

INTSIG 合合信息™

# 从文档解析到智能抽取， DocFlow助力百业高效运营

智能解决方案事业部



# 01 合合信息公司简介

# 公司发展历程

上海合合信息科技股份有限公司 是行业领先的人工智能及大数据科技企业，致力于通过智能文字识别及商业大数据领域的核心技术、C端和B端产品以及行业解决方案为全球企业和个人用户提供创新的数字化、智能化服务。

18年

深耕人工智能

23亿

全球累计用户下载量

125项

国内外发明专利

16项

顶级AI竞赛世界冠军

30个

行业服务经验



# 业务体系：深耕智能文字识别+商业大数据领域

## C端产品：深受全球用户喜爱的效率工具



名片全能王



扫描全能王



启信宝

总月活1.5亿  
覆盖全球200+个国家和地区

## B端服务：AI+大数据赋能数字化转型



金融对公风控&营销  
解决方案



供应链大数据  
风控解决方案



政企大数据  
治理解决方案



“启信慧眼”  
风控营销SaaS平台



名片全能王  
企业解决方案

服务30+个行业、2000+全球企业

## 平台产品：



智能文字识别服务平台

+



商业大数据技术与资产平台

## 底层技术：

AI

模式  
识别

图像  
处理

神经  
网络

深度  
学习

STR

NLP

+

大数据

数据  
治理

隐私  
计算

知识  
图谱

# 智能文字识别：16次AI学术世界冠军

2019 ICDAR  
国际文档分析识别大会

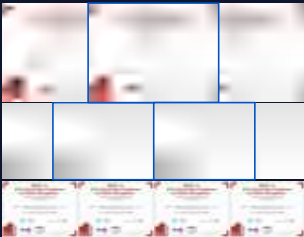
票据识别冠军  
表格识别竞赛冠军



2020 ICPR  
国际模式识别旗舰会议

获得11项冠军

具体包括：图片分类、文本检测识别、文本属性分类、坐标轴与图例分析、信息提取以及端到端信息提取等项目。



2023 ICDAR  
国际文档分析识别大会

“文本篡改检测”赛道冠军

2020 ICFHR  
国际手写识别前沿会议

签名识别与鉴别冠军  
姓名识别与鉴别竞赛冠军



2021 CSIG  
中国图象图形学学会

科技进步奖二等奖

\*鉴定委员会认为该项目在复杂场景文档图像识别与理解技术方面取得了创造性的成果，整体技术达到国际先进水平。



# 持续领跑：125项发明专利

## AI专利

125项  
发明专利授权

- 境内发明专利授权78项
- 境外发明专利授权39项



# 生态合作伙伴专业认证

SAP

银牌合作伙伴



微软

金牌合作伙伴  
MPN能力证书



华为

昇腾万里伙伴计划ISV伙伴  
COMPATIBLE技术认证

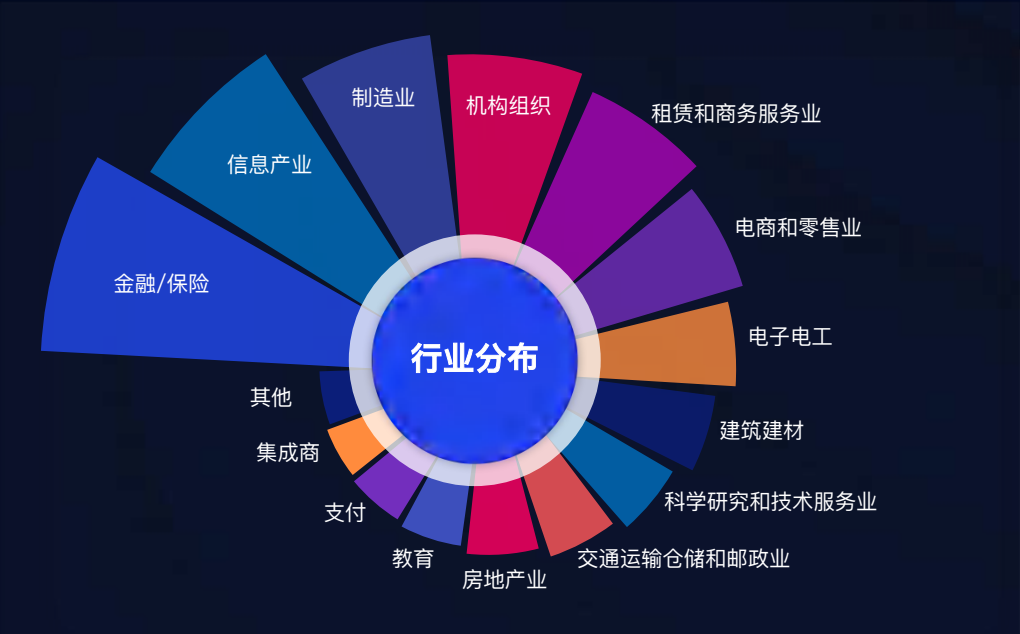


# 合合信息产品生态辐射全球市场

B端业务覆盖国家



企业行业分布



主要产品

Camcard for SF  
名片管理系统

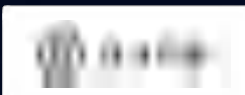
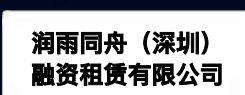
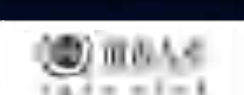
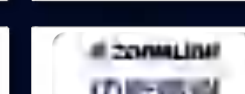
Camcard API/SDK  
名片识别 接口/工具包

Camscanner API/SDK  
扫描识别 接口/工具包

Camcheckout API/SDK  
银行卡识别 接口/工具包

ID Recognition  
证照识别

# 合作客户：已服务众多细分行业头部客户

|       |                                                                                     |                                                                                     |                                                                                      |                                                                                       |                                                                                       |                                                                                       |                                                                                       |                                                                                       |
|-------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------|
| 银行    |    |    |    |    |    |    |    |    |
| 证券    |    |    |    |    |    |    |    |    |
| 融资租赁  |    |    |    |    |    |    |    |    |
| 国央企   |    |    |    |    |    |    |    |    |
| 保险    |    |    |    |    |    |    |    |    |
| 制造业   |    |    |    |    |    |    |    |    |
| 医药医疗  |    |    |    |    |    |    |    |    |
| 能源    |   |   |   |   |   |   |   |   |
| 物流&地产 |  |  |  |  |  |  |  |  |
| 汽车    |  |  |  |  |  |  |  |  |

## 02 合合信息AI产品矩阵

# TextIn — 7+1+N 产品架构



# 通用基础文字识别

## 通用文字

- 支持**52**国语言的文字识别和提取，支持Latex公式识别、文本段落识别。



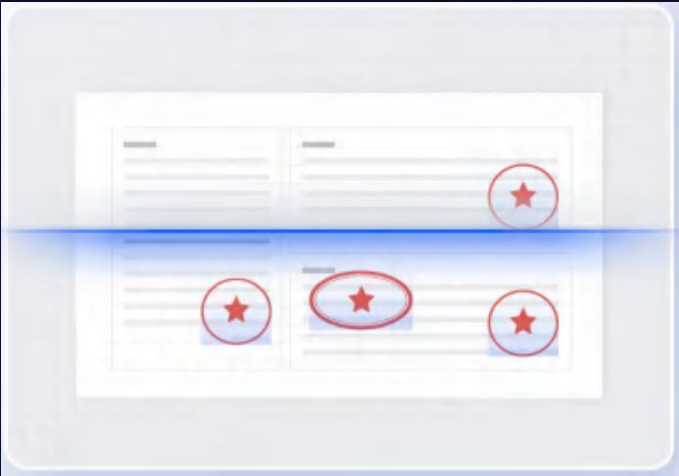
## 通用表格

- 支持识别图片/PDF格式文档中的表格内容，包括**有线表格、无线表格、合并单元格**表格。



## 印章检测

- 支持印章**存在性检测、形状识别、颜色识别、章面文字提取**



## 手写体识别

- 支持签名**存在性检测、中文签名**识别提取



# 标准卡证识别

- 特定场景的智能文字识别能力，针对场景证件的结构化提取有超高准确度，满足特定场景识别与结构化提取的业务需要，超过60+个标准智能文字识别引擎、超高识别精度，针对全社会通用的卡片、证件类型都做了特定的优化和提升。

