上海交通大学学位论文

基于多源信息融合的文字识别及在工业场景中的应用

姓　名：邓国伟
学　号：120032910121
导　师：陈彩莲教授
学　院：电子信息与电气工程学院
学科/专业名称：电子信息
申请学位层次：工程硕士

2023 年 3 月


20003506

A Dissertation Submitted to
Shanghai Jiao Tong University for Master Degree

Multi-Source Information Fusion-Based Text Recognition and Its Applications in Industrial Scenes

Author: Deng Guowei
Supervisor: Prof. Chen Cailian

Department of Automation,
School of Electronic Information and Electrical Engineering
Shanghai Jiao Tong University
Shanghai, P. R. China
March, 2023


20003506

摘 要

随着我国智能制造产业升级进程和数字化转型不断加快，基于机器视觉的文字识别需求日益激增，场景文字识别技术受到了学术界和产业界的广泛关注。由于工业现场环境复杂，获取高质量的文字图像存在极大困难。仅依靠单源图像输入中的信息难以准确、稳定识别工业文字，这使得工业场景的文字识别任务更具挑战。因此，本文致力于在文字识别方法中引入文字语义、多视角图像等多源信息，利用多源信息之间的互补性提取更准确的文字特征表示，并设计抗干扰能力强的工业文字识别方法。论文的主要工作总结如下：

其一，本文引入了文字语义信息，提出了一种基于工业语义与视觉融合的文字识别方法（ITScanner）。本文首先归纳总结了工业文字的分段式语义、段间独立性和段内关联性三种典型语义模式，解决了工业文本词库量大，信息密度高导致的语义建模难的问题。以此为理论基础，本文设计了一种基于工业语义与视觉融合的文字识别方法。该方法分别从语义信息和图像信息两方面提取文字特征，能够有效识别模糊图像中的工业文字。实验表明，ITScanner 在本文收集的多个工业场景数据集中均优于现有文献的方法。

其二，本文提出了一种基于多视角特征融合的文字识别方法（ITViewer）。为了区分和利用多视角图像之间的冗余性和差异性，本文提出了一种基于自注意力机制的方法来完成不同视角间特征的对齐与融合。由于多视角图像的采集和标注十分耗时耗力，本文基于相机成像的基本原理提出了一种多视角文字图像数据集的生成方法。实验表明，相比于现有方法对于单源图像的依赖性，该方法能够同时关注到多视角图像中的文字特征，具有较强的抗干扰性。

其三，本文立足于实际工业应用场景，以本文设计的工业文字识别算法 ITScanner 和 ITViewer 为核心步骤，设计了一套集文字检测、跟踪、识别的工业文字识别系统。为了解决系统实时运行的问题，系统采用了一种预识别加精识别的框架，通过预识别迅速、粗略地观察所有的文字区域，随后通过 ITScanner 和 ITViewer 精确识别高置信度区域，得到最准确的识别结果。

以上研究表明，基于多源信息融合的文字识别方法具有较强的鲁棒性和抗干扰性，在工业文字识别任务上具有显著的优势，是文字识别技术在工业场景落地的一次良好实践。

关键词：文字识别，工业现场，多源信息融合，文字语义，多视角图像

ABSTRACT

As China's intelligent manufacturing industry upgrade process and digital transformation continue to accelerate, the demand for machine vision-based scene text recognition is increasing, and scene text recognition technology has received widespread attention from academia and industry. Due to the complex environment of industrial sites, it is extremely difficult to obtain high-quality text images. In this case, it is difficult to accurately and consistently recognize industrial text by relying only on the information in a single-source image input, making it more challenging to recognize texts in industrial scenes. Therefore, this paper is devoted to introducing multi-source information such as text semantics and multi-view images into text recognition methods, extracting more accurate text feature representations using the complementary between multi-source information, and designing industrial text recognition methods with strong anti-interference capability. The main work of the paper is summarized as follows:

Firstly, this paper introduces text semantic information and proposes a text recognition method (ITScanner) based on the fusion of industrial semantics and vision. Three typical semantic patterns, including segmental semantics, inter-segment independence and intra-segment correlation of industrial text, are summarized in this paper to solve the problem of difficult semantic modeling caused by the large volume of industrial text lexicon and high information density. With this as the theoretical basis, this paper designs a text recognition method based on the fusion of industrial semantics and vision. The method extracts text features from both semantic modalities and image modalities respectively, and can effectively recognize industrial text in fuzzy images. Experiments show that ITScanner outperforms the existing methods in several industrial scene datasets.

Secondly, a multi-view feature fusion-based text recognition method (ITViewer) are proposed. In order to distinguish and exploit the redundancy and difference between multi-view images, self-attention mechanism is applied to accomplish the alignment and fusion of features between different views. Since the acquisition and labeling of multi-view images are time-consuming and labor-intensive, a

Multi-view text image generation method is presented in this paper based on the basic principles of camera imaging. Experiments show that ITViewer can focus on text features in multi-view images at the same time and has strong anti-interference compared with the dependence of existing methods on single-source images.

Thirdly, based on the actual industrial application scenario, this paper designs an industrial text recognition system integrating text detection, tracking and recognition with the industrial text recognition algorithms ITScanner and ITViewer designed in this paper as the core steps. In order to solve the problem of real-time operation of the system, the system adopts a framework of pre-recognition and fine recognition, which rapidly and roughly observes all text areas by pre-recognition, and then precisely identifies high-confidence areas by ITScanner and ITViewer to get the most accurate recognition results.

The above research shows that the text recognition method based on multi-source information fusion has strong robustness and anti-interference, and has significant advantages in industrial text recognition tasks, which is a good practice for the implementation of text recognition technology in industrial scenes.

Key words: text recognition, industrial scene, multi-source information fusion, text semantics, multi-view images

目 录

本文插图索引

本文表格索引

第一章 绪论

1.1 研究背景与意义

工业现场存在大量的文字，它们一般以数字、字母、号码、符号等形式被喷印或钢印在金属、塑料等各类材质的物料、产品和设备上，往往包含许多有价值的关键信息，如材料批次、产品型号、生产日期、生产车间编号等。工厂利用这些文字信息，能够精准快捷地控制生产流程和推进生产计划。目前，大部分工厂主要通过人工观察阅读这些文字，并手动录入到计算机系统中，需要消耗大量的人力成本和宝贵的生产时间。如果能通过计算机从图片、视频等多媒体感知信息中自动提取这些文本信息，就可以充分发挥机器视觉的优势，进而提升生产效率、降低生产成本。

为了解决这个问题，场景文字识别（Scene Text Recognition，STR）技术应运而生，近年来备受学术界与工业界的关注。借助深度神经网络的强大特征表达能力，现有的文字识别算法能够较准确地识别清晰图像中的文字。然而，由于复杂工业环境的多方面影响，场景文字识别技术在工业现场的落地仍然存在较大的困难。首先，工厂中常见的大功率设备如电焊、无线电发射、大电机、大继电器等产生的电磁辐射将较大程度地干扰视频监控系统中的信号传输。虽然数字滤波技术能够一定程度地抑制这些干扰，但滤波器的设计和参数的选择受环境影响较大，这导致在工业现场稳定传输高质量的图像仍有困难。其次，弱光环境下的曝光不足、高温高湿高粉尘环境导致镜头易受污染、元件流转和物流过程中的运动模糊和频繁遮挡等常见因素在物理层面阻碍了摄像机感光单元的曝光过程，使图像中相应区域的视觉信息模糊不清甚至完全丢失。图1-1展示了若干工业现场采集的文字图像示例。上述因素给场景文字识别在工业场景中的应用带来了严峻的挑

图 1-1 工业场景中文字图像
Figure 1-1 Text images in industrial scenes

战。一方面，由于环境噪声、光照变化等干扰因素的影响，文字识别方法对于字符区域的定位不够精准；另一方面，由于工业现场中收集的文字图像质量不稳定，现有方法提取的文字特征存在较大偏差。综合来看，已有的场景文字识别方法难以满足工业现场的实际需求。

通过广泛地文献检索和实验验证，本文现有方法的共同不足在于过度依赖于单一图像中的信息。在环境多变、干扰较强的工业现场中获取的单一图像具有较大的随机性和片面性，其中包含的像素信息容易被污染或遮挡，文字识别结果的可信度与完整性大大降低。然而，我们观测和描述一个文字目标的方式并不局限于单一信息来源。人们不仅能从不同时间、不同视角拍摄该文字目标获得多个图像，还能从语义规则、实际含义等非视觉方面来描述、识别和推测该文字目标。文字识别方法若能够考虑和利用多源信息之间的联系与区别，就能从不同信息源中提取出更加准确、鲁棒的特征表示，表现出更强的环境适应性和抗干扰性。因此，本课题旨在研究基于多源信息融合的文字识别方法，并将其应用于工业文字识别任务中，构建一种抗干扰能力较强的文字识别系统。

1.2 国内外研究现状

本节首先介绍了文字识别的主流范式，随后介绍了多源信息融合方法的相关研究进展，并介绍了文字识别领域中与多源信息融合相关的研究工作。

1.2.1 文字识别的主流范式

场景文字识别因其任务特性，兼具图像识别与序列识别两大特点。因此，主流文字识别方法大多同时采用适用于图像识别的卷积神经网络（Convolution Neural Networks，CNN）和适用于序列识别的循环神经网络（Recurrent Neural Networks，RNN）。总的来看，现有文字识别模型均包含如下四个阶段：

- **空间变换阶段**。现实场景中广泛存在的不规则文本对后续阶段的泛化性提出了较大的挑战。为了解决该问题，该阶段基于空间变换网络（Spatial Transformer Networks，STN）[1]将输入的图像变换为规则文本

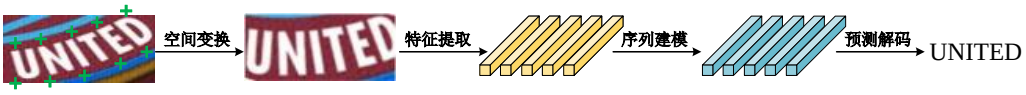


图 1-2 文字识别方法的主流范式[14]

