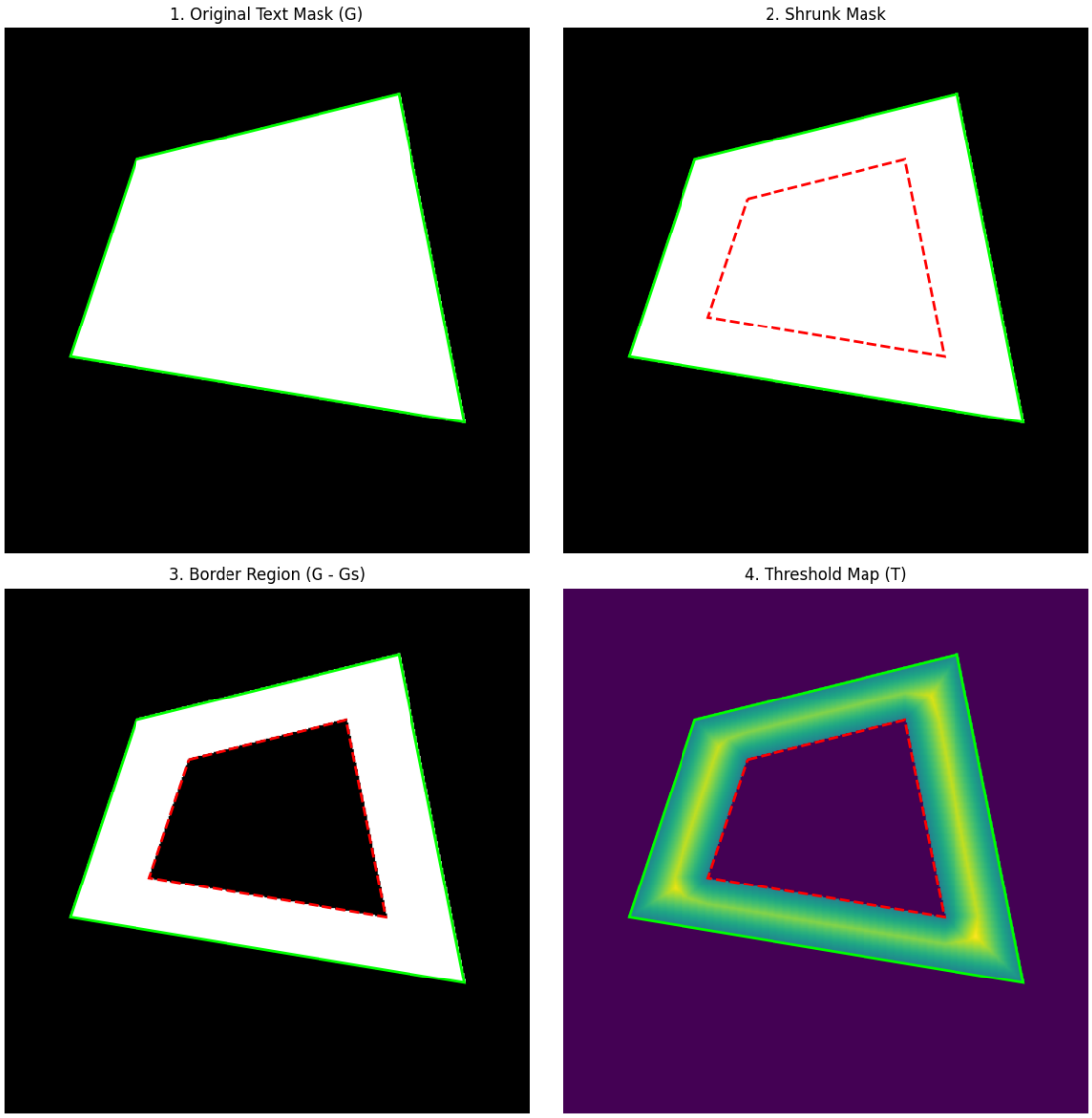

图 3 动态阈值图计算过程可视化演示

3. 可微分二值化：

传统的文本检测方法中，二值化通常通过阶跃函数实现，其表达形式如下：

$$B_{i,j} = \begin{cases} 1 \ if \ P_{i,j} > t \\ 0 \ otherwise \end{cases} \tag{2}$$

然而，阶跃函数在除阈值t处外的导数均为零，这会阻断梯度的反向传播，导致模型无法对阈值进行有效优化。此外，固定阈值在实际应用中往往难以适应复杂场景（如光照不均、低对比度文本等）。