## 个人卡证



## 企业证照



## 资产证照



# 票据识别

- 国内通用票据识别，是合合信息的票据识别产品系列拳头产品，支持**23大类**国内通用票据的识别与结构化信息抽取，提供高精度识别、快速识别、**多票混贴切分**、**旋转矫正**等一系列实用特性。

## 多票混贴，一次识别



## 复杂版式，快速识别，精准抽取



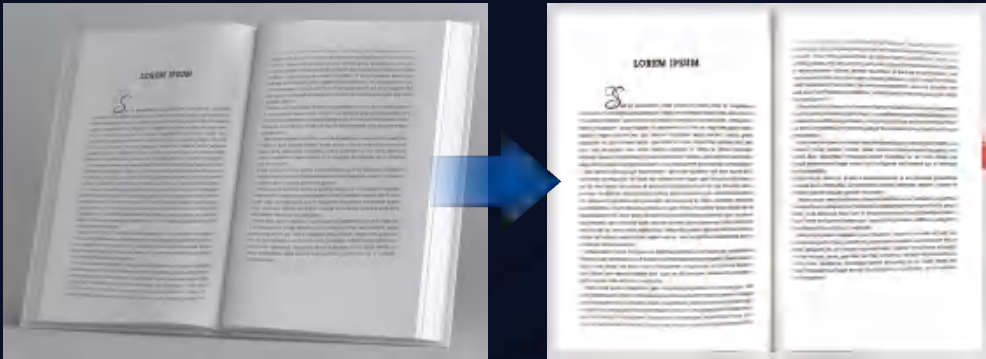
## 旋转票样，无忧识别



# 图像智能处理

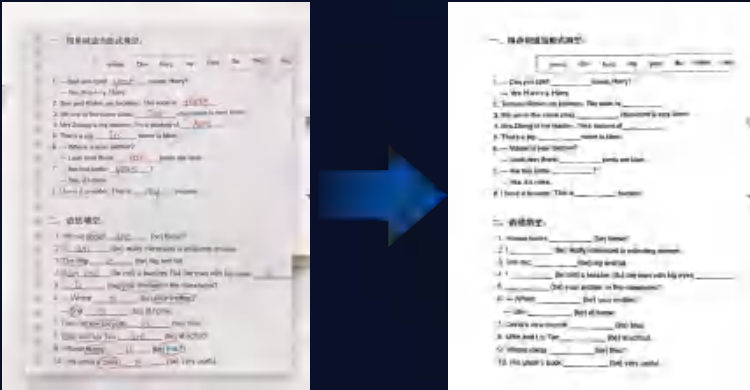
## 切边增强

- 图像切边增强产品，可以自动识别将图像中的重要内容，将其切出，并校正回矩形。
- 该产品不仅能处理拍摄视角带来的畸变，也能处理旋转、曲面图带来的畸变。



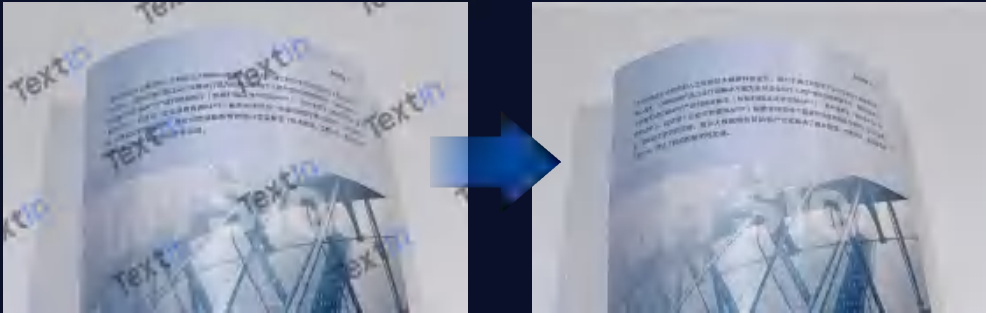
## 手写体擦除

- 手写擦除产品具备手写字体识别、擦除能力，针对特定的场景具有良好的效果。



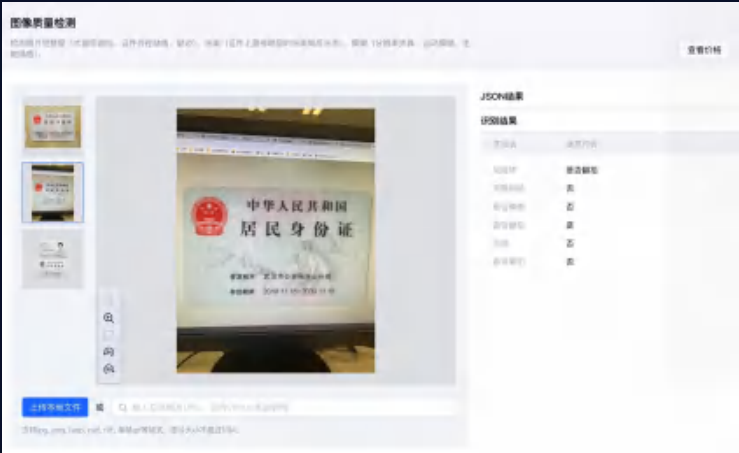
## 水印去除

- 水印去除产品可以去除图像中插入的水印，提高存档图像质量，同步提高后续识别流程的准确率。



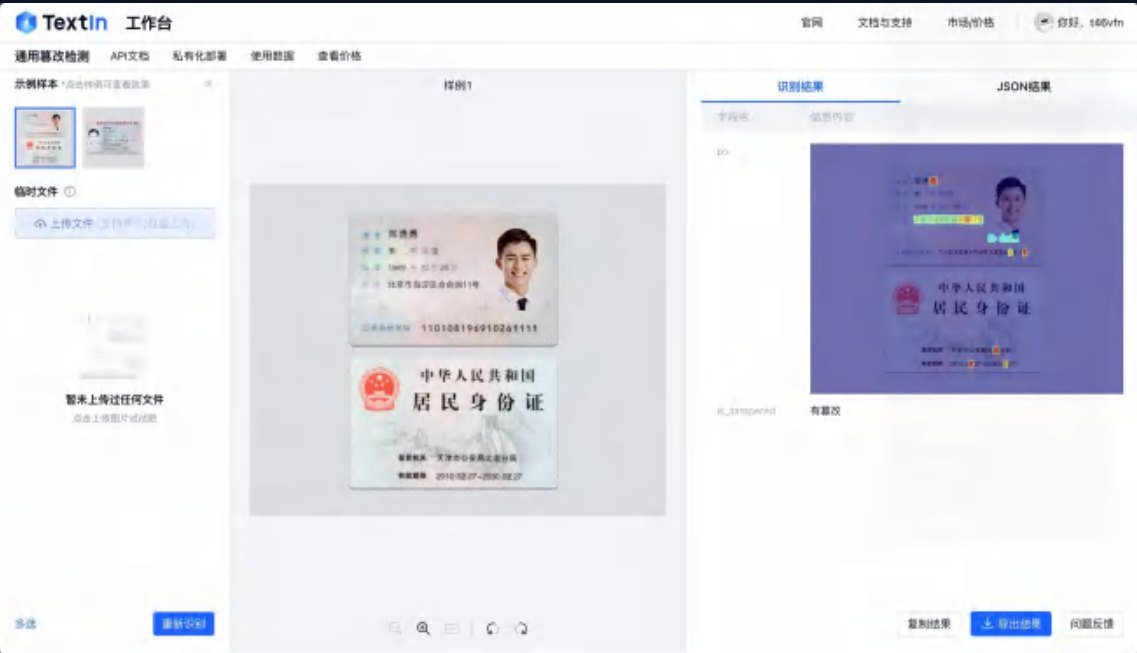
## 图像质检

- 过曝、欠曝识别
- 残缺识别
- 翻拍识别（屏幕翻拍、截图、打印复印）
- 模糊识别
- 光斑识别



