南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Knowledge-Enhanced Historical Document Segmentation and Recognition

En-Hao Gao, Yu-Xuan Huang, Wen-Chao Hu, Xin-Hao Zhu, Wang-Zhou Dai

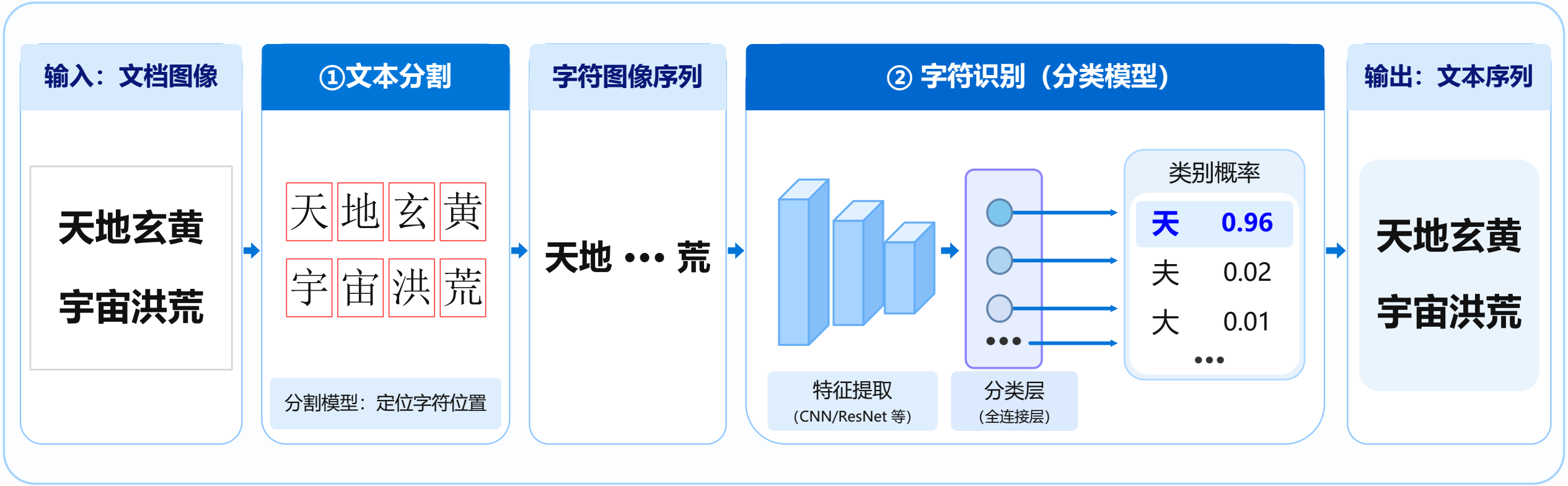
AAAI 2024

汇报人：王必广
汇报日期：2026.11.10

传统 OCR：基本流程

传统OCR的一般流程

- 通常采用“分割–识别”的两阶段流程：先使用分割模型定位并提取出每个字符的图像，再使用分类模型对每个字符图像进行类别预测，最终得到整段文本。



现代文档与历史文档的 OCR 对比

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

现代文档：OCR已较为成熟

人工智能技术及其应用

人工智能（AI）是研究、开发用于模拟、延伸和扩展人类智能的理论、方法、技术及应用系统的学科。近年来，人工智能在图像识别、自然语言处理等领域取得了显著进展，并在多个行业中得到广泛应用。

- **文字排列整齐**
  (版面规范、行列对齐)
- **字迹清晰规范**
  (印刷体为主)
- **标注数据充足**
  (通常具有字符级标注)

历史文档：仍面临多重挑战

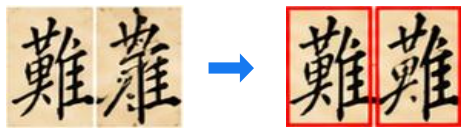


- **版面与文字布局不规整**
  (行列不齐、间距不一、倾斜等)
- **字形多样且易粘连**
  (手写体、异体字、笔画粘连等)
- **标注数据匮乏**
  (缺少字符级标注、标注成本高)
- **图像质量较差**
  (纸张破损、污渍、褪色、噪声等)

对OCR模型的影响

上述问题会导致传统OCR模型在分割、识别以及训练等方面表现不佳。

分割：文本区域划分不准



- 相邻字符粘连
- 字符边界模糊
- 易出现过分割/欠分割

识别：字符分类更困难



- 存在大量异体字、变体字
- 字形书写差异大，类间易混淆

训练：有效训练数据不足

- 缺少大规模、字符级标注数据
- 标注成本高、获取难度大
- 影响模型的泛化能力

方法总览（Overview of KESAR）

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

KESAR：面向历史文档的 OCR 方法，包含字符分割模型和字符识别模型，并通过三阶段学习策略将知识融入分割与识别，实现两者的**联合优化。**

推理流程（Inference Pipeline）

输入



历史文档图像

分割模型（Segmentation Model）

- 像素级预测三类区域：
  - 字符区域 (Character Region)
  - 字符亲和 (Character Affinity)
  - 背景区域 (Background)
- 利用分水岭算法聚合字符区域 (Watershed Aggregation)



字符区域

字符亲和

背景

文本行1 文本行2 文本行3 文本行4



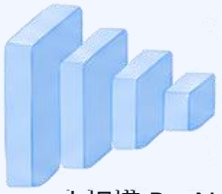
分割输出：字符位置 + 文本行归属

裁剪

根据字符框
裁剪单字符图像



识别模型 (Recognition Model)



小规模 ResNet
(Small-Scale ResNet)

输出

天 (行4)
地 (行4)
玄 (行4)
黄 (行4)
...