为了克服这一问题，DB 算法引入了一个可微的 Sigmoid 函数来近似阶跃函数，从而在保留二值化能力的同时，支持端到端的梯度传播：

$$\widehat{B_{i,j}} = \frac{1}{1 + e^{-k(P_{i,j} - T_{i,j})}} \tag{3}$$

其中k为放大因子（论文中默认$k = 50$），用于控制函数的陡峭程度，同时引入该因子的一

个重要作用是在使用交叉熵损失函数（Cross Entropy Loss） 时，可以增强对错分样本的惩罚（如图 5 所示），从而提升模型在边界区域的判别能力。

$$CELoss = -y * log[f(x)] - (1-y) * log[1-f(x)] \tag{4}$$

$$\begin{cases} l_+ = -log\left(\dfrac{1}{1+e^{-kx}}\right) \\ l_- = -log\left(1-\dfrac{1}{1+e^{-kx}}\right) \end{cases} \tag{5}$$

$$\begin{cases} \dfrac{\partial l_+}{\partial x} = -kf(x)e^{-kx} \\ \dfrac{\partial l_-}{\partial x} = kf(x) \end{cases} \tag{6}$$

其中l_+和l_-分别为错判为正样本和负样本的交叉熵损失函数值。

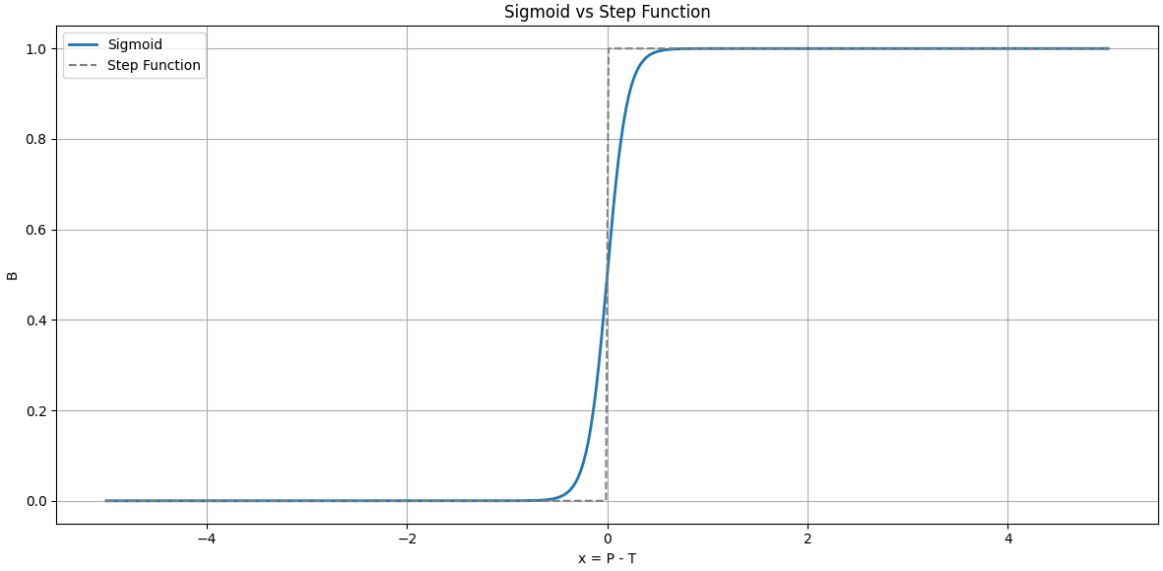

图 4 可微分二值化 DB 近似阶跃函数效果可视化展示

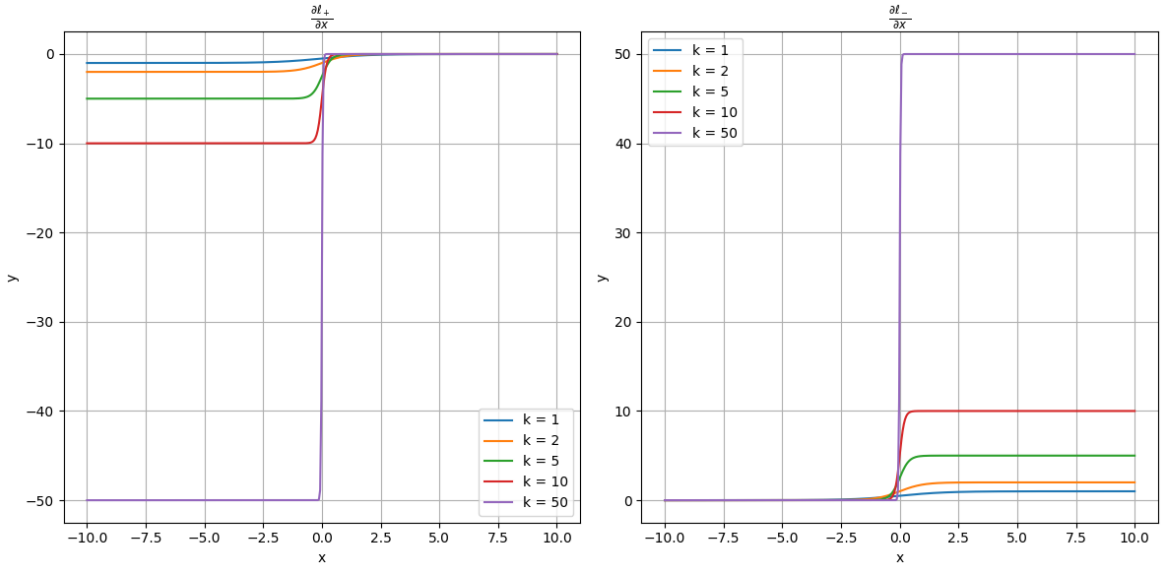

图 5 错判样本损失值函数的导数图像

2.1.3 损失函数

DBNet 训练过程的损失函数总共由三部分组成：

$$\mathcal{L} = \mathcal{L}_P + \alpha\mathcal{L}_T + \beta\mathcal{L}_B \tag{7}$$

1. 概率图损失 $\mathcal{L}_P$：二分类交叉熵损失，监督文本/非文本区域。

$$\mathcal{L}_P = -\frac{1}{N}\sum_{i,j}\left[y_{i,j} log P_{i,j} + \left(1 - y_{i,j} log\left(1 - P_{i,j}\right)\right)\right] \tag{8}$$