Figure 1-2 The dominant paradigm of existing text recognition methods

图像。具体地，该模块首先寻找包围文字上下边界的一组基准点，随后基于薄膜样条插值法（Thin Plate Spline，TPS）完成文字图像的规则化。

- **特征提取阶段**。该阶段大多采用卷积神经网络如 VGG[2]、ResNet[3]、EfficientNet[4]等或视觉 Transformer（ViT）[5]将输入图像映射为一个特征序列 V，其中每一个元素 $v_i \in V$ 表示图像中相应位置中文字的特征。该特征序列侧重于与字符识别相关的属性，不受字体、颜色、大小和背景等不相关因素的影响。

- **序列建模阶段**。由于卷积神经网络具有局部性，前一阶段的特征序列 V 并没有考虑文字序列的全局上下文信息。因此，许多工作采用了双向长短期记忆网络（Bidirectional Long Short Term Memory，BiLSTM）来提取上下文关系，以获得更加准确的特征序列 H[6-8]。近期工作中，学者更加青睐于使用 Transformer[9]网络，通过其中的自注意力机制来提取全局特征，取得了不小的性能提升[10-12]。另外，仍有工作如 Rosetta[13]去除了序列建模阶段以达到降低时间空间复杂度，提高运算速度的目的。

- **预测解码阶段**。该阶段主要任务是将前述特征序列 H 解码为输出的文本序列，是文字识别方法中最核心最重要的阶段。

按照解码方式的不同，文字识别算法可以分为以下三类：基于关联性时序分类（Connectionist Temporal Classification，CTC）[15]的方法、基于字符分割的方法、基于注意力机制（Attention Mechanism）解码的方法。

(1) 基于 CTC 的文字识别方法

在文字识别问题中要获得字符级别的标注信息存在许多困难，需要大量的人力时间资源。因此，输入的图像数据大多仅有单词级别的标注，这

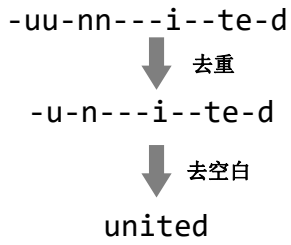


图 1-3 CTC 解码规则示例
Figure 1-3 A sample of CTC decoding

就需要文字识别模型去学习如何进行字符对齐。文字识别领域中的部分工作[6,13,16-17]采用了关联性时序分类（CTC）来处理未对齐序列的元素对齐问题。

为了完成序列对齐，CTC 首先定义了一个多对一映射 $\mathcal{B}$，通过去除原始序列 π 中的连续重复的元素与空白占位符元素的方式，将原始序列 π 映射成目标序列 l：$\mathcal{B}(\pi) = l$。图1-3展示了一个具体的 CTC 解码案例，其中‘-’代表空白占位符，其他字母代表有效元素。考虑到 $\mathcal{B}$ 为多对一映射，目标序列的预测概率为能映射到该序列的所有原始序列的预测概率之和。原始序列的概率则通过每个位置对应元素概率的累乘来计算，即[15]：

$$P(l|y) = \sum_{\pi:\mathcal{B}(\pi)=l} P(\pi|y) \tag{1.1}$$

$$P(\pi|y) = \prod_{t=1}^{T} P(\pi_t|y_t) \tag{1.2}$$

其中 T 是原始序列的长度，y 表示模型预测的概率序列，下标 t 表示对应位置。

为了文字识别中应用 CTC 方法，输入图像中的每一列像素视为一个图像帧。在经过空间变换、特征提取、序列建模三个阶段后，输入图像被降采样到高度为 1，宽度为 T，通道数为 C 的特征序列 $H \in \mathbb{R}^{1\times T\times C}$。按列从左往右看，每一个图像帧则对应于特征序列中的一个元素。模型在输出端采用 softmax 层预测每个元素对应的字符类别。根据 CTC 规则，预测的类别数除了所有有效字符类别数，还需加上一类空白占位符。在 CTC 解码规则的帮助下，整个模型无需字符级别的标注就能实现端到端的训练，极大地

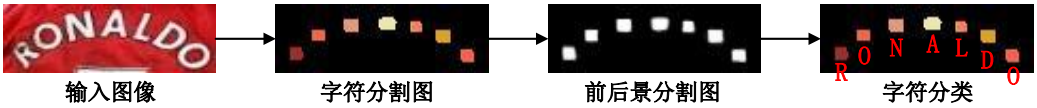


图 1-4 基于分割的文字识别方法[18]

Figure 1-4 Segmentation-based text recognition methods

降低了人工标注的工作量。

以卷积循环神经网络[6]（Convolution Recurrent Neural Network, CRNN）为代表的基于 CTC 的文字识别方法识别效率高，容易部署。然而该类算法的抗干扰能力不强，容易受到环境因素，如光照变化、图像模糊的影响，在许多情况仍然无法满足具体业务需求。

(2) 基于字符分割的文字识别方法

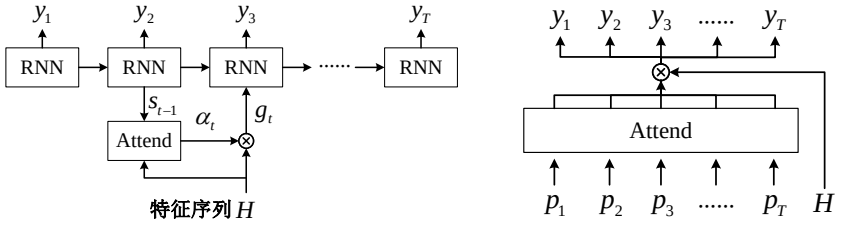
如图1-4所示，这类方法[18-21]将文字识别问题看作语义分割问题，指的是将图像中的每一个像素关联到一个字符类别标签上的过程。该类方法一般采用全卷积网络获取文本字符的分割图 $F \in \mathbb{R}^{h \times w \times c}$，其中 h,w 表示文字图像的高宽，c 值为字符类别数加上一个背景类别。采用的全卷积网络呈现 U 型，先通过卷积与池化等下采样操作提取图像特征，再通过反卷积反池化等上采样操作将特征还原为原始图像大小以获得字符分割图。在经过前后景分割、字符区域分类、序列生成等简单的后处理流程之后，该类方法就完成了从文字图像到文本序列的识别任务。

从原理上来看，该类方法从二维视角上完成特征提取以及文字分类，因此该方法能够应对任意形状的文字识别问题。然而，该类方法仍有两方面的缺点。一方面，为了训练该类模型，训练集的数据一般需要依靠字符级别的标注来生成符合语义分割要求的标签图。另一方面，该类方法所用的网络结构相对庞大，在面对中文这种大类别（上千类）的文字识别问题，所需生成的字符分割图 F 的通道数 c 较大，对设备的内存空间需求较高。

(3) 基于注意力机制的文字识别方法

在处理大批量信息和数据时，模型在许多情况下只需要关注小部分关键数据即可准确地完成任务。注意力机制即是一种估计数据重要程度的模块，它告诉模型哪一部分的数据需重点关注，哪一部分的数据可以忽略。注意力机制最早应用于自然语言处理领域中的机器翻译应用中，后续也被学者用于解决文字识别问题。如图1-5所示，按运算方式不同，文字识别方法

中采用的注意力机制可以分为串行[7,14,22-30]与并行[10-12,31-34]两种。



(a) 串行注意力机制解码 (b) 并行注意力机制解码

图 1-5 基于注意力机制的文字识别方法

串行注意力机制基于循环神经网络（如长短期记忆网络 LSTM 或门控循环单元 GRU），运算呈现一种自回归的特点。在解码过程中，每一个解码步的输入是来自于特征序列 $H = \{h_1, h_2, \ldots, h_T\}$ 和 RNN 上一步的隐藏层状态 s_{t-1}。首先，利用注意力机制计算该步字符的视觉特征表示 g_t；然后输入到 RNN 中，得到更准确的特征表示 y_t，并更新 RNN 隐藏层状态。依次循环，即[23]

$$e_{t,i} = f(s_{t-1}, h_i) \tag{1.3}$$

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{i'=1}^{n} exp(e_{t,i'})} \tag{1.4}$$

$$g_t = \sum_{i=1}^{n} \alpha_{t,i} h_i \tag{1.5}$$

$$(y_t, s_t) = rnn(s_{t-1}, g_t) \tag{1.6}$$

其中 $f(\cdot,\cdot)$ 函数称为对齐函数[1]，用于计算 s_{t-1} 在特征 h_i 上的响应值。随后，模型根据 y_t 完成字符分类，类别数除去有效字符种数，还需添加一类终止符。当输出一个终止符时，自回归解码过程停止，完成本次字符识别任务。

并行注意力机制的数学形式与前者类似，但其运算能够同时进行，各个解码步之间相互独立，互不影响。为了能够并行运算，该类模型舍弃了串行注意力机制中的 RNN 结构，并且使用字符阅读顺序的位置编码 $P =$

[1]不同方法中采用的对齐函数略有不同。

$\{p_1, p_2, \ldots, p_T\}$ 取代了串行注意力机制中的 RNN 隐藏层状态 s_{t-1}，即[10]：

$$e_{t,i} = f(p_t, h_i) \tag{1.7}$$

$$\alpha_{t,i} = \frac{exp(e_{t,i})}{\sum_{i'=1}^{n} exp(e_{t,i'})} \tag{1.8}$$

$$y_t = \sum_{i=1}^{n} \alpha_{t,i} h_i \tag{1.9}$$

串行与并行两种注意力机制方式各有优劣，在各自的适用范围中均取得了较好的效果。串行注意力机制由于 RNN 本身的特性，拥有一定的语义建模能力，在模糊文本识别上具有优势；然而由于其自回归的运算方式，其解码流程时间复杂度较高。并行注意力机制能够并行计算，运算效率高，但丢失了语义建模的能力，存在较大程度的注意力漂移问题。针对这一点，也有学者提出在模型下游增加一个显式的语义模型以弥补该方面的能力[10,12]，取得了可观的性能提升。总的来看，该类模型利用注意力机制对字符定位更加准确，能够在每一个解码步中关注到对应位置的特征，是目前精度最高的一类方法。