## 通用篡改检测

- 通用篡改检测，应对PS、黏贴、涂改等一系列围绕文档、卡证、票据的图像篡改行为



复制移动

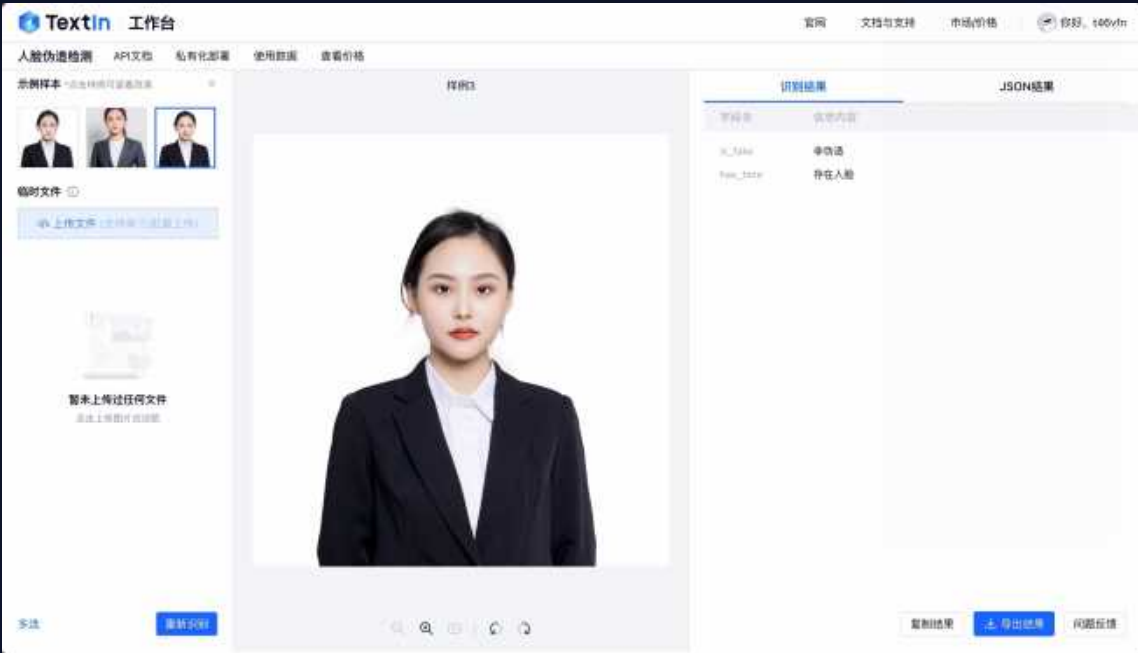
擦除

拼接

重打印

## 人脸伪造检测

- 人脸伪造检测，应对PS、黏贴、Deepfake技术等一系列围绕人脸的篡改和伪造行为



五官伪造

水印

贴图

绿边

# 大模型加速器：文档解析+ TEXT EMBEDDING+智能文档抽取



## 文档解析

支持将任意格式的文件（图片、PDF、Doc/Docx、网页等）解析为Json或Markdown格式，以对LLM最友好的方式呈现。

## Text Embedding

Embedding对于各种不同的下游NLP任务十分重要，如分类、聚类、检索、句子相似性判断等。合合信息发布的文本向量化模型 `acge_text_embedding` 在中荣获MTEB中文榜单（C-MTEB）第一名的成绩



## 智能文档抽取 (OpenKIE)

提供泛场景、开放域、开箱即用的图像文档信息抽取能力，其基于合合信息强大的文字识别能力，结合了文档解析、文档检索和自研大模型三项核心技术。众多场景超过gpt4和gpt4o

# 智能文档抽取

合合信息TextIn智能文档抽取产品，依托合合信息自研的垂直领域语义模型，并结合了合合信息强大的文字识别、文档解析、文档检索和文本生成四项关键技术，让计算机模拟人类的推理方式，来识别在训练阶段从未见过的新事物，实现开箱即用的“零样本”抽取，让AI触手可及。