三阶段学习框架（闭环优化）

基于结构知识的分割学习
(利用文本结构知识增强分割模型)

分割模型 (Segmentation Model)

识别模型 (Recognition Model)

分割与识别的联合优化
(字形知识改进分割，
改进分割结果促进识别)

基于溯因匹配的识别学习
(利用溯因匹配知识和未标注字符
训练识别模型)

知识驱动的联合优化（分割 ↔ 识别，相互促进）

基于结构知识的分割学习

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

核心思想：利用文本结构知识修正伪字符框，在仅有文本行标注的情况下增强分割模型训练。

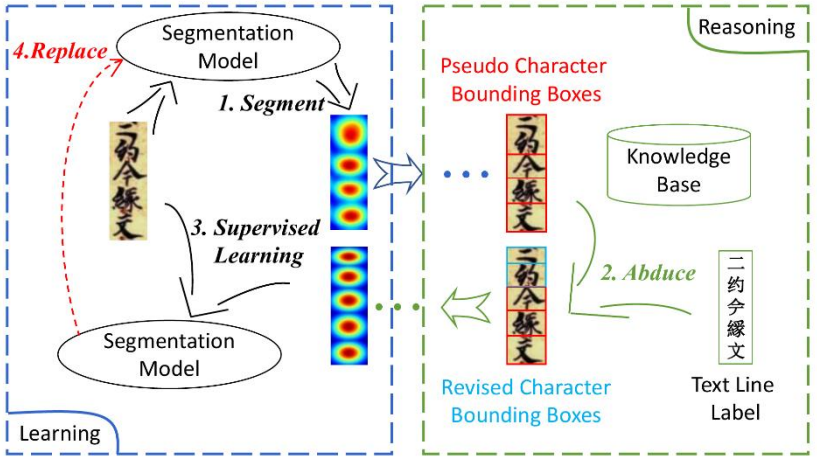
整体流程

输入



\+

文本行标注 (TL)



结构知识（FOL 规则示意）

1. $\text{reg_bbox}(B) \leftarrow \text{close}(\text{asp_rat}(B), 1)$
   - **字符框应接近方形。**
2. $\text{reg_textline}(TL) \leftarrow \text{bbox_seq}(TL) = \{B_1, B_2, \ldots\} \wedge \text{reg_bbox}(B_1) \wedge \text{reg_bbox}(B_2) \wedge \cdots \wedge \text{close}(\text{asp_rat}(B_1), \text{asp_rat}(B_2), \ldots) \wedge \text{close}(\text{size}(B_1), \text{size}(B_2), \ldots)$
   - **同一文本行中字符宽高比与尺寸应相近。**
3. $\text{false} \leftarrow \text{bbox_seq}(TL) = \{B_1, B_2, \ldots\} \wedge \text{horizontal_adjacent}(B_1, B_2, \ldots)$
   - **竖排文本中不应出现水平相邻字符。**