1.2.2 图像文字识别与多源信息融合

多源信息融合即基于多种 (同类或异类) 信息源，根据某个特定标准在空间或时间上进行组合，获得被测对象的一致性解释或者描述，并使得该信息系统具有更好的性能。为了充分利用多种信息源包含的特征，学者们提出了一系列的融合方法。总的来说，它们可以分为数据层融合、特征层融合和决策层融合[35]。即将原始数据的直接融合。数据层融合是将多个同类传感器采集的原始数据直接融合，也称早期融合。特征层融合首先提取数据源的特征信息，通过分析它们之间的关联性和差异性以获得更准确的特征表示，也称为中期融合。决策层融合基于特定的规则融合多个模型的输出结果，也称为晚期融合。在本文关注的工业文字识别任务中，由于工业现场采集到的文字图像质量良莠不齐，仅依靠单一来源的图像信息较难达到十分理想的识别效果。因此，本文从多源信息融合的角度出发，分别探究了结合文字语义的文字识别方法和结合多视角图像的文字识别方法。

1.2.3 结合文字语义的文字识别方法

文字本身包含丰富的语义信息。从生活经验出发，人们阅读模糊图像中的文字时往往会结合周边字符的情况，并基于构词法或者词组搭配等知识来推理预测模糊的字符。无独有偶，学者们在分析现有方法的失败案例时发现，许多错误都能利用文字中的语义信息加以避免。故近年来，结合文字语义的文字识别方法[10-12,36-45]越来越受到学者的关注。

由于待识别目标的基本单元是字符，文字识别任务主要关注字符级别的语义建模。令文字序列 $Y=\{y_1,y_2,\cdots,y_N\}$，则其语言模型在统计学上可表示为：

$$P(Y)=\prod_{i=1}^{N}P(y_i|y_1,y_2,\cdots,y_{i-1}) \tag{1.10}$$

在实际应用中，由于文字序列的长度 N 可能较大，这将导致模型参数呈指数级增长。为了解决该问题，研究者引入了马尔可夫假设，即当前字符仅与前 $n-1$ 个字符有关。故式(1.10)可简化为 N 元语言模型（N-gram model）：

$$P(Y)=\prod_{i=1}^{N}P(y_i|y_{i-n+1},\cdots,y_{i-1}) \tag{1.11}$$

基于 N 元语言模型，Nagai 等人[36]和 Wei 等人[37]分别提出了两种基于集束搜索（beam search）的晚期融合方法。该方法简单有效，能一定程度的提升识别准确率。然而，由于集束搜索是不可导的步骤，该类方法无法进行端到端的训练，故其语义建模能力仍然有限。

近期许多工作提出了一些结合文字语义的端到端可学习模型。Qiao 等人[38]提出的 SEED 模型采用全局语义信息而不是零状态来初始化串行解码器中的 RNN，迈出了显式语义建模重要的一步。具体地，SEED 模型通过两个线性层将特征序列 H 映射为全局语义向量 E，并使用一个预训练好的 FastText[39]语言模型来监督该语义向量。Yu 等人[10]提出的 SRN 模型包括一个基于并行注意力机制的视觉模块和一个基于 Transformer 的语言模块，并融合两者输出的特征，做出最终的预测。ABINet[12]提出了一种自主、双向、迭代式的文字识别模型。自主性使得该模型中的语义模块能够独立于整体，并且能从其他语料库中学习语义建模。同时，该模型提出一种类似完形填空的双向语义模型，在训练中抹去文本中的某一个字符，再让模型根据周

边字符去预测该字符，以获得更强的语义建模能力。该模型提出一种迭代式的预测方式来进一步减少预测错误。结合文字语义的方法近期取得了较大的成功，其在公开数据集上的准确率处在最高的水平。

受到公开基准数据集（如 ICDAR 2013、2015 等）的影响，现有工作针对日常场景的文字序列（如英语单词、词组等）提出了一些语义建模方法，大多通过一种类似 “背单词” 的方式学习其中的语义模式。不同于日常生活中使用的自然语言文本，工业文本往往不是某个单词或词语，而是使用一串由字母、数字、符号等组成的代码。考虑到这两者之间的语义模式区别，现有方法较难学习到准确的工业语义特征。一方面，工业文本的种类成百上千，每种工业文本又具有上万种可能情况，因此其词库大小要远高于自然语言文本[1]，对现有方法的单词记忆能力提出了较大挑战。另一方面，工业文本中蕴含的信息密度比日常文字要高，使用少量字符就能表达多方面的含义，这对现有方法的语义建模能力提出了较大挑战。本文旨在探究更符合工业文本的语义建模方法，获得更加准确的语义特征表示，以达到更准确的文字识别效果。

1.2.4 结合多视角图像的文字识别方法

在复杂的工业场景中，由于物体间的频繁遮挡、拍照设备的曝光不均、对焦不准以及传输过程的图像信息损耗等因素，工业现场采集的文字图像质量参次不齐，其中某些文字区域中的像素信息可能十分模糊甚至完全丢失。在上述客观因素的影响下，仅从单次拍摄的图像中就获得准确的识别结果较为困难。因此，许多研究者提出了许多结合多视角图像的视觉方法。

多视角图像具有多方面的优点。第一，多视角图像能够缓解图像采集过程中对焦不准、曝光不均等因素的干扰。第二，遮挡现象与拍摄视角息息相关，从多个视角拍摄同一目标能够大大降低遮挡范围，故多视角图像具有较强的抗遮挡特性。总之，相比于单视角图像，多视角图像能够携带更加完整的信息，在环境因素影响较大的情况下，即使某一视角的部分文字区域被遮挡，图像信息发生了缺失，不同视角之间也能够相互印证、相互补充，构建出较为完整的文字特征，有利于更准确的文字识别。

在现有工作中，多视角（Multi-view）图像常用于三维视觉任务中。在

[1]公开基准数据集的词库一般包含 9 万个常用英语单词。

3D 形状识别[46]、动作识别[47]、医学图像分割[48-49]等任务中，单一视角拍摄的图像无法观测到完整的目标物体，所以必须结合多视角的观测结果。一般来说，基于多视角图像的模型不仅需要考虑单视角图像的特征提取，还需考虑不同视角之间的特征对齐。Feng 等人[46]提出基于底层特征构建视角描述子，并根据不同视角下 CNN 提取得到的视觉描述子进行分组，将相类似的划分为一组，以完成对齐任务。Qiu 等人[47]以多视角成像过程为理论基础，首先使用 2D 关键点检测网络分别对各个视图做关键点检测，再基于视角之间的几何关系完成不同视角之间的特征融合。Chen 等人[49] 设计了一种基于多视角图像的形状表示网络来指导分割心脏解剖图像。

在文字识别领域中，传统的多视角方法[50-52]大多采用两阶段的识别流程：（1）它们会先将每个视角的图像分别输入到已有的文字识别模型中，获得多个对应的识别结果；（2）然后基于 KMP 字符串匹配等后处理方法将这些识别结果统合为一个整体。然而，这种方式既忽略了多视角图像之间的关联性，也无法进行端到端的训练。

为了解决上述问题，本文主要研究了一种端到端的多视角文字识别方法，通过在特征层面融合多视角图像中的信息以获取更准确的特征表示。

1.3 论文的研究内容与结构安排

本课题针对工业现场中环境多变、干扰较强的挑战，研究了基于多源融合的工业文字识别方法，并以此为核心设计了一套工业文字识别系统。图1-6展示了各章节内容和联系。

第一章主要介绍了工业场景中文字识别的重要性以及存在的挑战，简析了现有文字识别领域的发展现状，论证了多源信息融合对于工业文字识别的必要性，并简明扼要地说明了本文的研究工作和创新点。

第二章主要研究了结合语义的文字识别。该方法根据工业文本的分段式语义、段间独立性、段内关联性三个性质建立了工业语义的数学模型，为后文的模型实现提供了理论指导。其次，本文建立了结合工业语义的文字识别方法的数学模型，并利用卷积神经网络、自注意力机制模块、多层感知单元等网络结构实现了基于工业语义与视觉融合的文字识别模型。

第三章主要研究了结合多视角图像的文字识别。一方面，本文基于文字特点提出了一种基于多视角特征融合的文字识别方法。另一方面，为了

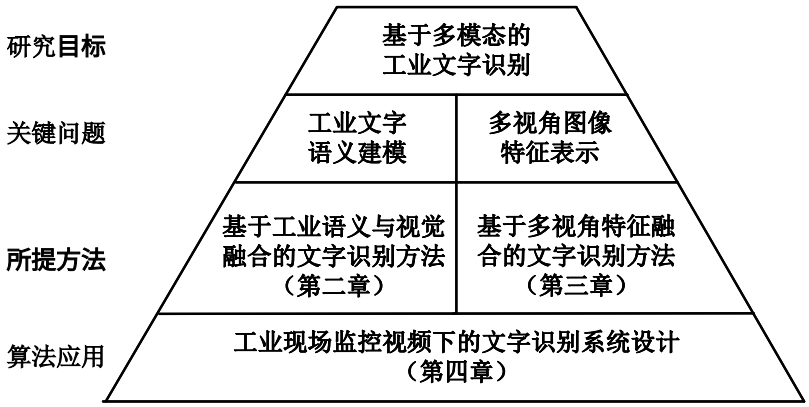


图 1-6 本文各章节内容和联系
Figure 1-6 Contents and contact of the thesis

使模型能够端到端识别，本文根据相机成像规律生成了两个多视角文字图像数据集 PlanarText 和 CylinderText。

第四章主要研究了基于多源信息融合的工业文字识别方法在现实场景的应用。本文立足于实际工业应用场景，以本文设计的工业文字识别算法 ITScanner 和 ITViewer 为核心算法，设计了一种集文字检测、跟踪、高精度识别的工业文字识别系统。