开箱即用

无需标注训练

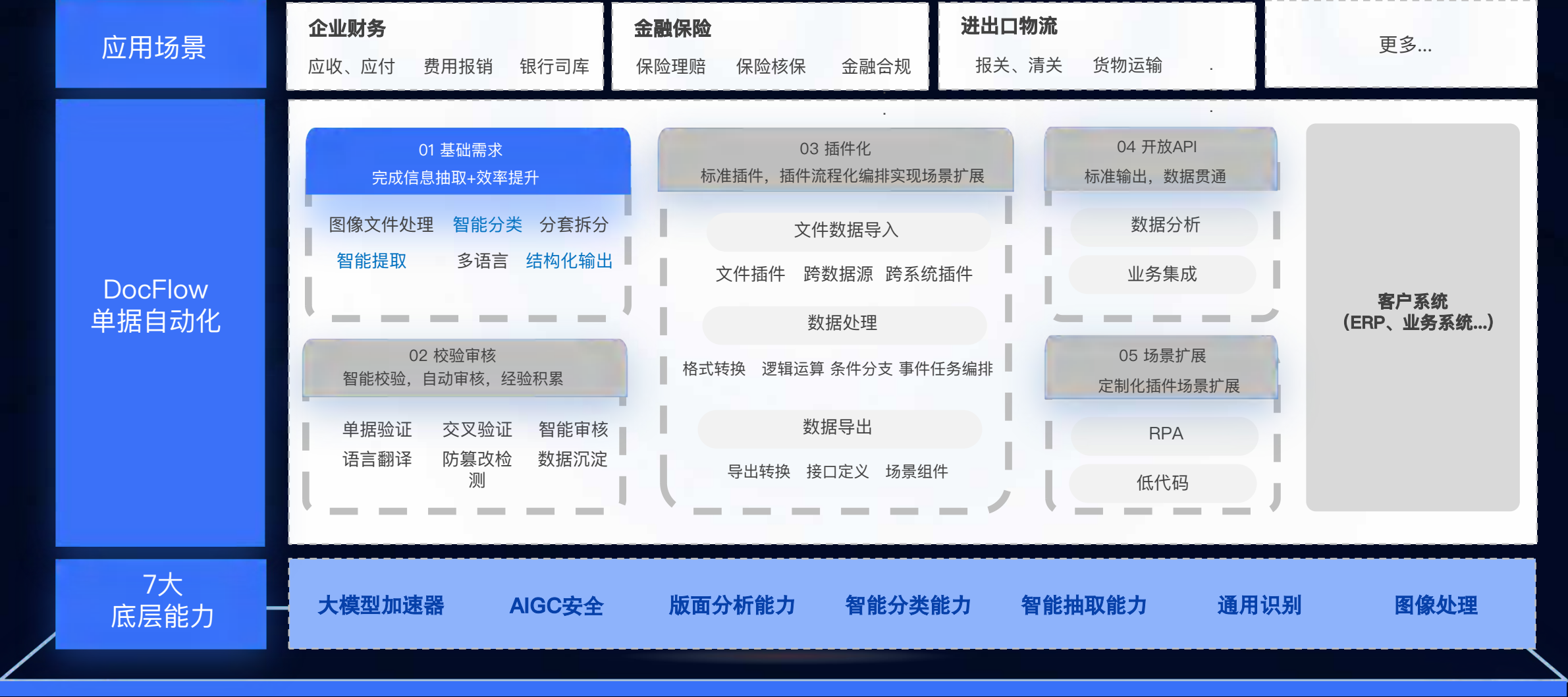
直接配置抽取字段





## 03 DocFlow 产品介绍

# DocFlow，真有用的智能文档数字化管理系统



# DocFlow产品特点

一站式智能单据数字化系统，解决各类文档分类及结构化，打通了信息化建设最后一公里

## 标准化，开箱即用

- 平台产品开箱即用
- 集中管理归档组织单据

## 分类、抽取字段灵活配置

- 分类样本、抽取字段自定义配置，免训练
- 分类结果路由至各业务系统

## 数据核验、导出

- 抽取字段准确性校验
- 基础核验规则配置
- 多格式导出适配

## 易于集成

- 单点登录，页面嵌入
- API接口开放项目周期短
- 快速集成ERP、RPA

# DocFlow：产品流程



# DocFlow: 文档分类

开箱即用，实现自动分类，只需上传少量样本/分类关键字设置，无需标注训练，即可实现文档自动分类

工作空间设置

名称

默认工作空间

描述

默认工作空间

创建分类

purchase\_order

Purchase Order

单据样本

文件格式: jpg, jpeg, png, pdf, 单个文件不超过10M, 单次上传不超过20个

| 序号 | 文件名 | 状态 | 操作 |
|----|-----|----|----|
|    |     |    |    |

invoice

Invoice

credit\_note

Credit Note

保存

单据类型详细设置

名称: purchase\_order

描述: Purchase Order

基本字段信息 (20)

Order Number

采购订单编号

Due Date

付款日期

Vendor Order ID

供应商订单编号

Supplier Name

供应商名称

Place of Supply

供应地点

Supplier Address

供应商地址

Supplier Email

供应商邮箱

Customer ID

客户ID编号

Customer Name

空间设置-配置文档类型、字段



Domestic Payment

VAT Fapiiao

Account Statement

Purchase Order

Receipt Report

Payment Application Form

International Payment

Duty Payment

Custom Declaration

Receipt

Monthly/ Annually Payment

Invoice

自动分类结果

# DocFlow: 文档拆分

- 对于混合扫描的文件，可自动切图/拆页，并按照预设的文档类型实现文档分类
- 支持人工通过在页面拖动的方式，完成文档分类复核



• 文档类型1

• 文档类型2

• 文档类型3

# DocFlow：识别结果&人机复核

- 按照预设的字段，抽取各类非固定版式单据中的关键信息
- 提供复核页面，人工可通过页面点击、框选或手动编辑的方式，实现快速的复核及一键补录

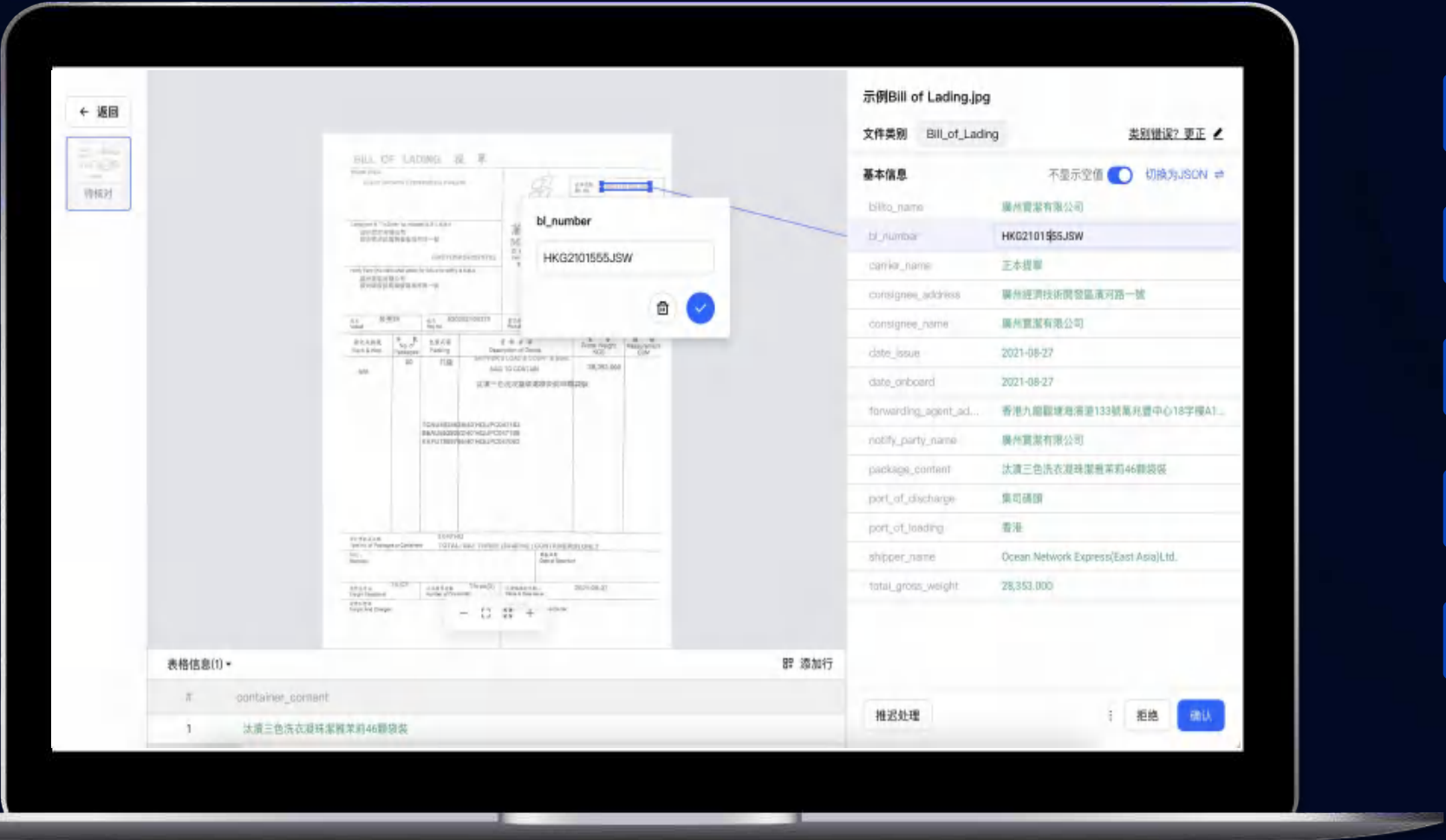
抽取结果展示

人工复核、调整文档类别

页面点击、框选补录

手工编辑抽取结果

确认、导出文档抽取结果



# DocFLow：强大的文档抽取能力

合合信息TextIn智能文档抽取产品，依托合合信息自研的垂直领域语义模型，并结合了合合信息强大的文字识别、文档解析、文档检索和文本生成四项关键技术，让计算机模拟人类的推理方式，来识别在训练阶段从未见过的新事物，实现开箱即用的“零样本”抽取，让AI触手可及。