伪字符框如何修正？（论文中的实际做法）

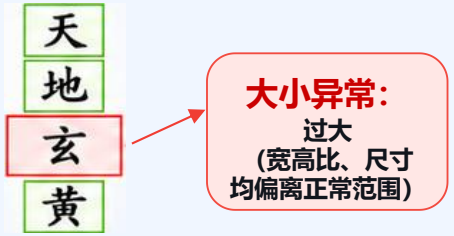
A. 合并水平相邻框

- 处理一个字符被错误切成左右两块。



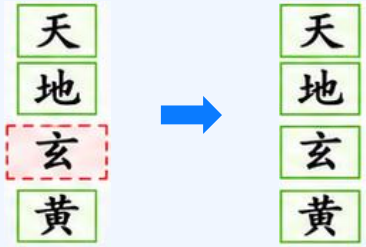
B. 找到异常框

- 依据文本行内字符的宽度、高度、宽高比识别不规则框。



C. 按正常框平均尺寸修正

- 将异常框调整到更合理的大小与形状。



修正依据：

- 参考同一行中正常字符框的平均宽高、尺寸
- 调整异常框的大小与位置，使其满足结构规则

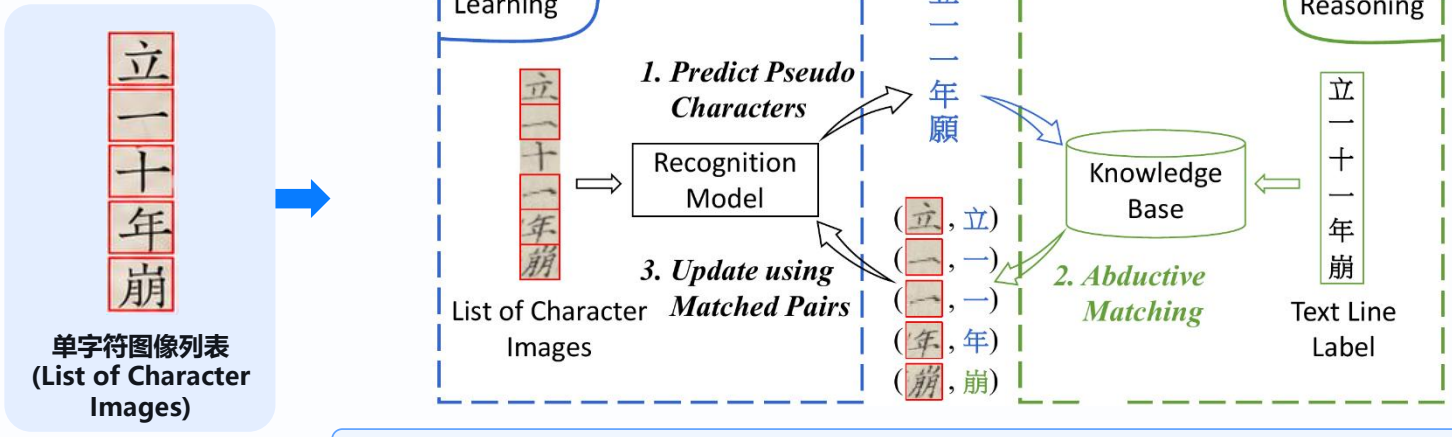
输出：更可靠的字符框 → 作为字符级监督信号，持续优化分割模型

基于溯因匹配的识别学习

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

核心思想：利用识别模型对单字符图像预测伪标签，再结合文本行标注与溯因匹配知识，构造可靠的“图像-标签”匹配对，用于迭代训练识别模型。

整体流程（Overall Pipeline）



目标：从不完全可靠的预测与文本行标注中，找到最大的一致匹配集合 Δ

溯因匹配算法（Algorithm 1）

输入： 预测字符串 $P = (p_1, p_2, \ldots, p_n)$
真实文本行 $G = (g_1, g_2, \ldots, g_m)$
最大匹配长度 max_len

输出： 最大匹配集合 Δ

1. 使用动态规划在 P 与 G 之间寻找最优匹配
2. 若 $p_i = g_j$ 则直接对齐
3. 允许长度不超过 max_len 的连续子串匹配
4. 回溯得到最大 grounding 集合 Δ

区别于普通 LCS：KESAR 强调连续匹配子串，更适合字符图像与文本行的对齐。

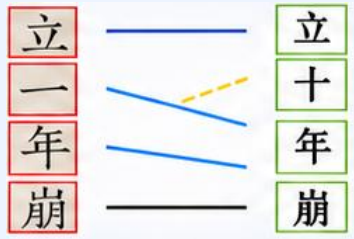
三种典型匹配关系（论文示例）

当识别误差导致字符数与文本行长度不一致时，溯因匹配仍可构造出可靠的图像-标签对。

案例 1：数量一致（Equal）



案例 2：预测更短（Shorter）



案例 3：预测更长（Longer）



图例（Legend）

- 黑线：预测对应 (Predict)
- 蓝线：按位置对齐 (Match by alignment)
- 黄虚线：溯因传播 (Match by abduction)

输出：匹配后的单字符图像-标签对 → 作为监督信号，持续优化识别模型

溯因匹配过程示例

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

核心思想： 在预测字符图像序列 P 与真实字符序列 G 中，寻找首尾相同、长度相等的连续子串，将可靠图像与真实标签配对，生成训练样本。

输入：原始行图像与分割结果

（1）原始行图像（单行文本）



（2）字符图像序列 P 与识别伪标签（示意）

故人	西	辞	黄	鹤	楼	烟花	三	月	下	扬	州
P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
放	西	词	皇	鸽	楼	姻	三	月	下	杨	州

$|P| = 12 < |G|$　　红字：示意错误预测

（3）真实文本行 G （字符序列）

故	人	西	辞	黄	鹤	楼	烟	花	三	月	下	扬	州
G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	G_{10}	G_{11}	G_{12}	G_{13}	G_{14}

$|G| = 14$

说明：分割模型把“故人”和“烟花”各合并为一个图像块，因此预测块少于真实字符，但原始行图像中的字符仍完整保留。

溯因匹配流程

1 搜索候选子串

在 P 与 G 中寻找首尾字符相同、长度相等的局部子串。

$P_2 \ldots P_6$: 西 词 皇 鸽 楼

$G_3 \ldots G_7$: 西 辞 黄 鹤 楼

候选 1（长度 = 5）

2 建立局部匹配关系

两端相同、长度相等，中间按位置自动对应，纠正“词、皇、鸽”。

P_2	P_3	P_4	P_5	P_6
西	辞	黄	鹤	楼
↓	↓	↓	↓	↓
西	辞	黄	鹤	楼

3 继续搜索（局部纠错）

继续搜索其他匹配子串，利用真实行的局部信息纠正识别伪标签。

$P_8 \ldots P_{12}$: 三 月 下 杨 州

$G_{10} \ldots G_{14}$: 三 月 下 扬 州

杨 → 扬

伪标签纠错示例：
P_{11} 的伪标签“杨”修正为真实标签“扬”。

4 生成图像-标签对

汇总所有可靠匹配关系，形成用于训练的样本集合。

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
故人 ✗	西	辞	黄	鹤	楼	烟花 ✗	三	月	下	扬	州
	↓	↓	↓	↓	↓		↓	↓	↓	↓	↓
	西	辞	黄	鹤	楼		三	月	下	扬	州