2. 阈值图损失 $\mathcal{L}_T$：L1 损失，仅计算文本边界区域的像素（通过膨胀/搜收缩真实标注生成边界掩膜）。

$$\mathcal{L}_T = \frac{1}{N_z}\sum_{i,j\in Z}\left|T_{i,j} - y_t^{i,j}\right| \tag{9}$$

其中Z为带状区域，$y_t^{i,j}$为高斯权重标签。

3. 二值图损失 $\mathcal{L}_B$：与 $\mathcal{L}_P$类似，但作用域二值化输出B。

2.2 DBNet++模型

2.2.1 DBNet++模型概述

DBNet++是在 DBNet 基础上进行重大改进的文本检测模型，专为复杂场景下的文本检测任务进行优化和增强。该模型的一大创新之处在于创造性地引入了**自适应尺度融合（Adaptive Scale Fusion，ASF）模块**，这是 DBNet++区别于传统文本检测方法的关键设计之一。ASF 模块的提出，旨在有效解决自然场景图像中因文本大小、形态和排列方式多样所带来的多尺度建模难题。

ASF 模块通过引入**动态融合机制**，能够根据图像中实际文本的空间分布和尺度特征，自主学习并调整来自不同尺度特征图的信息融合权重，从而实现更加灵活且具有针对性的特征整合。与 DBNet 中采用固定加权方式不同，ASF 可根据场景变化自适应地放大重要尺度的响应，抑制冗余或无关信息，极大地提升了模型对小尺度文本、弯曲文本、密集排布文本等复杂结构的感知与判别能力。

2.2.2 网络结构

DBNet++的网络架构在继承 DBNet 整体框架设计的基础上，通过引入自适应尺度融合（ASF）模块进行结构优化。整体网络仍保持主干特征提取网络 + 多尺度融合 + 可微分二值化分支的经典结构，其中 ASF 模块的加入显著提升了模型对多尺度文本的特征融合能力，使模型在复杂场景下具有更强的适应性和鲁棒性。图 5 展示了 DBNet++的详细网络结构。

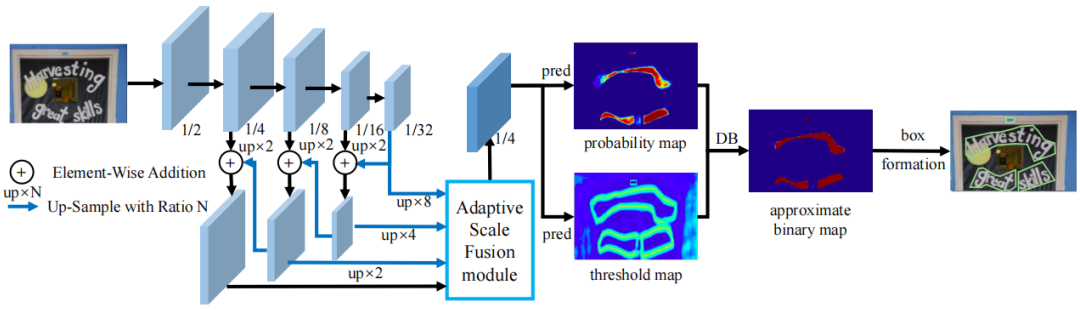

图 6 DBNet++网络结构图

2.2.3 自适应尺度融合模块（ASF，Adaptive Scale Fusion）

ASF 模块是 DBNet++ 中的核心创新，旨在解决传统多尺度特征融合（如 FPN 的简单相加或拼接）对不同尺度文本适应性不足的问题。其核心思想是动态学习不同尺度特征的权重，使网络能够根据输入文本的尺寸、形状和上下文环境，自动调整各层特征的贡献，从而提升检测的鲁棒性。