开箱即用

无需标注训练

直接配置抽取字段



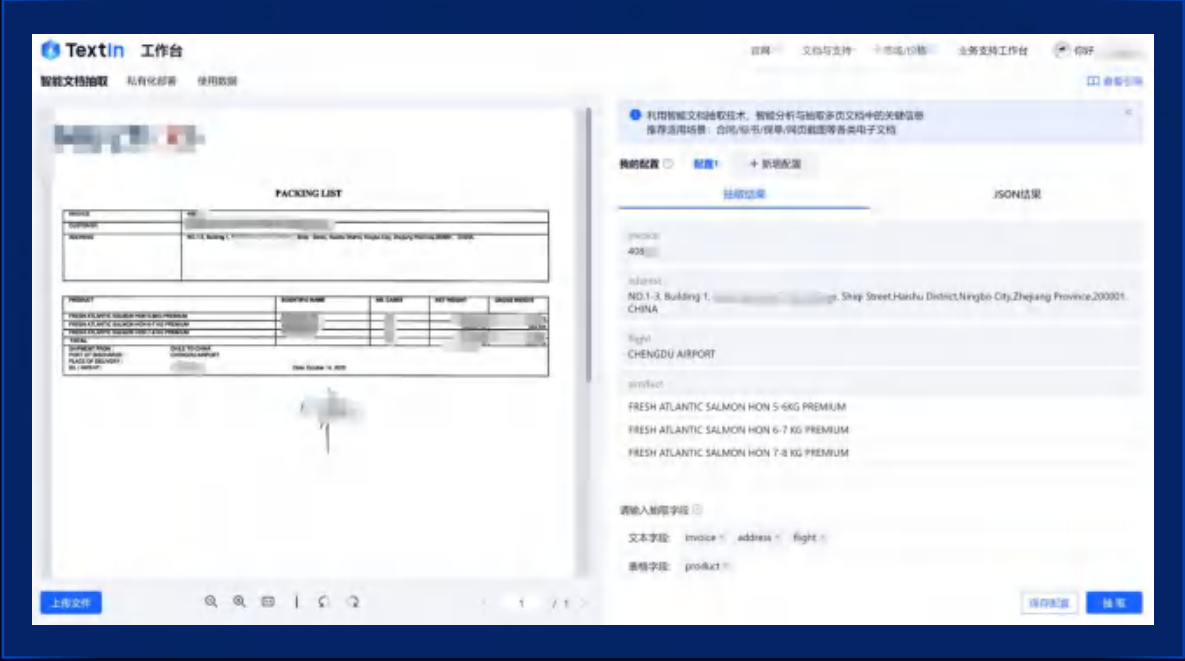
# 优秀的泛化性，兼容不同版式文档

合合信息提供“开箱即用”的方式处理国际结算业务单据抽取任务，具备处理能力强、高精度预测性能、强大的迁移学习能力等多重特点，仅需配置相关抽取字段，即可完成结构化抽取。

Packing list  
版式1



Packing list  
版式2



# 全要素抽取，兼顾短文本与长文本

TextIn智能文档抽取支持双层PDF电子件、拍摄件、扫描件等不同格式的文档，手写体、印章、表格等不同类型的元素的智能抽取，兼顾短文本与长文本，既支持单页的非标卡证、票据、表单，如海外invoice、国际信用证、电汇凭证、不动产权证等，也支持几十甚至上百页的长文档，如购销合同、借款合同、基金合同等。

纸质文档  
拍摄件

纸质文档  
扫描件

双层PDF  
电子件

WORD

手写体

印章

表格

50+语言文  
字

  
不动产权证

  
银行回执单

  
电汇凭证

  
商铺小票

  
进出口货物报关单

  
购销合同

  
借款合同

  
基金合同

  
融资租赁合

  
债权转让协议

  
海外电话单

  
海外账单

  
飞机行程单

  
国际信用证

  
银行流水单

  
展期协议

  
抵押合同

  
股权质押协议

  
财产保全申请书

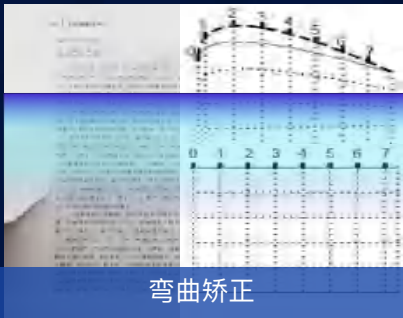
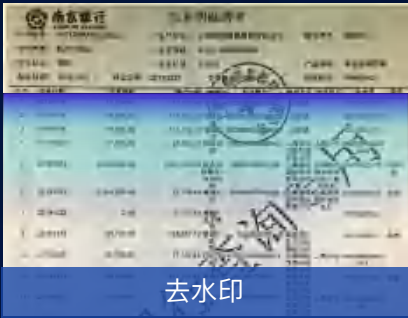
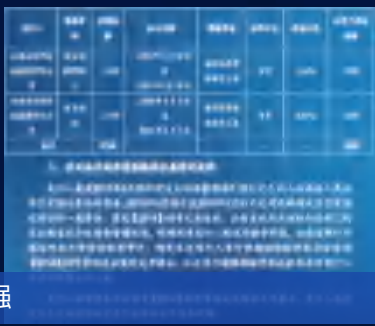
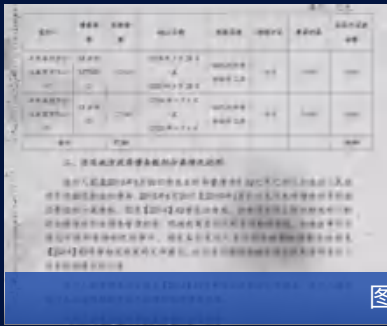
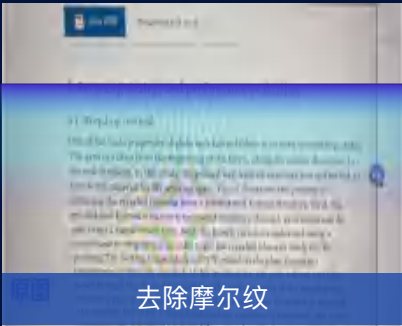
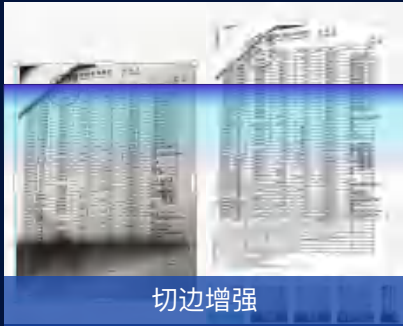
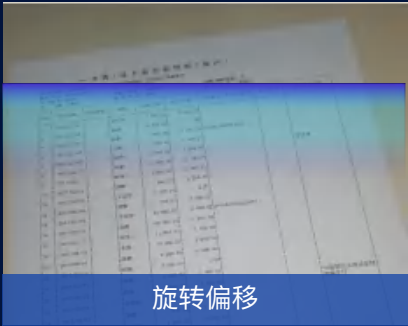
  