第五章简要总结了本文通过多源信息融合的思想设计的工业文字识别方法，并据此给出下一步的研究展望。


20003506

第二章　基于工业语义与视觉融合的文字识别方法

本章首先通过调查研究抽象表述了工业文字的语义模式，其次建立了结合工业语义的文字识别问题的数学模型，并通过深度神经网络结构构建了一种基于工业语义与视觉融合的文字识别模型（ITScanner）。通过实验，本文在若干个真实工业文本数据集上测试和验证了该方法各模块的作用，并讨论了本方法的创新性、有效性和局限性。

2.1　工业文字语义建模

在成熟的行业领域中，相关企业或组织（如国际标准化组织（International Organization for Standardization，ISO）将发布行业标准以统一工业文本的符号表示。根据其中相关规则，人们能够解释代码文本中的具体含义。以钢铁板坯编号为例，“21B40810603”中包含了生产年份、转炉号、生产日期与序列号等有价值的信息。

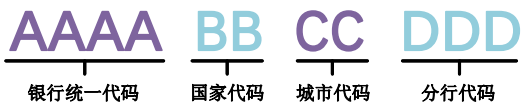
然而工业文本的种类成百上千，其遵循的行业标准与语义模式各不相同。图2-1展示了三个 ISO 发布的行业标准示例，表2-1展示了集装箱编号的国际标准。从以上图表可以看出，不同行业中使用的文字表示的语义不同，文字序列的模式也相差较大。为了准确描述其中的语义模式，本课题调研了大量现有的行业标准。本着找共性、谋全局的思想，本文首先分析了工业文字所遵循的行业标准的范式。总的来看，常见工业文字的行业标准主要规定了以下三个方面的内容：

- **文字组成**。其主要规定了该文字序列所包含的字符数量以及字符的取值范围。如图2-1(a) 展示的小型船只识别号标准规定了该识别号共使用 14 个字符来指代一种小型船只的型号，其中的字符可以为数字或字母。

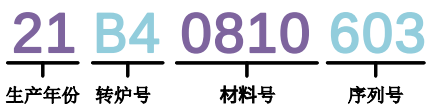
- **文字构造**。其主要规定了该文字序列的结构。如图2-1(b) 展示的银行身份码标准规定了一个银行身份码可分为 4 个文本片段。这些文本片段分别包含 4、2、2 和 3 个字符。

- **文字内容**。其主要规定了该文字序列中各部分所对应的含义。如图2-

(a) 小型船只识别号



(b) 银行身份码



(c) 钢铁板坯编号

图 2-1　ISO 发布的行业标准示例

表 2-1　集装箱编号标准
Table 2-1　ISO 6346-1995

片段序号	含义	字符范围	示例
1	箱主代码	三个大写字母	HTT, TBJ 等
2	设备识别码	U 或 Z 或 J	U, Z, J
3	箱号	六个阿拉伯数字	888208
4	校验码	一个阿拉伯数字	0-9

1(c) 展示的钢铁板坯编号标准规定了该文字序列个片段分别表示生产年份、转炉号、材料号和序列号。

基于上述范式，本文归纳总结了工业文本通用共有的语义模式。令 $Y=\{y_1,y_2,\ldots,y_N\}$ 为一个工业文本实例，则其满足以下特性：

- **分段式语义**。从整体来看，一串工业文本更类似于一句话而非一个单词，其可分为若干个文本片段。根据行业标准的规定，每个文本片段包含固定的字符个数。令文本片段个数为 m，每一段包含的字符为 $n_i, i\in\{1,2,\ldots,m\}$，则有

$$Y=\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_m\},\quad \mathbf{y}_i=\{y_i^1,y_i^2,\ldots,y_i^{n_i}\}$$

- **段间独立性**。文本片段之间的语义信息相互独立，互不影响，即：

$$P(Y)=\prod_{i=1}^{m}P(\mathbf{y}_i)$$

- **段内关联性**。同一文本片段中的字符需组成一个整体共同表达一个语义。因此，段内字符之间存在上下文的语义关联，即：

$$P(\mathbf{y}_i)=\prod_{j=1}^{n_i}P(y_i^j|y_i^1,\ldots,y_i^{j-1})$$

经过上述分析，工业文本语义模型可总结为：

$$Y=\{y_1,y_2,\ldots,y_N\}=\{\mathbf{y}_1,\mathbf{y}_2,\ldots,\mathbf{y}_m\} \tag{2.1}$$

$$P(Y)=\prod_{i=1}^{m}\prod_{j=1}^{n_i}P(y_i^j|y_i^1,\ldots,y_i^{j-1}) \tag{2.2}$$

其中 N 为文本的总字符个数，m 为该文本可分为的片段数量，n_i 表示属于第 i 个文本片段的字符个数。

从数学形式来看，式(2.2)将语义模型分为了若干个子模型，每个子模型独立地表示了对应文本片段内的语义信息。从文本的词库大小来看，该模型将一个大词库转化为了若干个小词库，降低了语言模型的记忆难度。从文本中的信息密度来看，该模型分别对工业文本中的多个语义单独建模，降低了语义建模能力的要求。本章将以该工业语义建模为基础，设计结合工业文字语义的文字识别方法。

2.2　结合工业文字语义的文字识别

结合语义的文字识别过程可分为两阶段：1）提取和分析输入图像的文字外观特征，得到视觉预测结果 Z；2）基于学习到的语义模式，修正和改良视觉预测结果，输出最终的预测序列 Y。令 $X\in\mathbb{R}^{c\times h\times w}$ 为高度为 h，宽

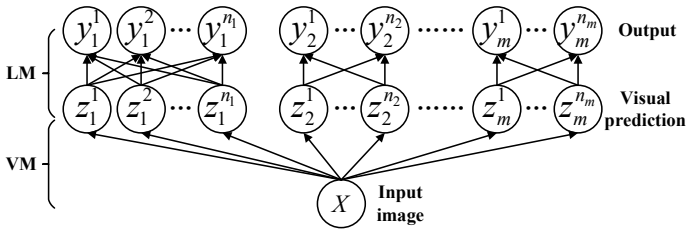


图 2-2　数学模型
Figure 2-2　The mathematical model

度为 w，通道数为 c 的输入图像，则该过程的数学模型可表示为：

$$\begin{aligned}\hat{Y} &= \underset{Y}{\operatorname{argmax}} P(Y|X) \\ &= \underset{Y}{\operatorname{argmax}} \sum_{Z} P(Z|X)P(Y|Z).\end{aligned} \tag{2.3}$$

其中 $P(Z|X)$ 为视觉模型，$P(Y|Z)$ 为语言模型。由于 Z 与 Y 均用于表示一个工业文字序列，故 Z 也具有式(2.1)的性质，即：

$$Z = \{\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_m\}, \quad \mathbf{z}_i = \{z_i^1, z_i^2, \ldots, z_i^{n_i}\}$$

视觉模型将文字视为无上下文关系的符号，仅从外观层面区分不同的文字类别：

$$P(Z|X) = \prod_{\forall i,j} P(z_i^j|X) \tag{2.4}$$

而语言模型主要考虑工业文字本身的上下文关系，结合式(2.2)，其可表示为：

$$P(Y|Z) = \prod_{\forall i} \prod_{\forall j} P(y_i^j|z_i^1, z_i^2, \ldots, z_i^{n_i}) \tag{2.5}$$

结合式(2.3)-(2.5)，该数学模型可总结为如图2-2所示的结构。总的来看，该模型是在结合语义的文字识别方法的基础上，引入了第 2.1 节提出的工业文字语义模型，并在数学上将文字识别过程分为了视觉模型和语言模型两部分，为后续的模型实现和训练提供了理论基础。

2.3　基于工业语义与视觉融合的文字识别模型

本节以前述数学模型为理论基础，利用卷积神经网络和注意力机制等网络结构，设计了一种基于工业语义与视觉融合的文字识别模型。依据式(2.3)-(2.5)，基于工业语义与视觉融合的文字识别模型的结构可分为视觉模型部分与语言模型部分。

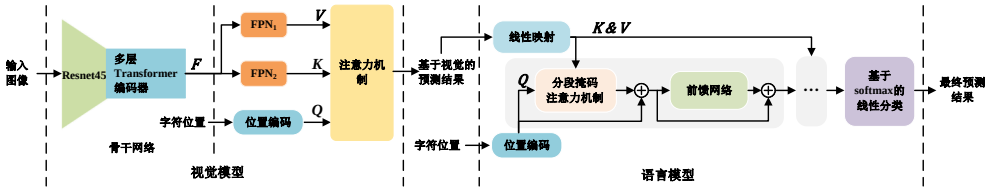


图 2-3　基于工业语义与视觉融合的文字识别模型的网络结构

Figure 2-3　The architecture of the text recognition model based on industrial-semantics-visual fusion

2.3.1　视觉模型

视觉模型的整体结构遵从章节1.2.1所述的主流范式，可分为特征提取、序列建模和预测解码三个阶段。特征提取阶段采用文字识别方法常用的ResNet45[3]网络结构。相比于 ResNet50，ResNet45 减少了三次下采样操作，提高了输出特征图的分辨率，更符合文字识别任务的需求。序列建模阶段采用多层 Transformer 编码器，利用其自注意力机制的全局感受野以分析不同区域之间的特征关联性，用于增强前一阶段输出的视觉特征。

$$F = \mathrm{Trm}(\mathrm{ResNet45}(X)) \in \mathbb{R}^{e \times \frac{hw}{16}} \tag{2.6}$$

预测解码阶段采用并行注意力机制的解码方式来预测字符类别。由于仅使用字符位置的编码作为注意力机制的查询向量，不同解码步骤之间互不干涉、相互独立，符合式(2.4)的模型定义。另外，该阶段采用两个特征金字塔网络[53]（Feature Pyramid Network，FPN）将特征图 F 映射到键向量空间和值向量空间，能够融合多尺度条件下的特征表示。这两个金字塔网络结构相同（如图2-4所示），但参数相互独立。