如图 7 所示，其结构包含特征对齐、空间注意力权重学习和加权融合三部分，具体流程如下：

1. 特征对齐：由于不同层特征的分辨率不同，ASF 首先对所有输入特征进行上采样或下采样，使其尺寸统一。

2. 动态权重学习（空间注意力）：先对每个尺度的特征图进行空间平均池化（Spatial Average Pooling），得到全局空间信息，随后通过卷积层（Conv-ReLU 和 Conv-Sigmoid）生成空间注意力权重图。

3. 加权融合：将输入特征图与对应的空间注意力权重相乘，突出重要区域，再对所有加权后的特征图求和，生成融合特征。

公式表示为：

$$S = \mathrm{Conv}(\mathrm{concat}([X_0, X_1, \ldots, X_{N-1}]))$$

$$A = \mathrm{Spatial\_Attention}(S)$$

$$F = \mathrm{concat}([E_0 X_0, E_1 X_1, \ldots, E_{N-1} X_{N-1}])$$

其中X_i表示第i张输入特征图，S表示中间特征图，E_i表示第 i 张特征图的动态权重值，F表示最终的特征图。

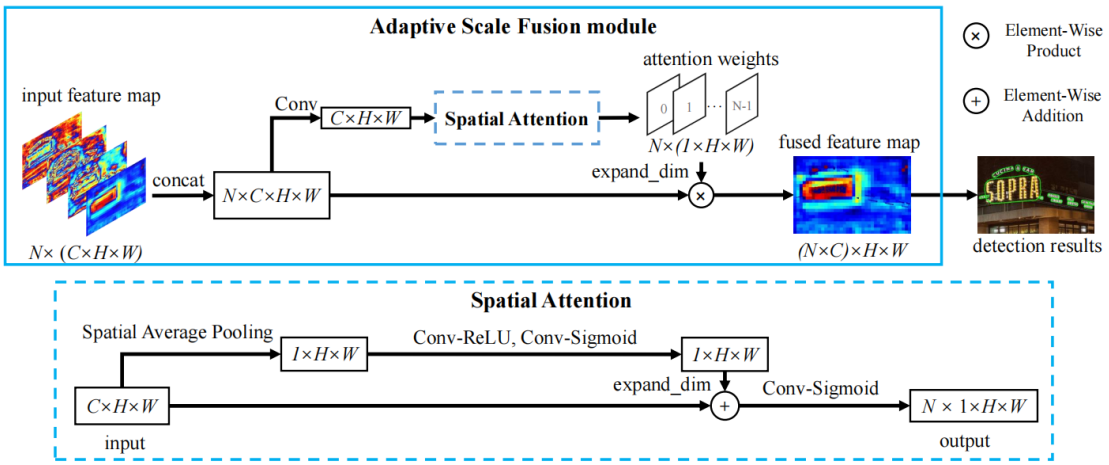

图 7 Adaptive Scale Fusion module

2.3 ABINet 模型

2.3.1 ABINet 模型概述

ABINet（**A**utonomous **B**idirectional **I**terative Network）是一种基于视觉-语言联合建模的文本识别模型，通过双向语言建模和迭代修正机制，解决复杂场景下的语义模糊与边界歧义问题。该

模型设计灵感来源于人类在阅读文本时所展现出的认知行为特征。在实际阅读过程中，人类通常不仅依赖于视觉感知获取字符形态信息，还会借助语言上下文来理解和修正模糊字符，尤其在面对字迹不清、遮挡、扭曲等复杂文本场景时，人类会不断进行前后文推理与主动修正，以获得准确的语义理解。该模型的核心创新点主要体现在以下三个方面：

1. 自主训练（**A**utonomous）：ABINet 中的视觉模型与语言模型采用分离式预训练机制，使二者在各自任务中分别学习稳定、鲁棒的特征表达，进而在融合阶段实现更加可靠的协同推理。这种设计有效避免了视觉误识别对语言模型造成的误差传导与累积。

2. 双向上下文建模（**B**idirectional）：语言建模模块不仅学习字符序列的前向依赖关系，还同时建模后向依赖，从而全面捕捉上下文语义信息。这种双向建模能力使得模型在遇到缺失、模糊或干扰字符时，能够基于完整语境进行合理推断和纠正。

3. 迭代优化（**I**terative）：ABINet 引入了类人认知行为中的“反复验证”思想，即通过多个预测-修正循环阶段，不断对初始识别结果进行优化。在每一轮迭代中，视觉感知与语言理解共同参与预测调整，从而逐步消除识别误差，提升最终输出的准确性。