起诉状

短文本（如：单页卡证、票据、表单）

长文本（如：合同、法律文书、政策文件）

# OCR-多种复杂场景下的高精度识别

## 解决传统挑战



- 抖动模糊
- 歪斜
- 反光
- 远距离低像素
- 低光照噪点
- 透视变换
- 对焦模糊
- 形变
- 光照不均
- 阴影
- 复杂背景
- 复杂字体

## 04. 应用场景

# 单据串联企业 业、财、法管理，打通业务系统间的信息流

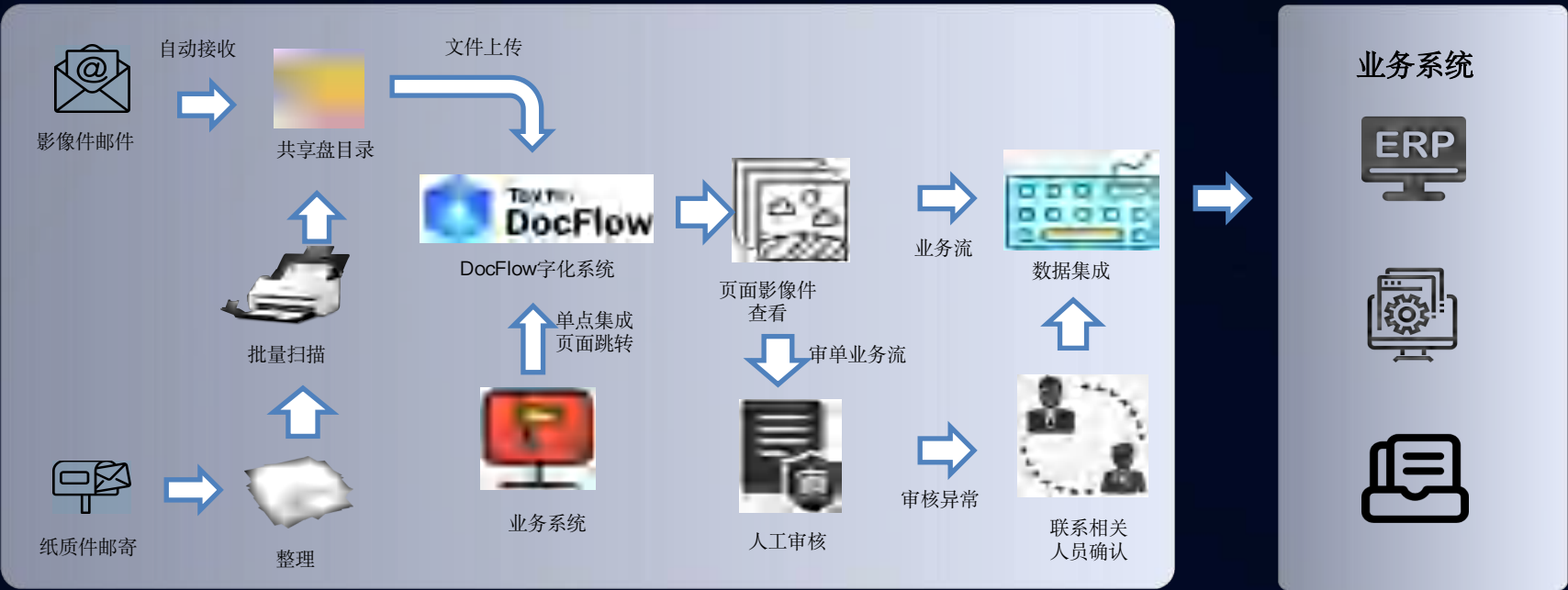


# 场景一：物流报关

Before



After



## Challenge

- 海量的非固定版式文件
- 多种来源混合输入
- 自动分类，自动校核



## Value

- 自动获取所有文档并进行结构化抽取
- 加快通关速度，危险货物申报数字化处理
- 加快运费审计和付款
- 99.5% 数字化流程优化

# 场景一： 物流报关—抽取能力

物流、报关行业需要处理大量非固定版式单据，包括提货单、INVOICE、空运舱单、交货单、运输标签、货物保单、报关单、装箱单、危险品申报单、PO订单以及许多其他非结构化文件类型。

## 例1： 装箱单

装箱单用于使用编码和每批货物中包装的重量、体积和数量信息来识别运输的包装，当卖家创建发票时，它是确认包裹与订单一致的关键工具



## 例2： 提单

提单中包含三个关键信息，即产品名称、商品收据以及条款条件。它是托运人和承运人之间具有约束力的法律协议。根据本法律协议的条款和情况，货物必须到达目的地。



| 文档类型    | 字段                                                         |
|---------|------------------------------------------------------------|
| 提单      | 船公司、提单号、收货人、被通知人、船名/航次/船旗、装货港、卸货港、装船日、更新日期                 |
| 交货单     | 供应商名称、供应商地址、供应商联系方式、收货人名称、收货人地址、物品数量、物品重量、物品描述             |
| INVOICE | 日期、结算方式、信用证/保函编号、来单币种及金额、开征日期、索汇币种及金额、期限、合同号、发票号、提运单号、合同金额 |
| 运输标签    | 收件人的姓名和地址、供应商/货运代理的名称和地址、联系电话号码、物品说明、其数量和/或重量、装运等级         |
| 货物保单    | 保险公司名称、地址、联系方式、保单编号、保险类型、保险金额、被保险人、保险期限、投保货物               |

| 文档类型   | 字段                                                            |
|--------|---------------------------------------------------------------|
| 报关单    | 原产国、货物描述、成本、保险和运费 (CIF) 详细信息、海关分类编号、关税等                       |
| 装箱单    | 起运时间、预计离港日期、运单号、预计到港日期、商品品种、商品数量、包装规格、商品净重、商品毛重、买方、运输方式、净重、总计 |
| 危险品申报单 | 寄件人和收件人的地址和联系方式、货物名称、数量、重量、包装组、放射性、处理信息                       |
| PO订单   | 购买编号、产品信息、产品数量、产品重量、订单日期、商品价格、付款条件                            |
| 空运货物舱单 | 运单编号、产品的详细信息、每批货物的重量/数量                                       |

# 场景二：物流集散

- 流程改造前

在传统的录单业务流程中，上港兴宝需要将最终客户提供的提货单扫描打印为纸质文件，分配给录单员，再由录单员在WMS系统中手动录入信息。由于采用“数字代码”的方式录入信息，需要录单员牢记常用字段对应的数字代码，新员工培训成本高，需要较长时间才可达到熟练录入程度。
- 流程改造后

系统可智能识别并抽取提货单中的关键字段信息及印章信息，自动录入WMS系统，录单员仅需进行复核确认，录入审核流程大幅度缩短，实现了智能化的单证信息采集、审核、确认与传输。经上港兴宝试运行效果验证，整体工作效率相较于流程改造前提升了80%。