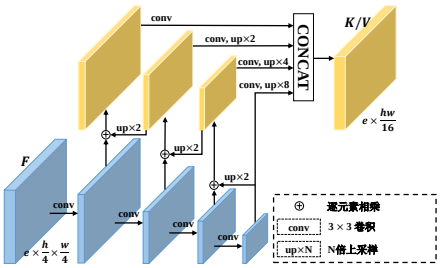


图 2-4　特征金字塔网络结构

Figure 2-4　Structure of the feature pyramid networks

$$\begin{aligned} K &= \mathrm{FPN}_1(F) = (k_1, k_2, \ldots) \in \mathbb{R}^{e \times \frac{hw}{16}} \\ V &= \mathrm{FPN}_2(F) = (v_1, v_2, \ldots) \in \mathbb{R}^{e \times \frac{hw}{16}} \end{aligned} \tag{2.7}$$

对于第 t 个字符，模型首先基于注意力机制从值向量中提取相关的特征 $\hat{v}_t$，然后输入到参数为 $W_v \in \mathbb{R}^{c \times e}, b \in \mathbb{R}^c$ 的线性分类层得到该字符的类别。

$$\tilde{w}_{t,i} = \frac{q_t^T k_i}{\sqrt{e}},\ w_{t,i} = \frac{exp(\tilde{w}_{t,i})}{\sum_{\forall i} exp(\tilde{w}_{t,i})},\ \hat{v}_t = \sum_{\forall i} w_{t,i} v_i \tag{2.8}$$

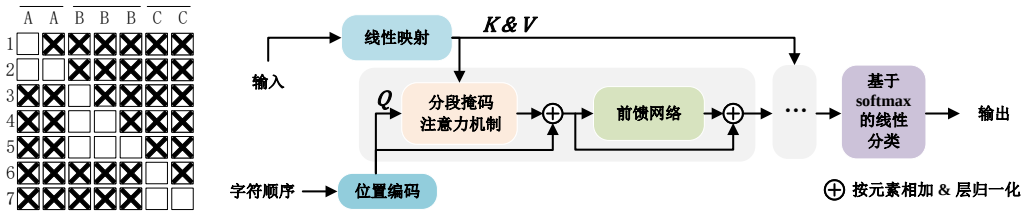
$$P(z_t|X) = P(z_t|\hat{v}_t) = f_{\mathrm{smx}}(W_v \hat{v}_t + b) \tag{2.9}$$

2.3.2　语言模型

从数学形式来看，式(2.2)可看作 m 个子语言模型的串联，每一个子语言模型 $\prod_{j=1}^{n_i} P(y_j|y_1, \ldots, y_{j-1})$ 用于表达一个文本片段内的上下文关系。为了表示上述语义模型，并联 m 个 LSTM 网络或并联 m 个 Transformer 解码器均为可行的实现方案。然而，多级并联的复杂结构将大幅度提升模型的参数量与运算复杂度，也不利于模型的训练学习过程。

为了解决该问题，本文提出了一种基于分段掩码的注意力机制模块来实现式(2.2)的语义模型，避免了多个子模型并联的复杂结构。其数学表示为：

$$f_{ga}(Q, K, V) = V f_{\mathrm{smx}}(\frac{Q^T K}{\sqrt{e}} + \mathbf{M}) \tag{2.10}$$



(a) 分段掩码　　(b) 网络结构

图 2-5　基于分段掩码注意力机制的语言模型

Figure 2-5　Group-wise mask attention-based language model

$$\mathbf{M}=\begin{pmatrix} M_{n_1\times n_1} & & -\infty \\ & \ddots & \\ -\infty & & M_{n_m\times n_m} \end{pmatrix} \tag{2.11}$$

$$M=\begin{pmatrix} 0 & \dots & -\infty \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix} \tag{2.12}$$

其中 $\mathbf{M}$ 为对角块为 $M_{n_i\times n_i}$ 其余为 $-\infty$ 的方阵，$M_{n_i\times n_i}$ 为下三角全为 0 其余全为 $-\infty$ 的方阵，n_i 为第 i 个文本片段包含的字符个数。从数学形式来看，通过在一般的注意力机制的基础上引入分段掩码 $\mathbf{M}$，该模块能够屏蔽冗余的段间关联性，保留有效的段内关联性。图2-5(a)展示了文本序列“AABBBCC”所对应的分段掩码。

基于分段掩码注意力机制，本文提出了结构如图2-5(b)的语言模型。在输入端，模型分别采用线性映射与位置编码的方式将输入序列 $Z\in\mathbb{R}^{c\times N}$ 与字符顺序 $t\in\{1,\dots,N\}$ 映射到同一个线性空间 $\mathbb{R}^e$ 中，然后输入到分段掩码注意力机制模块中：

$$K=V=W_cP(Z|X) \tag{2.13}$$

$$\begin{cases} Q_{t,2i} & =\sin(t/10000^{2i/e}) \\ Q_{t,2i+1} & =\cos(t/10000^{2i/e}) \end{cases},i\in\mathbb{Z},0\le i\le\frac{e}{2}. \tag{2.14}$$

其中 N 为文字序列的最大长度，c 为字符类别数，e 为注意力机制中的向量

维度，$W_c \in \mathbb{R}^{e\times c}$ 为可学习的权重矩阵。

为了提高模型的语义特征表示能力，本文采用了一种深层网络结构，每一层均包含分段掩码注意力机制模块、前馈网络、残差连接与层归一化结构。前馈网络是一种包含两层全连接层的网络结构，用于衔接前后两层的注意力机制。残差连接结构在网络中引入了一条恒等映射的分支，通过将一部分的前一层的信息无差的传递到下一层，能够很好地缓解深层网络退化的问题。层归一化对神经网络中隐藏层的输入进行归一化，从而使得网络更容易训练，是自然语言处理领域中常用的归一化方法。令模型总层数为 L，则其运算过程如下：

$$\begin{aligned} F_0 &= Q \\ F_l' &= f_{ga}(\mathrm{LN}(F_{l-1}),K,V)+F_{l-1}, l=1,\ldots,L \\ F_l &= \mathrm{FFN}(\mathrm{LN}(F_l'))+F_l', \qquad l=1,\ldots,L \end{aligned} \tag{2.15}$$

$$\mathrm{FFN}(x) = W_2(\max(0,W_1x+b_1))+b_2 \tag{2.16}$$

经过多层结构的语义编码后，我们获得了一个文字语义特征表示 F_L。通过一个基于 softmax 的线性分类层后，模型输出文字序列的类别。

$$\hat{Y} = f_{\mathrm{smx}}(WF_L+b) \tag{2.17}$$

2.3.3　损失函数

参考现有工作的通用做法，我们将文字识别问题看作一种序列多分类问题，并采用交叉熵损失函数计算预测值 $\hat{Y}$ 与标签 Y 之间的距离。

$$\mathcal{L}(\hat{Y},Y) = -\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{n_m}\ln(P(\hat{y}_i^j = y_i^j|X)). \tag{2.18}$$

2.4　实验及分析

本节将介绍实验所用数据集、模型训练方式、各模块作用分析和与之前相关工作的性能对比。

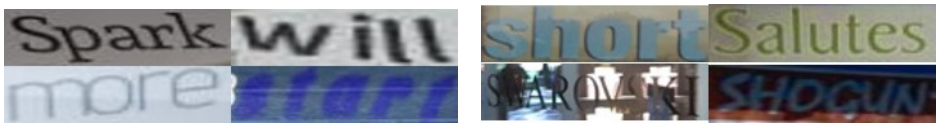
2.4.1　数据集简介

(1)　日常场景文字数据集

虽然本文主要关注工业场景的文字识别任务，但本方法的视觉模型具有较高的独立性，能够单独训练，并完成许多场景的文字识别任务。因此在实验中我们也使用了一些日常场景的文字识别数据集作为训练集和测试集来验证其性能。

训练集 MJSynth（MJ）[54]与 SynthText（ST）[55]是两个常用的合成数据集。MJSynth 数据集包含 890 万个文本图像样本，覆盖了 9 万多个常用的英文单词。其数据合成步骤包括文字渲染、边界与阴影渲染、着色、透视变形、真实图像混合和加入噪声。SynthText 合成数据集原本被提出于针对文字检测，每一张图片包含丰富的背景和多个文字实例。文字的内容均来自于新闻稿合集库 Newsgroup[1]。为了使渲染的文字融入于背景中，该方法通过语义分割和深度估计的算法获得放置文字的区域、方向和角度。在文字识别任务中，学者裁剪出图像中存在的文本区域，获得了 800 万个文本图像样本。

测试集使用了三个常用的真实场景的英文文本数据集。ICDAR 2013[56]（IC13）数据集包括 1015 张图片，本文筛去了包含非字母数字和文字长度小于 3 的样本，使用剩余的 857 张图像用于测试。IIIT5K[57]数据集包含 5000 张图像，本文仅使用了其测试集的 3000 张图像。SVT[58]包括 647 张街景文字图像，存在一些低分辨率和模糊的图像。



(a) 合成数据集　　(b) 真实场景数据集

图 2-6　日常场景文字数据集示例
Figure 2-6　Samples of daily scene text datasets

(2)　工业场景文字数据集

训练集采用合成数据集 SynthIT。该数据集的合成方式与与 SynthText 数据集较为类似，但数据源均来自于工业场景。首先，该数据集所用背景

[1]http://qwone.com/ jason/20Newsgroups/

图像均为工业场景照片，如工厂车间、门禁入口等。其次，文字的内容均来自于本文收集的工业文本语料库，包含了大量真实的工业文字内容。另外，为了仿真工业文字图像的高噪声、低分辨率等特性，文字的渲染过程中加入了较高程度的高斯噪声、运动模糊、颜色变换等步骤。

测试集使用了两个真实场景的工业文本数据集，均来源于工厂的监控视频。集装箱数据集（CIN）包括 367 个样本。由于表面的凹凸不平，大部分文字都发生了一定程度的形变。受到环境影响和长时间的磨损，部分字符十分模糊，难以辨认。钢铁板坯编号数据集 (SB) 包含 2153 个图像。受到网络传输限制，该数据集中的图像的分辨率较低，故大部分图像都十分模糊。此外，由于生产环境因素，钢铁板坯上的喷印字符容易出现氧化脱离、字符粘连等情况，容易造成许多误识别情况。图2-8展示了这两个数据集的示例。