2.3.2 网络结构

ABINet 的网络结构围绕“视觉感知 + 语言建模 + 迭代优化”的核心思想展开，如图 8 所示，整体架构由视觉编码器（**Vision Model**）、语言建模器（**Language Model**）、融合模块（**Fusion Module**）与迭代推理模块（**Iterative Correction**）组成，形成一个端到端可训练的文本识别框架。

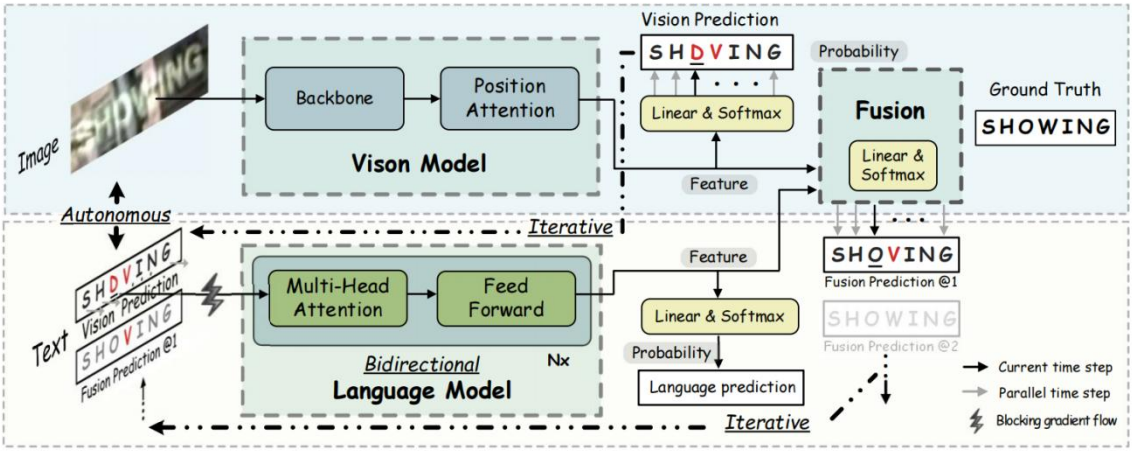

图 8 ABINet 网络结构图

1. Vision Model

该模块负责从输入图像中提取稳定且具有区分性的视觉特征，是整个 ABINet 框架中至关重要的组成部分。为了同时兼顾局部细节捕捉与全局上下文建模，ABINet 采用了 ResNet 与 Transformer 结合的混合视觉编码策略，有效融合了卷积神经网络的结构感知能力和自注意力机制的长程依赖建模能力。具体而言，原始图像首先通过 ResNet 网络提取多尺度二维空间特征图，这一阶段主要负责保留字符边缘、纹理及几何结构等视觉关键信息。随后，这些特征被输入到 Transformer 编码器中进行进一步处理，通过多头自注意力机制建立图像区域间的远程关联，从而增强模型对复杂字符形态、干扰背景及低质量图像的鲁棒性。

为进一步提升模型对空间位置的感知与语义聚焦能力，ABINet 在视觉特征提取后引入了位置注意力机制（Position Attention Module），该机制能够自适应地关注文本图像中的关键区域，并强化重要位置的特征表达。最终，通过一个 softmax 层将处理后的特征转化为每个字符位置的分类概率分布，输出视觉模块的初步识别结果。

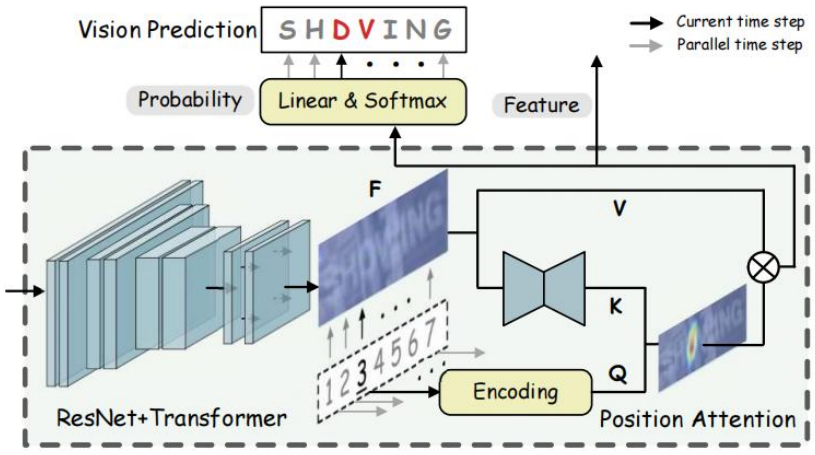

图 9 ABINet 的 Vision Model 结构图

2. Language Model

ABINet 中的语言建模模块的目标是通过学习字符序列中的上下文依赖关系，从语义层面对视觉识别结果进行补全与纠错。该模块基于 Transformer 架构构建。与传统单向语言模型不同，ABINet 采用了双向上下文建模（Bidirectional Context Modeling）策略，即模型在预测某个字符时，会同时考虑其前文与后文的信息。这一机制模仿了人类在阅读过程中的语义理解方式——我们在阅读时并非仅从前到后线性地解析文本，而是会根据上下文进行前后关联的推理，从而提高对模糊或不完整字符的识别能力。