# 场景三：物流仓储

在零售行业售后退货场景中，退货商品外包装常常存在破损，需要按照B品升A品流程，对商品外包装进行更换，重新上架。

**流程改造前** 人工采集或向品牌方索要信息-批量匹配信息-批量打印标签-分播标签-根据标签挑选新包装。该流程主要依靠人工，存在数据库信息不完整、数据信息有误、与品牌方沟通时效长、手工输入效率低等问题。

**流程改造后** 依托合合信息智能文档抽取能力，仅需4步，就能完成货品外包装标签信息的识别、抽取、校核与更新。

## 1.图像采集

进入系统页面创建对应任务，系统自动调度高拍仪采集外包装标签图像，可兼容多组照片。



## 2.信息抽取

对采集到的图像进行关键信息识别（货号、logo、颜色、材料、生产日期、经销商、制造商等）和字段逻辑比对。



## 3.信息校核

用户可针对系统标识出的不匹配字段进行二次框选识别或手动编辑，保存后数据同步更新至品牌基础标签数据库。



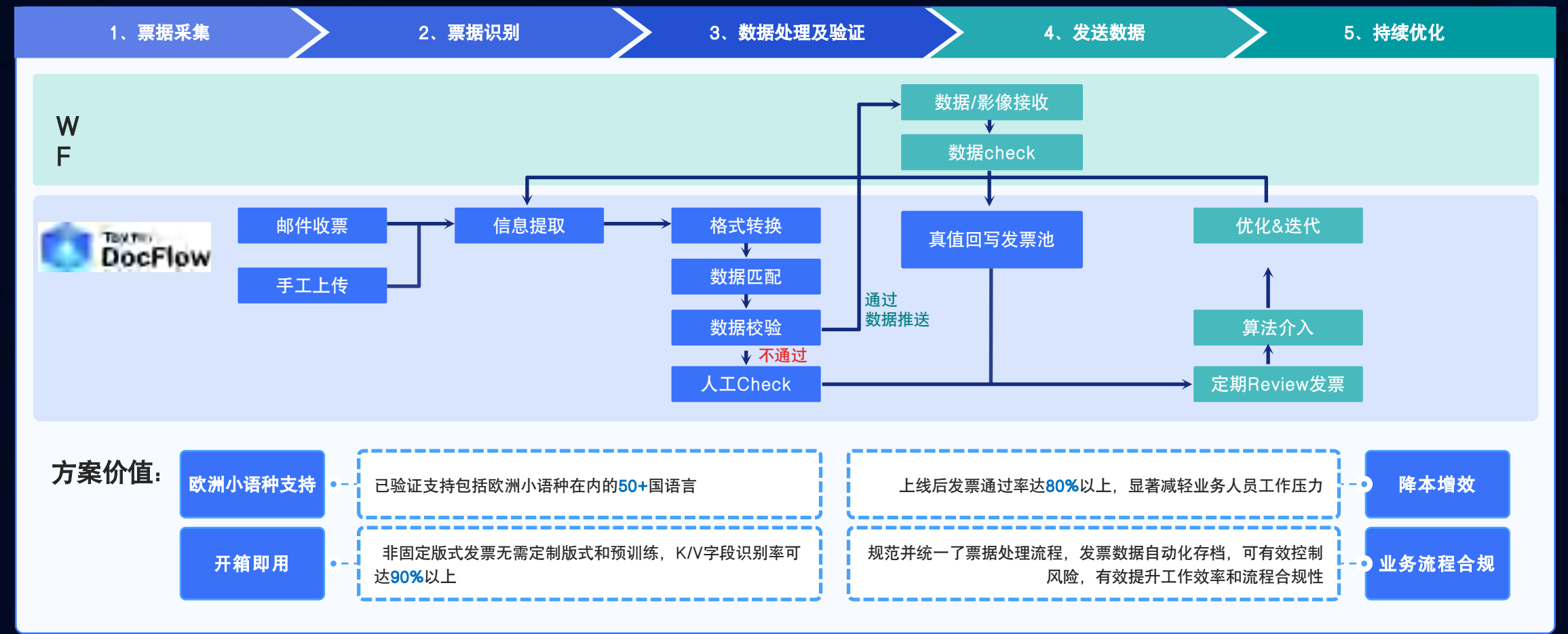
## 4.数据生成

支持按照企业标签系统的excel模板生成数据，将数据导入至企业标签系统生成对应的包装标签，完成打印及贴标流程。



# 场景四：财务票据自动化

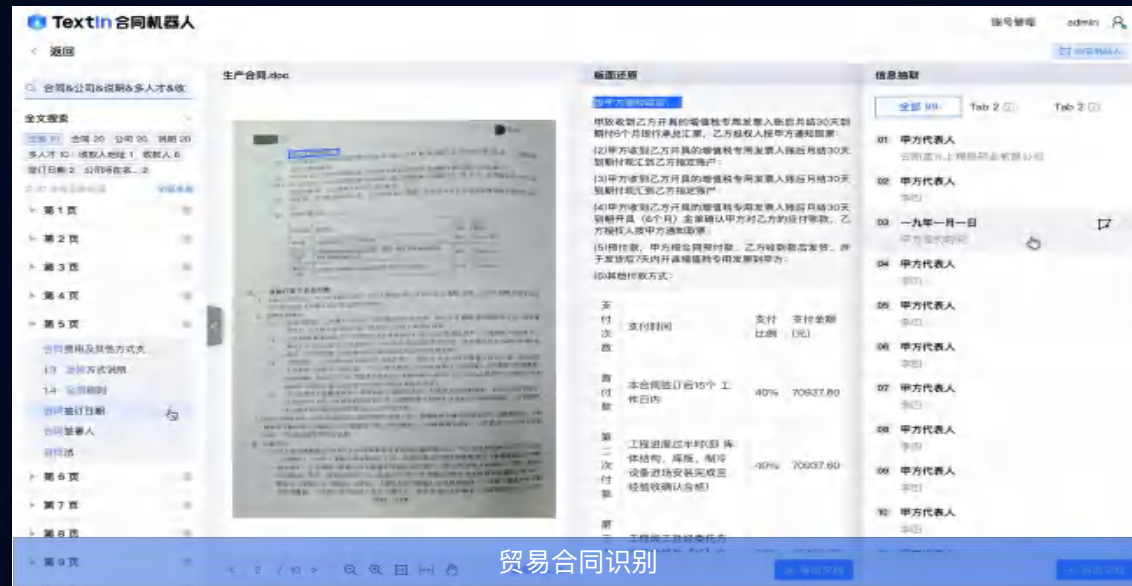
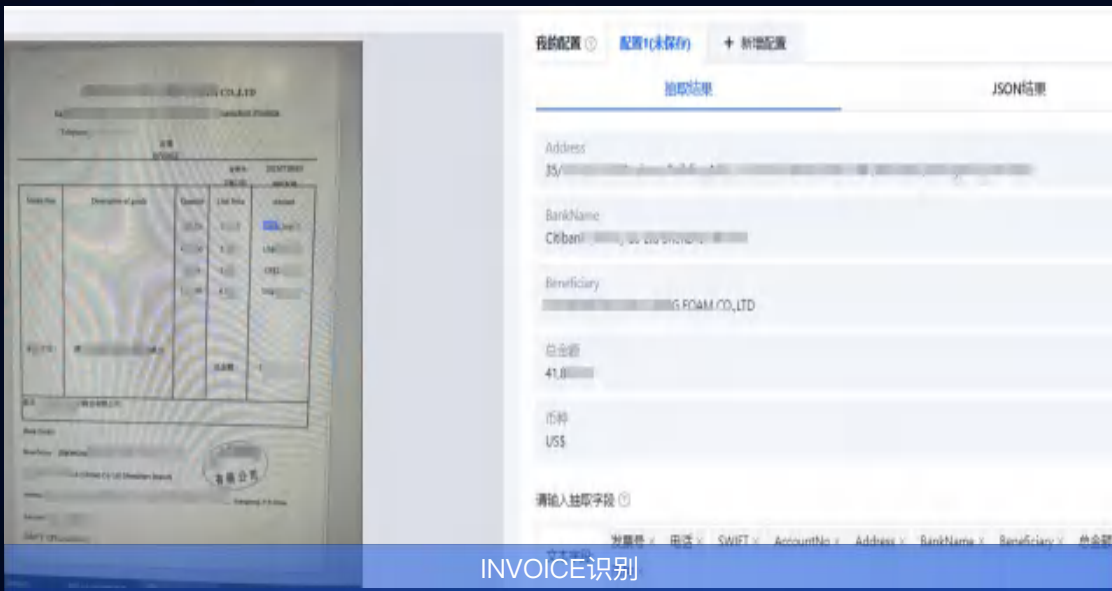
合合信息票据智能处理解决方案，成功为企业打破票据处理过程中语言、版式的障碍，并实现从传统手工录入到全流程自动化管理的转变，为某国产新能源汽车挺进全球化“深海”深度赋能、持续助力。



## 场景五：国际结算

**进出口托收/代收  
业务抽取**

## 国际保理业务抽取



# INVOICE



- 日期、结算方式
- 信用证/保函编号
- 来单币种及金额
- 开征日期
- .....

## 提单



- 船公司
- 提单号
- 收货人、被通知人
- 船名/航次/船旗
- .....