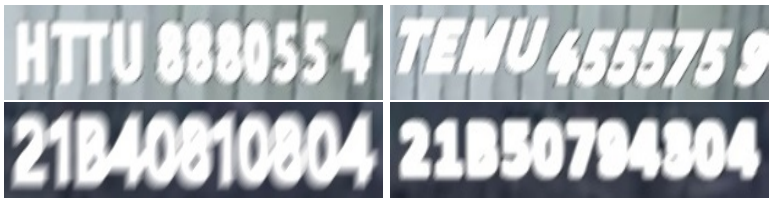


图 2-7　SynthIT 数据集示例
Figure 2-7　Samples of SynthIT datasets

(a) 集装箱编号　　(b) 钢铁板坯编号

图 2-8　真实场景工业文本数据集示例
Figure 2-8　Samples of real world industrial scene text datasets

2.4.2　训练方式

(1)　视觉模型预训练

视觉模型的精确性和泛化性很大程度上依赖于训练集的多样性和丰富性。为了增强模型的特征表示能力，本文使用了 SynthText 和 MJSynth 来预

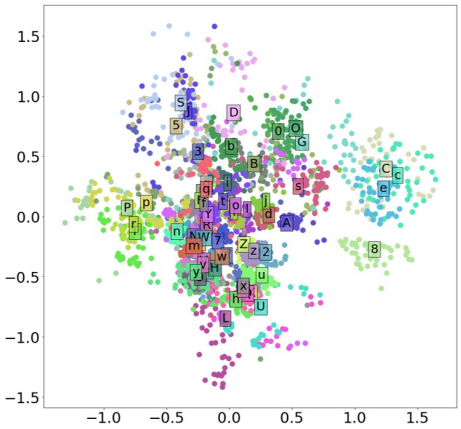


图 2-9　字符表征向量的可视化
Figure 2-9　Visualization of character's representation

训练 ITScanner 的视觉模型。虽然这两个数据集中的文本并不符合工业文本的语义模式，但这些图像包含颜色各异、字体多变、形状多样的文字，具有丰富的外观多样性，有助于文字识别模型学习到更加准确泛化的视觉特征表示。预训练过程中，本文采用了一种在线数据增广方法，按一定概率随机施加运动模糊、高斯模糊、几何变换、对比度调节等方式。此外，本文采用 ADAM 优化器来优化模型参数，每一个训练批次大小为 192。模型总共训练了 8 个迭代周期，前 6 个周期中学习率为 10^{-4}，后两个周期下降为 10^{-5}。

(2)　语言模型预训练

本文提出的语言模型可以看作一种文本纠错模型。对于一个优秀的文本纠错模型，若输入一个正确的序列，则输出的序列不变；若输入错误的序列，则输出纠正后的正确序列。故语言模型预训练的过程，本质上是训练该模型的文本纠错能力。

SynthIT 数据集包含的 10 万个文本标签被用来作为预训练任务的训练数据。为了扩充训练数据，本文提出了一种基于分段式语义的文本数据增广方法。具体做法是：1）以文本片段为单位来收集语料库；2）从语料库中采样，并按顺序拼接为新的工业文本。若某一种可分为三段的工业文本的两个样本“AA BBB CC”和“DD EEE FF”。通过上述的数据增广方法，我们总共可以获得 8 个符合要求的文本数据，如“AA EEE FF”等。

为了构造文本纠错任务的训练样本对，需对正确的文本施加随机的字符扰动。为了使这种扰动更符合真实情况，我们参考了 Chen 等人[59]的工

作，根据外观形状构造了字符的表征向量。图2-9展示了字符表征向量的可视化结果。外观相似度越高，其在图中距离较近，如“c”“e”和“z”“2”。因此，我们使用表征向量之间的欧氏距离作为字符之间的相似度量，并在取负之后使用 softmax 函数归一为概率分布，作为字符扰动的依据，即：

$$w_{ij} = ||v_i - v_j||^2, \quad p_{ij} = \frac{e^{-w_{ij}}}{\sum_{j \neq i} e^{-w_{ij}}} \tag{2.19}$$

另外，考虑到文本检测框的影响，错误识别问题更容易发生在序列的首尾。在选择扰动的字符时，应给予首尾字符更大的概率比重。

总的来说，语言模型的预训练任务步骤如下：1）文本数据采样；2）随机选择是否扰动字符；3）若选择扰动，随机选择一个字符；4）按式(2.19)的概率分布随机选择一个字符替换第 3 步中选择的字符。整个预训练任务中，我们使用 SGD 优化器来优化模型参数，共优化 100 个迭代周期，前 20 个周期学习率设置为 10^{-4}，之后降为 10^{-5}。

(3)　端到端训练

端到端训练是指深度学习模型直接学习从原始数据到期望输出的映射的过程。在经过前述的两个预训练步骤之后，我们使用合成的工业文本数据集 SynthIT 训练整个文本识别模型。在此步骤中，我们同样采用(1)节中所述的数据增广方法，并使用 ADAM 优化器来优化模型参数，学习率设置为 10^{-5}，共训练 8 个迭代周期。

2.4.3　实验分析

本节通过实验评估了视觉模型和语言模型的具体性能，并讨论了这两个模块中一些超参数的影响。另外，本节还通过实验探讨了预训练方式对于模型性能的提升效果。

(1)　视觉模型性能评估与讨论

视觉模型是 ITScanner 的重要组成部分。为了验证视觉模型的特征提取能力，本文分别使用日常场景的测试集 IC13、IIIT5K、SVT 和工业场景的测试集 CIN、SB 测试其识别精度。本文还讨论了不同方式的序列建模对于性能的影响。

如表2-2[1]所示，视觉模型在日常场景的数据集上的准确率均在 80% 以上，而在工业场景的数据集上的准确率均低于 50%。这说明了识别工业场景文字的难度远高于日常场景，仅凭视觉信息难以准确获得正确的识别结果。另外，随着 Transformer 编码器层数的增加，视觉模型的识别性能越高，运算速度越低。由于过深的模型的优化难度大大增加，在层数高于 3 之后，模型的精度提升幅度趋于平缓。因此，在今后的实验中，ITScanner 均采用 3 层 Transformer 编码器的配置。

表 2-2　不同设置下的视觉模型性能测试
Table 2-2　Evaluation of the vision model with different configurations

方法	数据集准确率 (%)					运算速度
	IC13	IIIT5K	SVT	CIN	SB	（ms）
Trm-0	89.15	83.04	81.72	40.14	22.10	10.3
Trm-1	91.74	90.05	87.28	42.58	24.93	13.7
Trm-2	93.58	94.32	89.47	44.17	26.52	17.2
Trm-3	**94.92**	**94.56**	90.41	**44.44**	**26.59**	19.1
Trm-4	94.73	94.49	**90.56**	43.81	26.37	22.0

(2)　语言模型性能评估与讨论

为了证明本文提出的语言模型的有效性，本节在工业场景数据集 CIN 和 SB 上进行了两方面的实验。第一，保持模型的视觉部分不变，更改语言部分，观察准确率的变化情况。第二，将本文语言模型引入到现有方法中，观察前后准确率的变化情况。

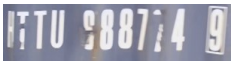
在探究实验一中，我们不仅实验了无语言模型的情况，还使用其他的语言模块替换掉本文提出的语言模型。GSRM 提出于 SRN 方法中，是一种全局语义推理模块，适用于日常文字的语义模式。表2-3的实验结果表明，在图像模糊或存在遮挡的情况下，工业文字中所蕴含的语义信息能较大程度地提升文字识别模型的准确率。并且，相比于 SRN 方法中为日常文字提出的语义模块 GSRM，本文提出的基于分段掩码注意力机制的语言模型更加符合工业文本的语义模式，故准确率比 GSRM 高 18.03%。表2-4列举了一些

[1]Trm-n 表示模型在序列建模阶段使用了 n 层的 Transformer 编码器

表 2-3　语言模型探究实验一
Table 2-3　Ablation study 1 of the proposed language model

语言模型	在 CIN 数据集上的准确率（%）
无	44.44
GSRM[10]	56.63
本文语言模型	**74.66**

表 2-4　探究实验一中的代表案例
Table 2-4　Typical cases in ablation experiment 1

输入图像	视觉预测结果	GSRM	本文语言模型
	XTTU8886526	HTTU8886526	HTTU8886526
	ITTU9887I49	INTU8887749	HTTU8887749
	21850100q04	21B50000904	21B50800904
	21BS079950b	21BS0799503	21B50799503

代表案例。从中我们可以看出，虽然 GSRM 模块具有一定的纠错能力，但当视觉预测结果中存在多处错误时，该模型不仅无法完成纠错任务，还会由于错误信息的传递反而修改识别正确的文本。本文提出的语言模型，由于其分段掩码的作用，视觉预测结果中的错误仅会被本片段中的其他字符注意到，因此避免了错误的全局传递，具有很大的优越性。

在探究实验二中，我们将本文提出的语言模型接入到 CRNN、Rosetta 和 CA-FCN 三个不考虑文字语义的文字识别方法的后端，利用其学习到的语义特征提升这三种方法的语义识别效果。表2-5的实验结果显示，在引入工业语义后，这三种方法在 CIN 数据集上分别提升了 18.79%、19.98%、15.67%，在 SB 数据集上分别提升了 15.92%、14.76%、13.67%。这不仅表明了本文提出的语言模型具有较强的语义特征能力，还具有一定程度的独立性，能够单独作为一个模块协助其他文字识别算法以获得更准确的识别结果。

(3)　预训练对模型性能的影响

表 2-5　语言模型探究实验二
Table 2-5　Ablation study 2 of the proposed language model

数据集	现有方法	原准确率（%）	引入本文语言模型后的准确率（%）	差值
CIN	CRNN[6]	41.92	60.71	**+18.79**
	Rosetta[13]	43.97	63.95	**+19.98**
	CA-FCN[18]	42.22	57.89	**+15.67**
SB	CRNN[6]	26.24	42.16	**+15.92**
	Rosetta[13]	28.04	42.80	**+14.76**
	CA-FCN[18]	25.86	39.53	**+13.67**