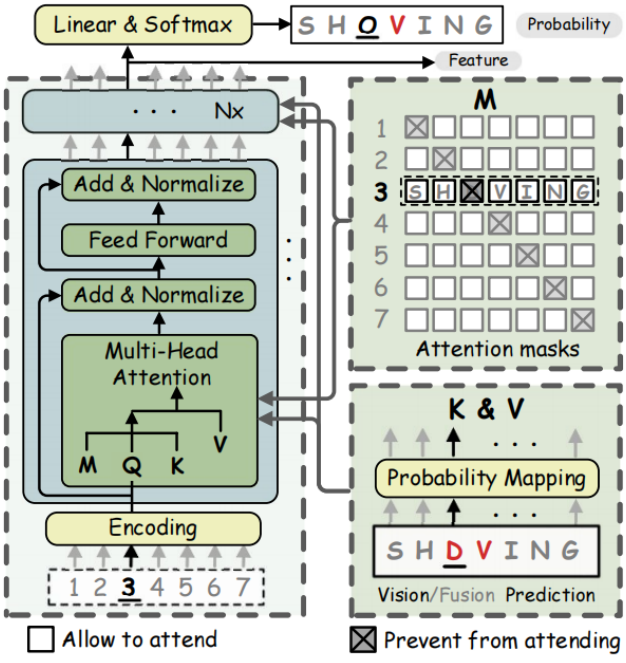

图 10 ABINet 的 Language Model 结构图

3. Fusion Model

在 ABINet 架构中，视觉模块与语言模块分别从图像和文本两个不同模态中独立学习特征。尽管这两种特征都能在一定程度上反映字符信息，但它们本质上来自异构模态：视觉特征专注于字符的形态与结构，而语言特征则强调语义一致性与上下文关系。为了实现两者之间的信息对齐，并做出更准确的最终识别决策，ABINet 设计了一个融合模块（Fusion Module），用于整合来自视觉与语言模型的输出，该模块采用了门控机制（Gated Mechanism），以自适应方式动态调控视觉特征与语言特征的贡献权重。具体公式为：

$$\mathrm{G} = \sigma([\mathrm{F_v}, \mathrm{F_l}]\mathrm{W_f})$$

$$F_f = G \odot F_v + (1 - G) \odot F_l$$

其中 $\mathrm{F_v}$ 和 $\mathrm{F_l}$ 分别表示视觉模块和语言模块的预测结果，σ表示 Sigmod 函数，$\mathrm{W_f}$是可学习的参数矩阵，G表示融合权重，F_f 表示最终的预测结果。

4. Iterative Correction

在 ABINet 中，语言模型虽然具备强大的上下文建模能力，但其输入往往来自视觉模块的预测结果，而这些预测在实际场景中可能受到模糊、遮挡、背景复杂等干扰因素的影响，导致存在大量噪声字符。这类噪声会破坏语言模型的语义理解过程，进而引起错误传播，严重时甚至会导致整个识别序列的结构紊乱。为了解决这一问题，ABINet 提出了迭代式语言模型（Iterative Language Model）机制，通过多轮语言推理对视觉预测进行逐步修正，实现更精细的字符预测优化。具体而言，语言模型会被重复执行 M 次。在首次迭代中，其输入为来自视觉模块的预测，记为 y^1。在随后的每一次迭代中，输入则更新为前一轮由融合模块输出的结果 y^t，即：

$$y^t = FusionModel(y^{t-1}), t = 2,3,\ldots,M$$

这种迭代方式可以不断调整和强化语言模型对语义的把握能力，从而逐步消除视觉预测中的错误信息。例如，在图像中原始视觉输入预测 “APPLA” 时，语言模型在理解上下文 “APPLE” 的过程中识别出字符 “A” 属于噪声，进而将其纠正为更合理的 “E”。

2.3.3 损失函数

为了实现视觉信息、语言信息与融合结果之间的协同优化，ABINet 采用了多任务联合训练策略（Multi-task Joint Training）。该策略通过对多个模块施加监督信号，使模型在训练阶段能够同时学习视觉感知、语言建模以及跨模态融合的能力，从而提升整体识别性能与稳定性。