最终的图像-标签对 Δ （共 10 个）：

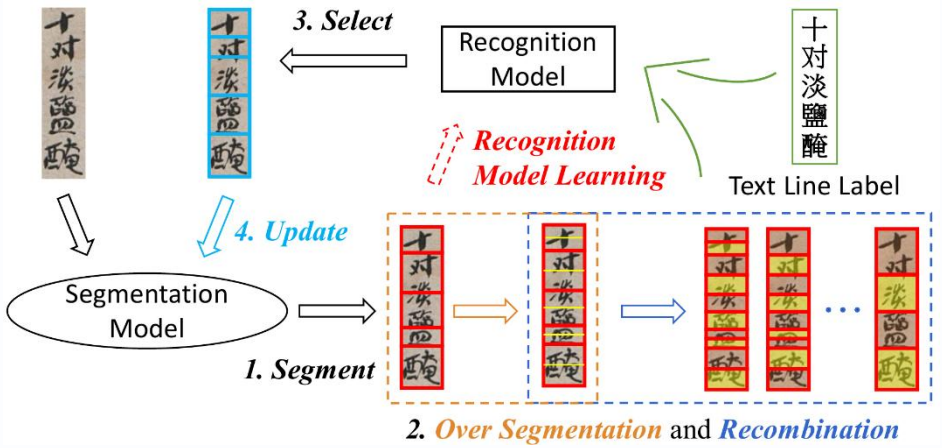
$\Delta = \{(P_2, 西), (P_3, 辞), (P_4, 黄), (P_5, 鹤), (P_6, 楼), (P_8, 三), (P_9, 月), (P_{10}, 下), (P_{11}, 扬), (P_{12}, 州)\}$

输出：匹配后的单字符图像-标签对 Δ → 作为监督信号继续训练识别模型

联合优化

核心思想：利用识别模型的高置信度结果反向优化分割模型，通过“过分割 + 重组合”纠正分割错误，并实现分割与识别的协同提升。

整体流程（对应论文 Fig. 3）



联合优化算法（Algorithm 2: Over-Segmentation and Recombination, OSR）

输入： 识别模型 f ； 字符框序列 $B = (B_1, ..., B_n)$ ；
字符标签序列 $C = (c_1, ..., c_m)$ ； 最大重组数 r

输出： 重组合后的字符框序列 D

1. 将 B 中每个字符框按高度中点切分，得到过分割结果 OB ；
2. 遍历相邻片段，计算组合识别得分 new_score ；
3. 若 new_score > score[j] （原得分），则合并两个片段，
并更新匹配长度 match_len 与回溯指针 trace ；
4. 在最多 r 次重组合范围内，回溯得到最优的字符框序列 D 。

关键变量

- score[j] ： 前 j 个片段的最佳匹配得分；
- match_len[j] ： 当前最佳匹配的长度；
- trace[j] ： 回溯指针，用于恢复最优组合。

过分割与重组合（OSR）

思想：先将每个字符框过分割为更细的片段，再基于识别得分选择性地将相邻片段重组合，纠正漏分/粘连等错误。

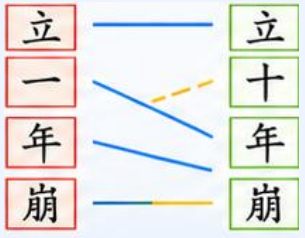
OSR 处理流程

1. **过分割（Over-Segment）**
将字符框按高度中点切分；
2. **重组合（Recombination）**
尝试合并相邻片段；
3. **评分与选择**
基于识别模型的得分，
动态优化组合方案。

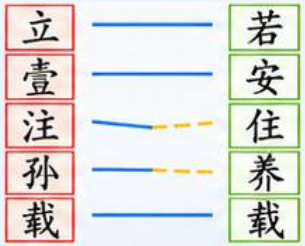
案例 1：正确分割（无需重组）



案例 2：漏分（过分割后合并）



案例 3：粘连（过分割后重组）



图例

- 预测对应 (Predict)
- 按位置对齐 (Match by alignment)
- 基于溯因的匹配/重组 (Match by abduction)

输出：优化后的分割与识别模型（相互促进、迭代提升）

实验部分（一）：实验设置与数据集

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

数据集：特点、标注与实验用途

依据论文 §5.1

TKH　用于预训练

Yang et al., 2018 · HCIILAB ·《高丽大藏经》

1. **规模**：1,000 张历史文档图像
2. **标注**：同时提供字符级与文本行级标注
3. **特点**：文本行规整、字符大小较均匀，字符类型相对有限
4. **用途**：因有字符标注，随机选取 600 张作为 KESAR 的预训练数据

MTH　更具挑战

Ma et al., 2020

1. **规模**：2,200 张历史文档图像
2. **标注**：同时提供字符级与文本行级标注；本实验仅用文本行标注验证方法
3. **特点**：文本行普遍扭曲、版面复杂，有时包含插图
4. **划分**：随机按 7:3 划分训练集与测试集

GBACHD　最具挑战

2022 大湾区（黄埔）国际算法算例大赛

1. **规模**：2,000 张图像，超过 15,000 字类（含大量罕见字与异体字）
2. **标注**：仅提供文本行级标注
3. **特点**：严重扭曲、尺度变化；污渍、模糊笔迹、印章等多源噪声
4. **划分**：随机取 1,400 张训练，其余测试

训练流程：分割与识别分别预训练、微调，再联合优化

依据论文 §5.2

模型	① TKH 预训练	② 分别微调	③ 联合优化
CRAFT（ResNet50 骨干）分割模型	TKH 训练数据 预训练 320 epochs	MTH 微调 180 epochs；GBACHD 微调 80 epochs	联合优化 10 epochs
ResNet34 识别模型	TKH 训练数据 预训练 25 epochs	使用分割模型生成的字符图像 再微调 25 epochs	