章节2.4.2中提出，在端到端训练之前先使用额外的数据集预训练视觉模型和语言模型，能够获得准确率更高的模型。本文通过实验验证了该方法的有效性。表2-6的实验结果指出：1）预训练视觉模型能使模型在 CIN 和 SB 数据集上的准确率分别提升 3.62% 和 5.87%，这说明在 SynthText 和 MJSynth 两个数据集上预训练能够帮助模型学习到更加泛化的特征表示；2）预训练语言模型能使模型在 CIN 和 SB 数据集上的准确率分别提升 1.33% 和 1.26%，意味着本文提出的基于分段增广和字符扰动的预训练方式存在一定的有效性；3）两者均完成预训练任务时，模型的准确率分别在 CIN 和 SB 数据集上提升 5.75% 和 7.45%。

表 2-6　预训练对模型性能的影响
Table 2-6　The effect of pre-training

预训练		CIN	SB
视觉模型	语言模型		
		68.91	42.94
✓		72.53	48.81
	✓	70.24	44.20
✓	✓	**74.66**	**50.39**

2.4.4　与现有方法的对比

表 2-7　与现有方法的对比
Table 2-7　Comparison with state-of-the-arts

方法			CIN	SB
不考虑文字语义	CRNN[6]	TPAMI 2016	41.92	26.24
	Rosetta[13]	KDD 2018	43.97	28.04
	CA-FCN[18]	AAAI 2019	42.22	25.86
考虑文字语义	SAR[25]	AAAI 2019	50.89	35.02
	SRN[10]	CVPR 2020	56.47	38.62
	VisionLAN[11]	ICCV 2021	49.01	34.36
	ABINet[12]	CVPR 2021	57.21	38.63
	ITScanner（本文方法）		**74.66**	**50.39**

为了验证本文提出的 ITScanner 在识别工业文本上的创新型和有效性，本文测试了现有方法在工业场景数据集 CIN 和 SB 上的识别性能。为了公平比较，这些方法的特征提取部分均使用 ResNet45 网络，并都使用 SynthText、MJSynth 与 SynthIT 三个合成数据集进行训练，其中 SynthText 和 MJSynth 为日常场景文字数据集，SynthIT 为工业场景文字数据集。表2-7列出了所有方法的准确率，其中 CRNN、Rosetta 与 CA-FCN 三个方法在识别文字时不考虑文字语义，其余 4 个方法和 ITScanner 考虑文字语义。从表中结果可以看出：

- 考虑文字语义的方法的准确率普遍高于不考虑语义的方法，这说明在复杂的工业环境下，工业文字中的语义信息能较大程度地协助文字识别任务，帮助获得更准确的识别结果。
- 相比于现有方法，ITScanner 在 CIN 和 SB 两个数据集上分别提升了 17.45% 和 11.76%，这说明本文提出的基于分段掩码注意力机制的语言模型更加适用于工业文本的语义模式，能够更加准确地提取工业文本中的上下文关系。

2.4.5　算法适用性讨论

表 2-8　识别失败样例展示
Table 2-8　Failure cases

输入图像	真实标签	视觉预测结果	最终预测结果
	1B50800902	1B50100*02	1B50100*02
	HTTU8889356	HTIU81893*6	HTTU88893*6
	HTTU8880512	HTTU8880510	HTTU8880510
	21B40810004	21B40810804	21B40810804

本文提出的 ITScanner 虽然在大部分情况下表现良好，但仍然不适用于一些极端情况。第一，由于不同种类工业文本的语义模式相差较大，分段情况各不相同，故本文提出的语言模型在参数固定的情况下仅能处理单种工业文本。若需识别多种文本，则需在不同数据集上分别训练，以获得多套参数，并在实际使用中根据输入的工业文本种类来选择对应的参数和分段掩码。第二，若视觉预测结果中存在多字或少字的错误识别情况，这将使得字符位置和分段掩码无法对齐，并大大影响语言模型的作用。第三，工业文本中存在一些不包含语义信息的部分，如集装箱编号中的箱号和钢铁板坯编号中的序列号。当错误识别发生在这些文本片段中时，本文提出的语言模型也无法纠正。表2-8[1]展示了一些识别失败的样例。

2.5　本章小结

本章提出了一种基于工业语义与视觉融合的文字识别方法。该方法根据工业文本的分段式语义、段间独立性、段内关联性三个性质建立了工业语义的数学模型，解决了工业文本词库量大，信息密度高导致的语义建模难的问题。其次，本文建立了结合工业语义的文字识别方法的数学模型，为之后的模型实现提供了理论指导。另外，本文利用卷积神经网络、自注意力机制模块、多层感知单元等网络结构实现了基于工业语义与视觉融合的文

[1]标红的字符为错误字符；* 为占位符，代表漏识别的字符。

字识别模型。最后，本文进行了大量的实验来验证本文方法的有效性，并讨论了 ITScanner 各模块的作用。实验结果说明，在恶劣条件下的工业文本识别任务中，本文提出的 ITScanner 具有一定的优越性。

第三章　基于多视角特征融合的文字识别方法

在复杂的工业场景中，由于物体间的频繁遮挡、拍照设备的曝光不均、对焦不准以及传输过程的图像信息损耗等因素，工业现场采集的文字图像质量参次不齐，其中某些文字区域中的像素信息可能十分模糊甚至完全丢失。在上述客观因素的影响下，仅从单次拍摄的图像中就获得准确的识别结果较为困难。

为了更加准确鲁棒的识别结果，本文探究了一种基于多视角图像的文字识别方法。多视角图像是从不同视角对同一目标多次拍摄的图像集合，一般为不同位置的多个摄像头拍摄获得。在实际应用中，若某一摄像头与拍摄目标存在持续的相对运动，则该摄像头在不同时间拍摄的目标图像亦可称为多视角图像。本章基于多视角图像中的文字特点，提出了一种基于多视角特征融合的端到端文字识别方法（ITViewer）。为了使模型能够进行端到端训练，本文以相机成像原理为理论基础，提出了一种多视角文字图像的生成方法，并生成了两个数据集 PlanarText 和 CylinderText。实验表明，ITViewer 在多个真实工业现场的数据集 SB-M 和 CIN-M 表现出较强的准确性和鲁棒性。

3.1　多视角文字识别方法设计

3.1.1　问题定义

与一般文字识别任务不同，多视角文字识别方法须同时提取多个输入源中的文字特征表示，通过区分和利用它们之间的关联性和差异性完成多视角之间的特征融合，最终完成文本序列的分类。给定一个对应于文字序列 Y 的多视角图像集合 $\{X_1, X_2, \cdots, X_m\}$，本文研究的文字识别方法可表示为：

$$\hat{Y} = \underset{Y}{\operatorname{argmax}} P(Y | f(X_1, X_2, \cdots, X_m)) \tag{3.1}$$

其中 $f(*)$ 为特征融合的方式。基于上述特点，本文提出的文字识别方法包括基于自注意力机制的多视角特征融合和基于跨视角注意力机制的预测解码，其整体框架如图3-1所示。

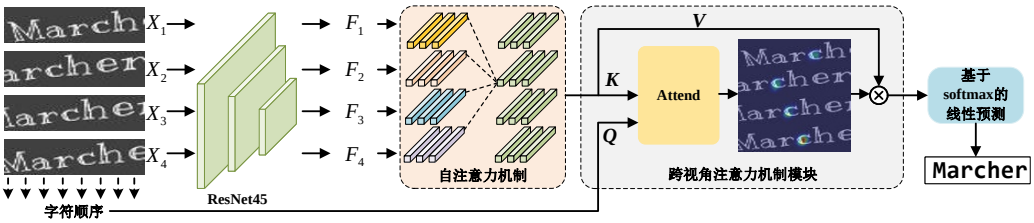


图 3-1　基于多视角特征融合的文字识别方法
Figure 3-1　Text recognition method based on multiple view fusion

3.1.2　多视角特征融合

作为计算机视觉任务，文字识别方法常用卷积神经网络来提取输入图像中的特征。然而，卷积神经网络仅能作用于单一图像的局部区域，无法从不同图像中寻找关联性和差异性。为了解决该问题，本文采用自注意力机制来实现多视角特征融合任务，具体结构如图3-2所示。

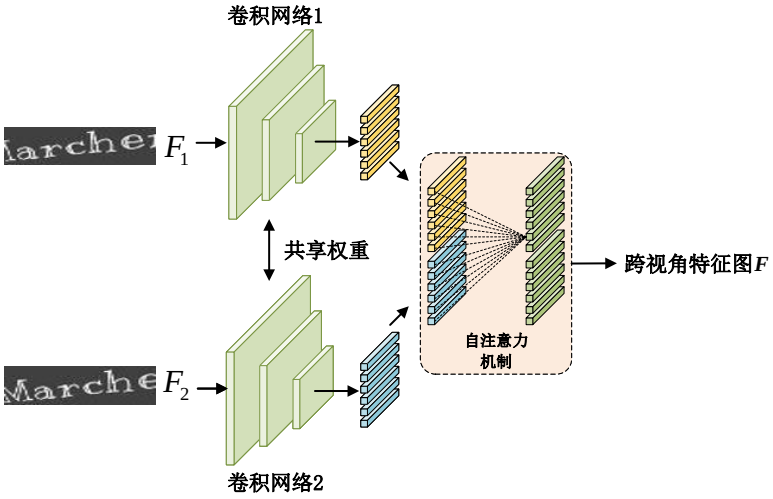


图 3-2　基于自注意力机制的多视角特征融合
Figure 3-2　Multiple view feature attention based on self-attention

首先，我们基于卷积神经网络 $\varphi(*)$ 构造了一种 Siamese 网络结构，以将输入图像映射至同一个特征空间中。Siamese 网络也称孪生网络，具有多个相同架构、参数和权重的相似子网络，常用于目标跟踪、人脸识别等任务中。令 $X \in \mathbb{R}^{c\times h\times w}$ 表示为高度为 h，宽度为 w，通道数为 c 的输入图像，则