模型的总损失函数形式如下：

$$\mathcal{L} = \lambda_{\mathrm{v}}\mathcal{L}_{\mathrm{v}} + \frac{\lambda_{\mathrm{l}}}{\mathrm{M}}\sum_{\mathrm{i=1}}^{\mathrm{M}}\mathcal{L}_{\mathrm{l}}^{(i)} + \frac{1}{\mathrm{M}}\sum_{\mathrm{i=1}}^{\mathrm{M}}\mathcal{L}_{\mathrm{f}}^{(i)} \tag{11}$$

其中：$\mathcal{L}_{\mathrm{v}}$ 表示来自视觉模块的交叉熵损失（Cross-Entropy Loss），用于监督模型从原始图像中提取准确的字符预测；$\mathcal{L}_{\mathrm{l}}^{(\mathrm{i})}$表示第 i 次迭代中，语言模型对字符序列的预测损失；$\mathcal{L}_{\mathrm{f}}^{(\mathrm{i})}$ 表示第 i 次迭代中，融合模块输出结果的损失，用于对视觉-语言联合特征的最终预测进行监督；M为语言模型迭代次数；λ_{v} 和 λ_{l} 是两个用于平衡视觉损失与语言损失权重的可调超参数。

3 实验部分

3.1 场景文字检测（Scene Text Detection, STD）

3.1.1 DBNet++模型复现与验证

为评估 DBNet++ 模型的文本检测性能，本研究基于公开数据集 ICDAR2015 对其进行了复现与训练。

数据集说明：

ICDAR2015 作为场景文本检测领域广泛使用的基准数据集，包含 1000 张训练图像和 500 张测试图像，具有多方向文本、复杂背景等特点，能有效检验模型在实际场景中的性能表现。

实验设置：

实验在 PyTorch 框架下完成了全流程实现，具体环境配置与训练参数如下表所示：

表 3 复现环境配置及训练参数

项目	配置/参数
操作系统	Ubuntu-22.04
Python 版本	3.13.2
开发框架	PyTorch 2.6.0
CUDA 版本	CUDA 11.8
GPU	NVIDIA GeForce RTX 3060 （单卡）
损失函数	DBLoss，$\alpha = 1$，$\beta = 10$
学习率	初始 1e-3，Warmup+Poly 衰减
优化器	Adam（amsgrad=True）
训练 epoch	1200
评价指标	Precision、Recall、F-measure