## 保险单据



- 日期、结算方式
- 信用证/保函编号
- 来单币种及金额
- 开征日期
- .....

## 贸易合同



- 日期、结算方式
- 信用证/保函编号
- 来单币种及金额
- 开征日期
- .....

## 装箱单



- 日期、结算方式
- 信用证/保函编号
- 来单币种及金额
- 开征日期
- .....

# 报关单



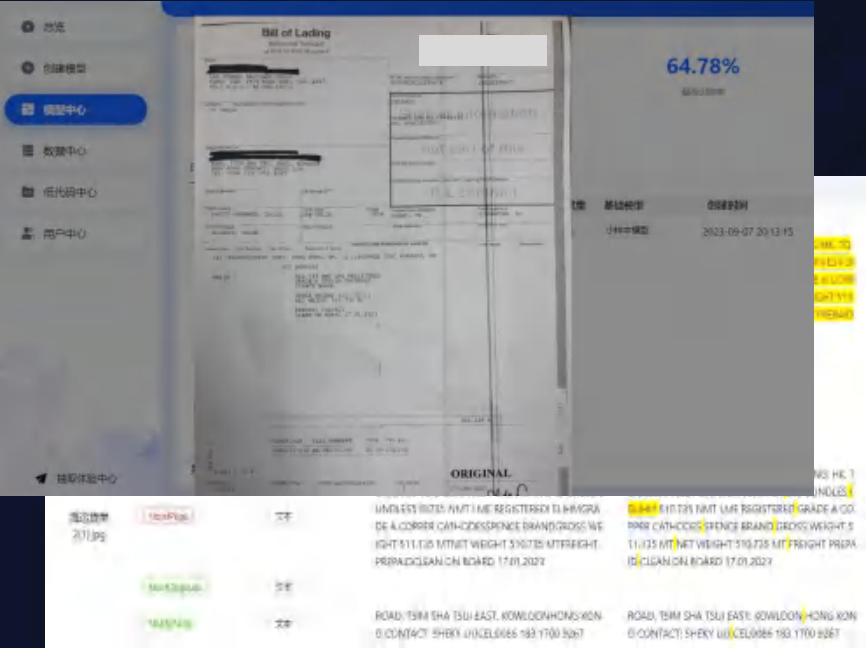
- 日期、结算方式
- 信用证/保函编号
- 来单币种及金额
- 开征日期
- .....

# 场景五：国际结算

## 最佳实践某银行交易银行部-基于LLM的国际结算单据处理

### 基于深度学习训练平台的建模方式

识别率：国际物流业务场景下，样本字段识别率60%~70%  
训练成本高、周期长

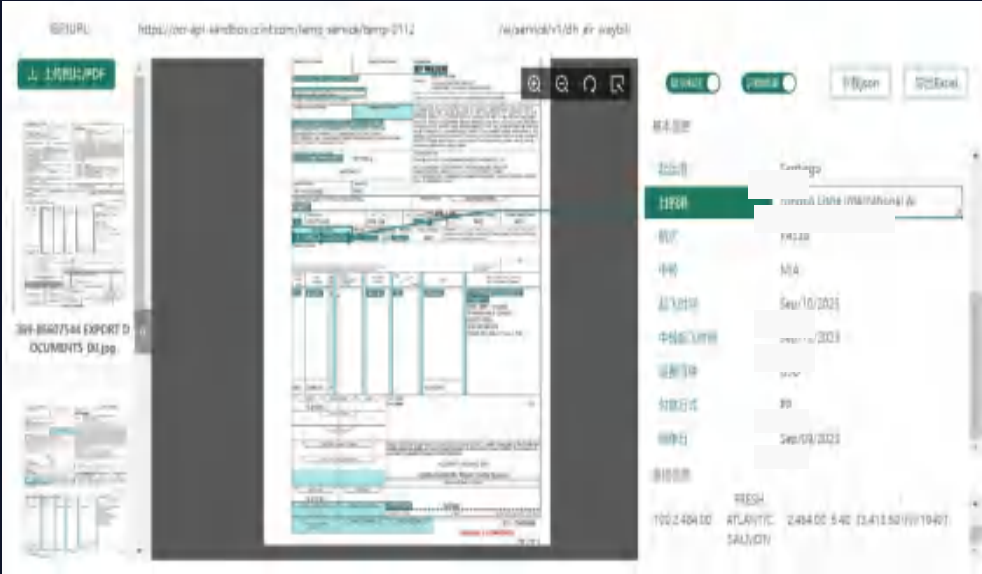


- 提单、运单、PO订单、海外INVOICE等

对比

### 基于大模型+OCR+规则的通用抽取能力

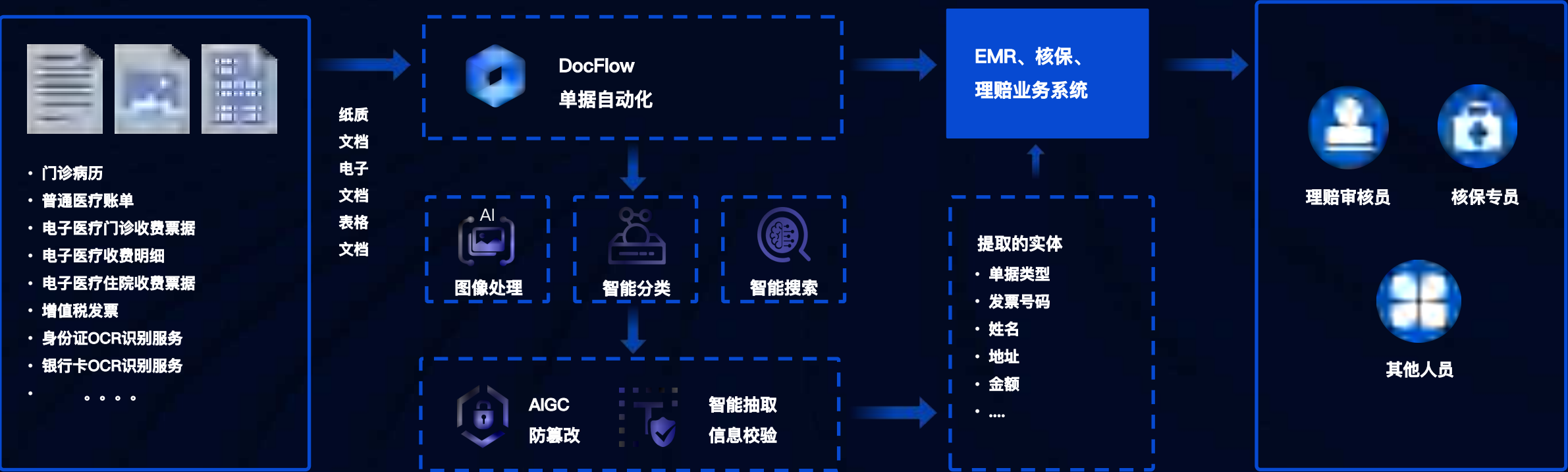
识别率：国际物流业务场景下，样本字段识别率85%以上  
少样本或者零样本、周期短



- 提单、运单、PO订单、海外INVOICE等

# 场景六：医疗险核保、理赔

## • 医疗保险场景中的单据“流”



### 方案 优势

- 提供开箱即用和用户自定义处理能力，支持自动定义配置
- 综合图像处理能力如图像增强、多图切分、去除水印、摩尔纹等多项技术解决采集图像质量，提升识别精度
- 通过Embedding特征向量，少量样本实现精准快速的分类，结合Generative AI快速精准的从文件中提取结构化数据
- 便捷的插件扩展和开放的集成API

INTSIG 合合信息™

感谢聆听