完整训练流程约 15 小时

训练时长为论文报告；更多实现细节见附录。

实验部分（二）：分割实验结果与消融

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Table 1 | 文本行分割对比与消融（论文原表）

论文 §5.3–5.4 · PDF 第 6–7 页

Method	MTH			GBACHD			TKH		
	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)	R(%)	P(%)	F(%)
PSENet	88.4	87.3	87.8	73.2	82.2	77.5	**97.5**	90.7	94.0
FCENet	78.8	81.6	80.2	71.1	69.0	70.0	84.1	84.8	84.4
CRAFT	85.9	93.4	89.5	82.5	93.2	87.5	97.3	**98.3**	**97.8**
KESAR (w/o JOPT)	92.1	**94.0**	93.1	93.5	93.4	93.5	-	-	-
KESAR	**93.1**	93.4	**93.2**	**94.4**	**94.0**	**94.2**	-	-	-

R：召回率　P：精确率　F：F-measure；表中单位均为 %，以下增量为百分点。

w/o JOPT：不含联合优化

① 分割对比：与谁比、结果如何

PSENet：渐进扩张；FCENet：傅里叶轮廓。
CRAFT：KESAR 使用的基础分割模型。

MTH：KESAR 的 F 为 93.2，
较 PSENet / FCENet 高 5.4 / 13.0。
GBACHD：F 为 94.2，为表中最高。

② 消融：结构知识的贡献

CRAFT → KESAR（w/o JOPT）

用知识库修正伪字符框，尚未联合优化。

MTH：F 89.5 → 93.1（+3.6）
GBACHD：F 87.5 → 93.5（+6.0）

两数据集 R、P、F 均提高；推理速度同 CRAFT。

③ 消融：联合优化的贡献

KESAR（w/o JOPT）→ KESAR

MTH：F 93.1 → 93.2（+0.1），R↑、P↓；
行内字符通常分离清楚，因此增益有限。

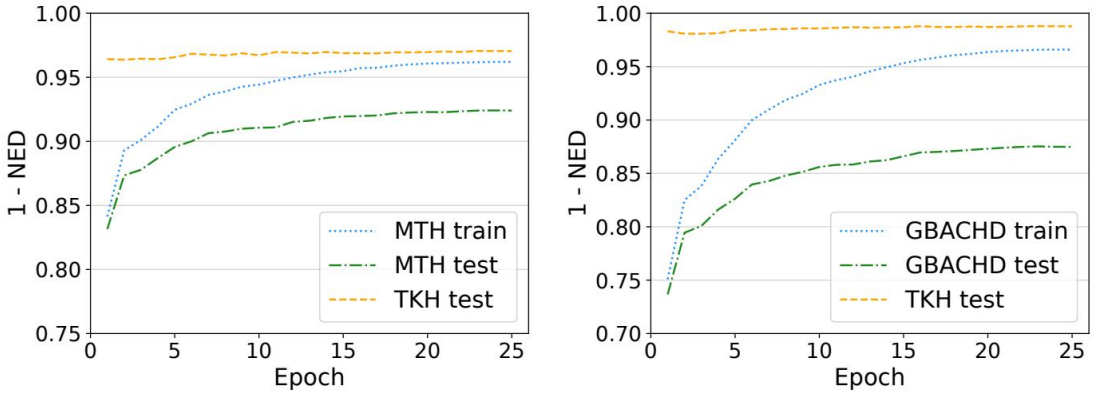
GBACHD：F 93.5 → 94.2（+0.7），
R、P、F 均升；识别模型帮助分开密集字符。

TKH 限制　字符级标注已用于预训练，未报告两种 KESAR 结果；CRAFT 与其他方法的训练标签不同，TKH 比较仅供参考。

实验部分（三）：识别实验结果与分析

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Figure 4 | 识别模型学习曲线



(a) MTH　　(b) GBACHD

蓝：训练集　绿：测试集　黄：TKH 测试集

指标为 1-N.E.D.，越接近 1 越好。
曲线显示识别性能随训练提升并逐渐趋稳。

Table 2 | 识别性能对比（1-N.E.D.）

Method	MTH	GBACHD	TKH
RobustScanner	0.905	0.722	0.991 / **0.989**
ABINet	0.900	0.718	**0.992** / 0.988
KESAR	**0.924**	**0.875**	0.970 / 0.988

Table 3 | 成功溯因率随训练轮次变化

Epoch	1	2	4	10	15	25
MTH	0.949	0.989	0.992	0.994	0.995	0.996
GBACHD	0.919	0.977	0.983	0.988	0.990	0.992

① 识别提升与方法对比

MTH：0.831 → 0.924（+0.093）
GBACHD：0.737 → 0.875（+0.138）

在 MTH、GBACHD 上优于两种对比方法；
GBACHD 比最强对比方法高 0.153。
TKH 上表现接近，但并非最优。

② 溯因匹配是否有效

成功溯因率：正确溯因得到标签的比例，
用于评估字符图像与标签的匹配效果。

第 25 轮：MTH 0.996；GBACHD 0.992。
匹配率较快收敛，支持生成字符级训练对。

注意：成功溯因率不等于识别准确率。

③ 比较条件与 TKH 列含义

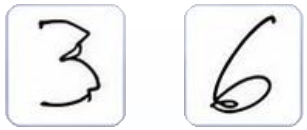
1. 对比方法也加入 TKH 数据进行训练。
2. KESAR 用预测文本行；对比方法用
   真实文本行作为训练数据。
3. TKH 两值依次对应 MTH+TKH、
   GBACHD+TKH 训练后，仅在 TKH 测试。