本实验采用文本检测任务中常用的三项评价指标进行性能评估，分别为：

1. Precision（精确率）：检测出的文本区域中，真正为文本的比例。

2. Recall（召回率）：所有真实文本区域中，被正确检测出的比例。

3. F-measure（F1 值）：精确率与召回率的调和平均，用于衡量整体检测性能。

3.1.2 实验结果与分析

训练过程监控：

如图 11 所示，模型在训练过程中表现出明显的阶段性特征：

1. 快速上升阶段（0-200 epoch）：三项指标均呈现显著增长趋势。此阶段模型快速学习文本区域的基础特征表示。

2. 渐进优化阶段（200-800 epoch）：检指标增速明显放缓，呈现渐进式提升。反映模型进入细粒度特征优化期。

3. 稳定收敛阶段（>800 epoch）：各项指标波动幅度显著减小，F-measure 在 84.9%-85.7%区间窄幅震荡。表明模型达到收敛状态。

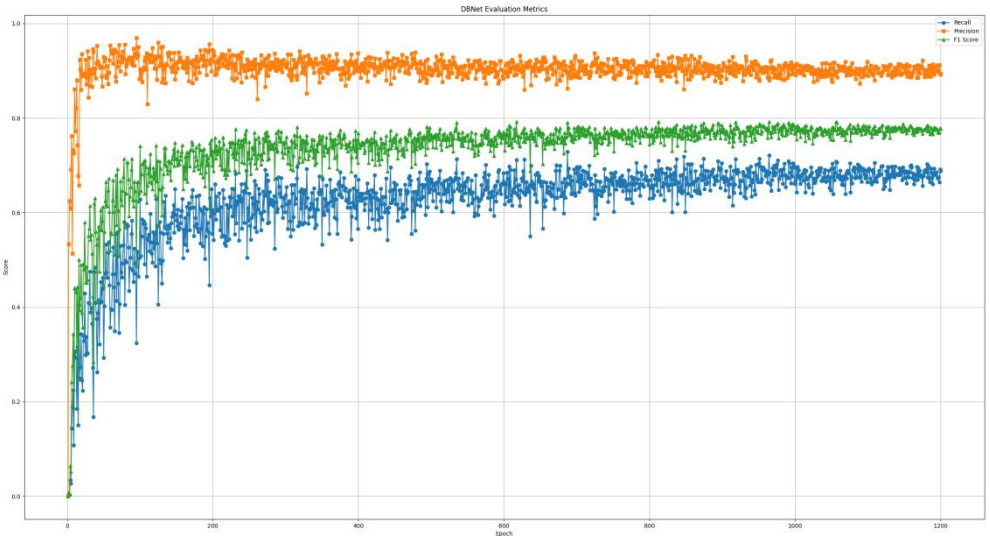

图 11 模型关键指标变化图

可视化检测结果：

图 12 展示了模型在 ICDAR2015 测试集上的典型检测效果：

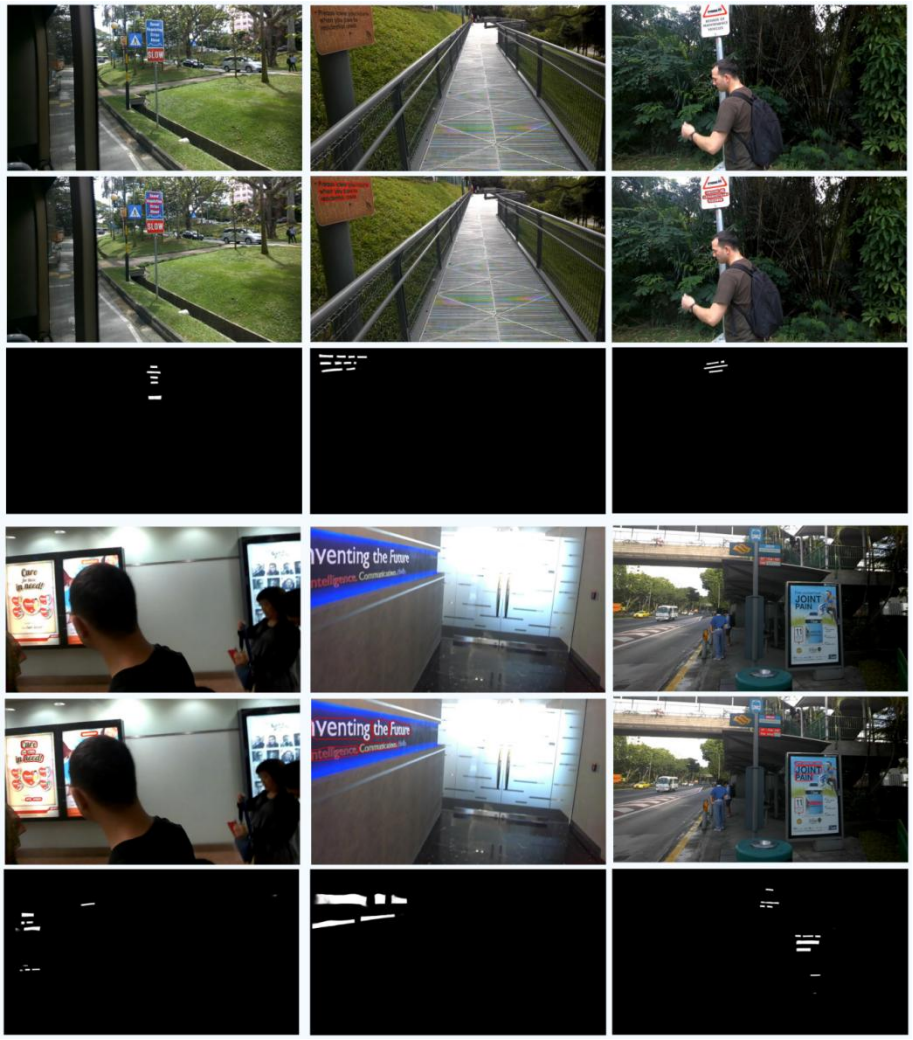

图 12 模型测试结果(上图为原图，中图为加入文本检测框后的预测图，下图为文本预测区域二分图)

参考文献

[1] Liao, M., Wan, Z., Yao, C., Chen, K., & Bai, X. (2020, April). Real-time scene text detection with differentiable binarization. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 11474-11481).

[2] Fang, S., Xie, H., Wang, Y., Mao, Z., & Zhang, Y. (2021). Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7098-7107).

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[4] Liao, M., Zou, Z., Wan, Z., Yao, C., & Bai, X. (2022). Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE transactions on pattern analysis and machine intelligence, 45(1), 919-931.

[5] Ye, Q., & Doermann, D. (2014). Text detection and recognition in imagery: A survey. IEEE transactions on pattern analysis and machine intelligence, 37(7), 1480-1500.

[6] Long, S., Ruan, J., Zhang, W., He, X., Wu, W., & Yao, C. (2018). Textsnake: A flexible representation for detecting text of arbitrary shapes. In Proceedings of the European conference on computer vision (ECCV) (pp. 20-36).

[7] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[8] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[9] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

[10] Wan, Q., Ji, H., & Shen, L. (2021). Self-attention based text knowledge mining for text detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5983-5992).

装 订 线