来源：论文 §5.3 溯因匹配、§5.4 识别对比；Figure 4、Table 2、Table 3 均为 PDF 第 7 页原图裁剪。

研究问题背景（一）：传统 ABL 基本流程

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

任务输入与监督信息



监督标签：仅知道和为 9

- 输入：两张手写数字图像
- 监督：只有加法结果标签

感知模型初始预测（可能出错）



初始预测 / 伪标签候选 “3”：0.96



初始预测 / 伪标签候选 “6”：0.93

- 模型对每张图片仅给出初步预测（伪标签候选）。
- 由于模型尚未充分训练，这些预测可能不准确。

知识约束

加法规则：

$$a + b = 9$$

$$a, b \in \{0, \ldots, 9\}$$

结合规则库与预测结果，生成满足约束的候选标签序列。

候选标签序列与选择

汉明距离（与预测结果最近）

候选序列：(3,6) (4,5) (2,7) (1,8) (0,9) …

置信度（组合概率最大）

候选序列（组合概率）：(3,6) 0.89 | (4,5) 0.92 | (2,7) 0.48 | (1,8) 0.31 | (0,9) 0.07 …

模型更新与迭代优化

- 将选中的伪标签当作硬标签（hard label），
- 用于监督感知模型再训练。
- 循环迭代，不断优化。

初始预测 → 选择伪标签 → 作为硬标签再训练

传统 ABL 的核心问题

1 候选满足知识规则 ≠一定是真实标签

$a + b = 9$ → (3,6) ✓ (4,5) ✓ (2,7) ✓ …

都满足规则，但只有一个是真实标签！

2 选择由预训练的感知模型决定

预训练感知模型（固定、可能有偏差）→ (3,6) 0.89 (4,5) 0.92

3 错误选择会导致误差逐步累积

初始迭代选错伪标签（如 (4,5)）→ 用其作为硬标签训练 → 模型被误导，后续迭代错误不断放大

研究问题背景（二）：传统 ABL 的硬选择局限与 A³BL 改进

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

知识约束： $a + b = 9$

候选序列及感知模型预测概率（示意）：

(3,6)	(4,5)	(2,7)
$p = 0.50$	$p = 0.40$	$p = 0.10$

- 三个候选均满足规则；
- 最高得分 0.50 不代表其他候选概率为 0！
- 上述概率为感知模型的预测结果（示意），并非真实标签。

传统 ABL：只保留最高分（硬监督）

	候选标签	感知模型置信度 p（示意）	
✓	(3,6)		0.50
×	(4,5)		0.40
×	(2,7)		0.10

选中 (3,6)

候选权重（硬标签）$[1,0,0]$

用单一伪标签训练
• 硬标签 / CE

⚠ **过早丢弃剩余 0.50 的可能性** ((4,5) 仍有 0.40、(2,7) 仍有 0.10)

A³BL：保留多候选（软监督）

He et al. • ICML 2024

	候选标签	感知模型置信度 p（示意）	
✓	(3,6)		0.50
✓	(4,5)		0.40
✓	(2,7)		0.10

第1张图（a）：		
3		0.50
4		0.40
2		0.10

第2张图（b）：		
6		0.50
5		0.40
7		0.10

软标签训练： $\mathrm{CE}(q,p)$ ；q 固定时等价于最小化 $\mathrm{KL}(q \,\|\, p)$

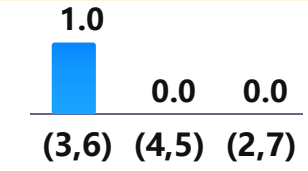
- A³BL 使用软标签交叉熵（Eq. 9），并通过 Eq. 7 和 Eq. 8 将候选概率归一化并聚合为每张图像的标签分布。

损失函数对比：硬标签（CE）vs. 软标签（KL）

硬标签（传统 ABL）

$$\mathcal{L}_{\mathrm{CE}}(y,p) = -\sum_k y_k \log p_k$$

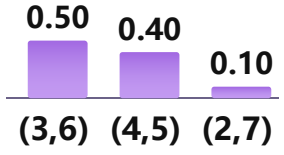
$y = [1,0,0]$



软标签（A³BL）

$$\mathcal{L}_{\mathrm{KL}}(q \,\|\, p) = \sum_k q_k \log \frac{q_k}{p_k}$$

$q = [0.50,0.40,0.10]$



A³BL 的核心改进

保留多候选的信息： 将单一的硬标签替换为由候选概率构成的软标签，充分利用感知模型的分布信息。

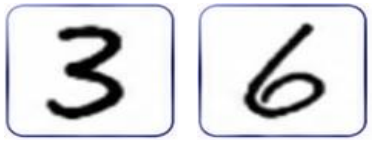
更合理的训练目标： 采用 KL 散度（等价于软标签交叉熵）进行优化，使模型在多候选上获得更平滑、稳定的监督信号。

研究问题背景（三）：A³BL 的局限——错误偏好仍会被放大

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

A³BL 的改进： 保留多候选，不再只选一个硬标签 | 但候选权重仍由感知模型给出 | 若初始判断有偏，软标签也会被带偏

一个具体例子（加法任务）



真实标签（隐藏）：(3,6)

任务标签：和为 9 | 知识规则：$a+b=9$

满足规则的所有候选序列：(3,6) (4,5) (2,7) (1,8) (0,9) …

上述候选均满足 $a+b=9$，但只有 (3,6) 是真实标签。

A³BL 如何赋权（依赖感知模型）

初始感知模型：尚不可靠

	4	3	2	1
第1张图 (3)	0.55	0.30	0.10	0.05
第2张图 (6)	0.55	0.30	0.10	0.05

候选权重（由上述预测得到）

(4,5)	0.55 → 错误但权重最高
(3,6)	0.30 → 真实但权重更低
(2,7)	0.10
(1,8)	0.05

诱导的软标签（用于监督训练）

	4	3	2	1
图1软标签 q_1:	0.55	0.30	0.10	0.05
图2软标签 q_2:	0.55	0.30	0.10	0.05

作为软标签监督，反映用于再训练

错误偏好如何被逐步放大

	迭代 t_0（初始）	迭代 t_1（一次训练后）	迭代 t_2（多次训练后）
(4,5)	0.55	0.68	0.82
(3,6)	0.30	0.22	0.12
(2,7)	0.10	0.07	0.04
(1,8)	0.05	0.03	0.02
	初始偏向错误候选	软标签训练进一步强化了 4 和 5	闭环反馈导致错误偏好放大

反馈闭环：赋权（选候选）→ 软标签（构造 q）→ 再训练（更新模型）→ 再赋权（更新模型）

GABL 的启示

来自 GABL

感知模型初始能力 / 置信度 —— 存在阈值 —— 阈值 τ

- 高于阈值（$\geq\tau$）→ 伪标签质量较高，模型可逐步改进
- 低于阈值（$<\tau$）→ 错误伪标签增多，ABL 闭环会让模型越来越差

当初始能力低于阈值时，ABL 会越训越差

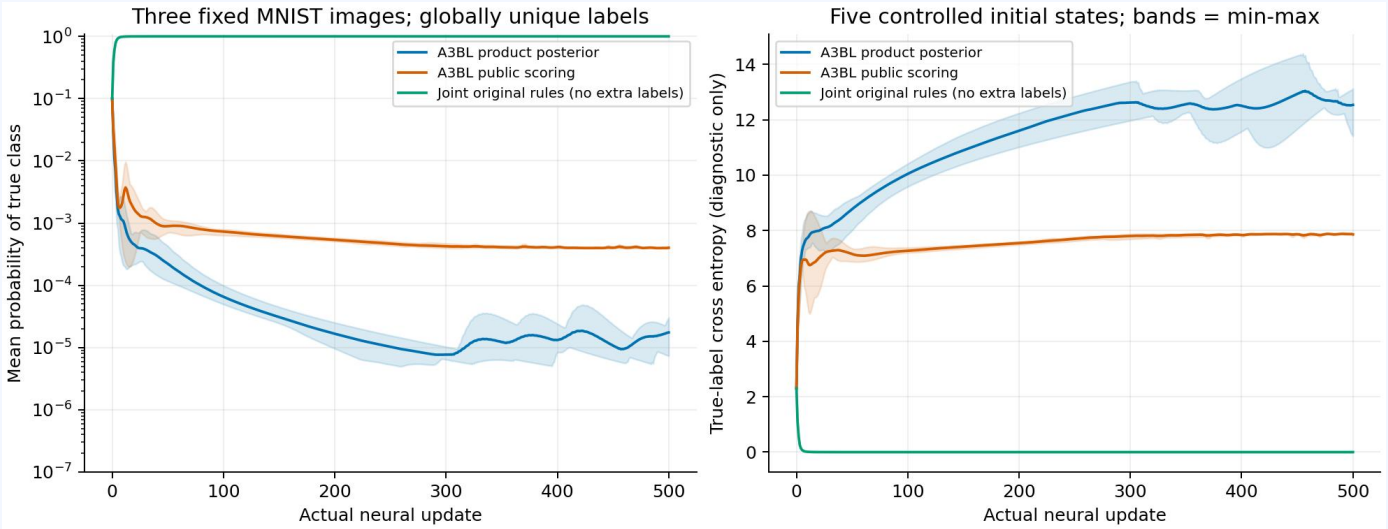
实验一：初始错误偏好会被进一步强化

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

三张MNIST图（真值1、2、3），三条加法约束；人为设置偏差后正常训练500步。

正确候选始终保留；两种赋权与联合原约束对照使用相同信息。

真类概率下降，诊断损失上升



5个匹配受控初态，均继续正常更新模型参数。

500步后的恢复结果

方法	三图全部恢复
正文赋权	0 / 5
公开赋权	0 / 5
联合原约束	5 / 5

同一真值1图像的概率

$0.10 \rightarrow 1.08\times10^{-6}$

正确候选仍在，但赋权偏向错误解释，实际训练会进一步强化这一偏差

受控小任务验证机制；完整MNIST普通加法随机冷启动仍达到98.86%。

实验二：关系偏斜导致整体性能下降

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

实验设置（控制变量）

- 训练图固定 50,000 张：0–9 每类各 5,000 张
- 每组训练 30,000 张方程，训练 50 轮
- 每类数字在每组中均出现 6,000 次
- 每张训练图曝光次数一致（1 或 2 次）
- 测试集、网络结构、训练预算保持不变
- 仅改变数字配对关系分布

为什么这样设置?

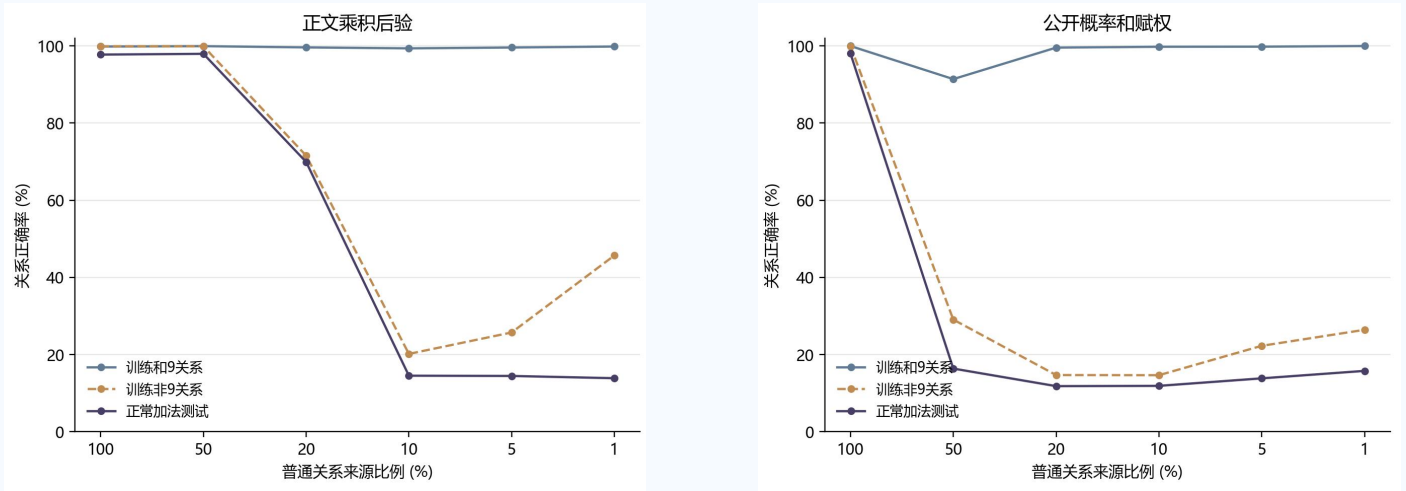
- 目的是隔离 “关系分布偏斜” 的影响。
- 实验中 0–9 各类图片数量保持不变，每类仍为 5,000 张。
- 改变的不是图像数据分布，而是监督信息分布：普通关系减少，“和为9”关系增多。
- 因此可排除 “性能下降仅因为数据分布不均匀” 这一解释。

只改变监督信息分布，
不改变图像类别数量。

六档设置

普通关系来源：	100%	50%	20%	10%	5%	1%
实际“和为9”比例：	10%	55%	82%	91%	95.5%	99.1%

六档关系分布：训练关系与测试任务



横轴：普通关系来源比例；每档为 3 个 seed 均值。

关键结果（以正文赋权为例）

数字识别准确率： 98.88% → 19.92%

正常配对加法准确率： 97.72% → 14.53%

- 关系越偏向“和为9”，
  对正常配对测试的泛化越差。
- 多数“和9”关系接近满分，
  但非9关系与正常测试显著下降。

结论：这是“关系分布偏斜”而不是“数字类别长尾”；在数字类别完全平衡时，A3BL 仍会因高频“和为9”关系导致整体性能显著下降。

为什么关系偏斜会暴露 A3BL 的问题?

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

机制分析： 高频弱约束关系 + 候选赋权依赖感知模型

1. 为什么“和为9”特别敏感?

- **对一位数加法，观测和 = 9 时，合法候选最多。**
- 例如：(0,9)、(1,8)、(2,7)、(3,6)、(4,5)、…、(9,0)
- 知识规则只能判断这些候选都合法，难以唯一确定哪一组才是真实标签。

观测和	合法候选组数
和 = 0	1 组
和 = 1	2 组
和 = 9	10 组（最多）
和 = 18	1 组

和为9：10组候选，消歧能力最弱

2. A3BL 在这类样本上如何工作?

图片1 图片2



观测：两图之和

→ 知识库给出全部合法候选

(0,9)、(1,8)、(2,7)、(3,6)、(4,5)、…、(9,0)

→ A3BL 按感知模型概率为候选赋权

(3, 6) （正确候选） 权重：0.26

(4, 5) （错误候选） 权重：0.58

→ 对每个位置求边缘，形成软标签

→ 用软标签继续训练感知模型

当知识约束本身不能充分消歧时，候选赋权会更依赖感知模型当前预测。

3. 为什么会出现错误强化?

1. 早期感知模型不可靠。
2. 在大量“和为9”样本上，错误候选也满足规则。
3. A3BL 给错误候选较高权重，生成偏向错误的软标签。
4. 再次训练后，模型更相信该错误候选。
5. 形成错误伪标签反馈循环。

例：

真实：

若模型偏向 4 和 5，则错误候选 (4,5) 也满足

4. 为什么非9关系纠正不过来?

高频“和为9”关系：占训练主导 ≈ 91%

低频非9关系：纠错信号不足 ≈ 9%

- 普通关系减少后，能排除错误映射的样本变少。
- 模型即使在“和为9”上表现很好，也可能没有真正学会正确的数字映射。
- 因此会出现：训练和9关系接近满分，但数字识别与正常加法测试显著下降。

满足知识规则 ≠ 学会真实概念

谢谢聆听