有

$$\varphi(X) = F \in \mathbb{R}^{e \times n} \tag{3.2}$$

其中特征图 F 表示 n 个感受野对应的特征向量，维度为 e。

理论上说，外观相似的感受野区域在经过卷积神经网络之后应获得相似的特征表示。根据该特性，自注意力机制通过计算图像之间的相似度来关联不同图像间的感受野。令带有尺度变化的内积函数

$$\lambda(f_1, f_2) = \frac{f_1^T f_2}{\sqrt{e}} \tag{3.3}$$

为相似度函数，则图像 X_1 上某一感受野的特征 f 与图像 X_2 的特征图 $F = (f_1, f_2, \cdots, f_n)$ 的关联性可表示为：

$$\tilde{w}_i = \frac{(W_q f^T)(W_k f_i)}{\sqrt{e}}, \quad w_i = \frac{exp(\tilde{w}_i)}{\sum_{\forall i} exp(\tilde{w}_i)} \tag{3.4}$$

其中 W_q 和 W_k 为两个可学习的权重矩阵。若 f 所对应的感受野为图像 X_1 上的一字符区域，w_i 越接近于 1，说明 X_2 的第 i 个感受野中包含了同一字符。如图3-3的示例中，不同视角图像中关于同一字符“A”的感受野在经过卷积神经网络后的特征向量相似度较高，而关于字符“A”和“P”的感受野对应的特征向量相似度则较低。

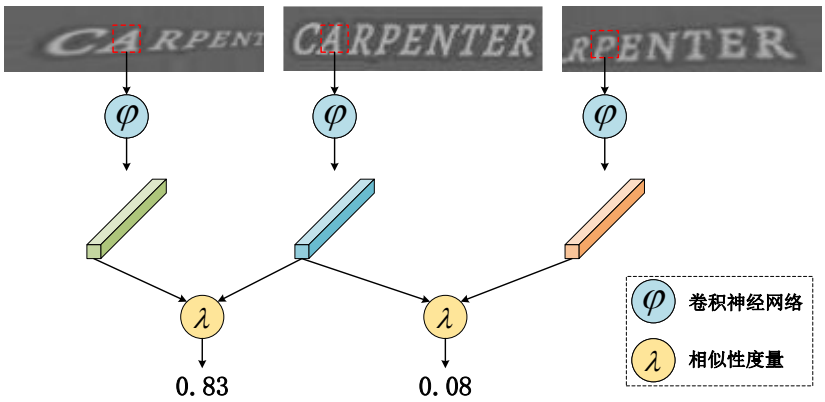


图 3-3　多视角图像特征间的关联性

Figure 3-3　The correlation between multiple view images' features

通过感受野特征间的相似度，自注意力机制建立了多视角图像间的关

联性，然而，不同的拍摄视角客观导致了图像间仍存在许多差异，其中包括了不少相互印证、相互补充的信息。为了融合不同视角间的特征，自注意力机制基于图像间的关联性，通过加权平均的方式从其他视角中寻找有价值的特征信息，即：

$$f = f + \sum_{i=1}^{n} w_i f_i \tag{3.5}$$

对于 N 个输入视角图像，为了运算效率，我们用矩阵形式来表示上述运算，得到了了跨视角特征图 $\mathbf{F}'$：

$$F_i = \varphi(X_i), i = 1, 2, \cdots, m \tag{3.6}$$

$$\mathbf{F} = (F_1, F_2, \cdots, F_m) \tag{3.7}$$

$$\mathbf{F}' = \mathbf{F}\text{softmax}(\frac{(W_q\mathbf{F})^T(W_k\mathbf{F})}{\sqrt{e}}) + \mathbf{F} \tag{3.8}$$

3.1.3　基于跨视角注意力机制的预测解码

在预测解码阶段，注意力机制是常用的解码方法，常用的有一维注意力机制和二维注意力机制。在多视角文字识别任务中，文字的形状和排布通常呈倾斜和弯曲状，一维注意力机制较难准确地定位字符，识别效果不够理想。二维注意力机制能从二维视角关注到字符的具体位置，但无法从多个视角中关注同一字符，存在严重的注意力漂移问题。为了在预测解码阶段能够关注到多个视角中的文字特征，本文提出了一种跨视角注意力机制。

跨视角注意力机制是并行注意力机制的一种变体，区别在于前者能够处理更高的特征维度。我们首先基于跨视角特征图 $\mathbf{F} \in \mathbb{R}^{e \times n \times m}$（$m$ 为视角数，n 为单视角图像中的感受野数）构建键值对：

$$k_{ijk} = v_{ijk} = \mathbf{F}_{ijk} \tag{3.9}$$

我们使用字符顺序的位置编码作为注意力机制的查询向量，即：$q_t = f_o(t), t \in [0, 1, \cdots, T]$ 其中 T 表示文字序列的最大长度。也就是说，当我们使用 q_t 去

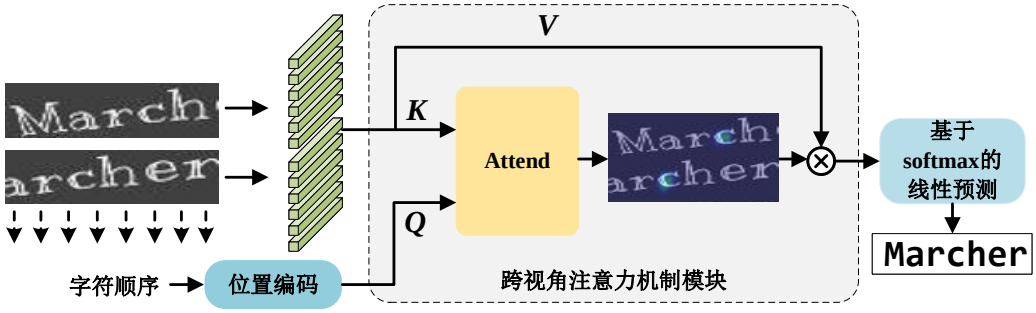


图 3-4　基于跨视角注意力机制的预测解码

Figure 3-4　The prediction based on cross-view attention mechanism

查询前述的键值对时，我们想要获得图像中的第 t 个字符对应的特征。在该步骤中，我们使用了求和式的注意力图生成方法：

$$\begin{cases} e_{t,ijk} = W_e^T tanh(W_q q_t + W_k k_{ijk}) \\ \alpha_{t,ijk} = \dfrac{\exp(e_{t,ijk})}{\sum_{\forall i,j,k} \exp(e_{t,ijk})} \end{cases} \tag{3.10}$$

图展示了跨视角注意力图的可视化结果，直观地表现了该方法的关注区域。从中我们看出，该方法能够从多个视角的图像特征中关注到准确的文本区域。

随后，我们采用加权平均的形式获取到第 t 个字符所对应的字符特征，并通过基于 softmax 的分类层来预测该字符的类别：

$$g_{t,ijk} = \sum_{\forall i,j,k} \alpha_{t,ijk} v_{ijk} \tag{3.11}$$

$$P(y_t | X_1, X_2, \cdots, X_m) = \text{softmax}(W_o g_{t,ijk}) \tag{3.12}$$

3.2　多视角文字图像数据集生成

众所周知，图像数据集是计算机视觉任务的基础，良好的公开数据集是驱动课题发展的动力源。然而，相比于一般文字图像数据，收集和标注多视角文字图像数据需花费成倍的人力物力。目前的公开数据集如 MJSynth (MJ)[54]、SynthText (ST)[55]、ICDAR 2013[56]等数据集只提供单视角的文字图像。在现有工作中，我们暂未检索到公开的多视角文字图像数据集，无

法满足第 3.1 节所述方法的实验验证需求，这给本文的实验验证带来了困难。为了解决这个问题，本节首先分析了相机成像的原理，以此为理论基础提出了两个多视角图像的生成方法，并生成了两个数据集 PlanarText 和 CylinerText。

3.2.1　相机成像原理与多视角图像

数码相机图像拍摄的过程是一个光学成像的过程，可以分为四个步骤：刚体变换、透视投影、畸变校正和数字化图像。一般来说，一个无镜头畸变的相机的成像数学模型可表示为：

$$z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = K \underbrace{\begin{bmatrix} R & t \\ 0_3^T & 1 \end{bmatrix}}_{T} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{3.13}$$

其中 $(x_w, y_w, z_w)^T$ 为世界坐标系下一点的物理坐标，$(u, v)^T$ 为该点对应的在像素坐标系下的像素坐标。$K \in \mathbb{R}^{3\times4}$ 为内参矩阵，只与相机本身固有的参数有关，例如焦距 f、像元尺寸等。$T \in \mathbb{R}^{4\times4}$ 为外参矩阵，取决于相机坐标系和世界坐标系的相对位置，$R \in \mathbb{R}^{3\times3}$ 表示旋转矩阵，$t \in \mathbb{R}^{3\times1}$ 表示平移矢量。

从数学上说，使用数码相机从多个视角拍摄同一个物体，实际上是通过不同的外参矩阵 T 和内参矩阵 K 将世界坐标系下的同一个点 $(x_w, y_w, z_w)^T$ 变换到不同图像平面上。

若文字所在表面为平面，则根据式(3.13)可有如下推论[60]：从不同位置拍摄同一平面物体所得图像上的像素点 $(u_1, v_1)^T$ 和 $(u_2, v_2)^T$ 存在平面单应性（Planar Homography），并可以用投影变换表示，即：

$$\begin{bmatrix} u_1 \\ v_1 \\ 1 \end{bmatrix} = H \begin{bmatrix} u_2 \\ v_2 \\ 1 \end{bmatrix}, \quad H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix}, \tag{3.14}$$

其中 $H \in \mathbb{R}^{3\times3}$ 为一种投影变换矩阵，亦称为单应性矩阵。这表明，理论上

我们可以基于单应性矩阵变换将某一视角拍摄得到的平面文字图像变换到其它视角。

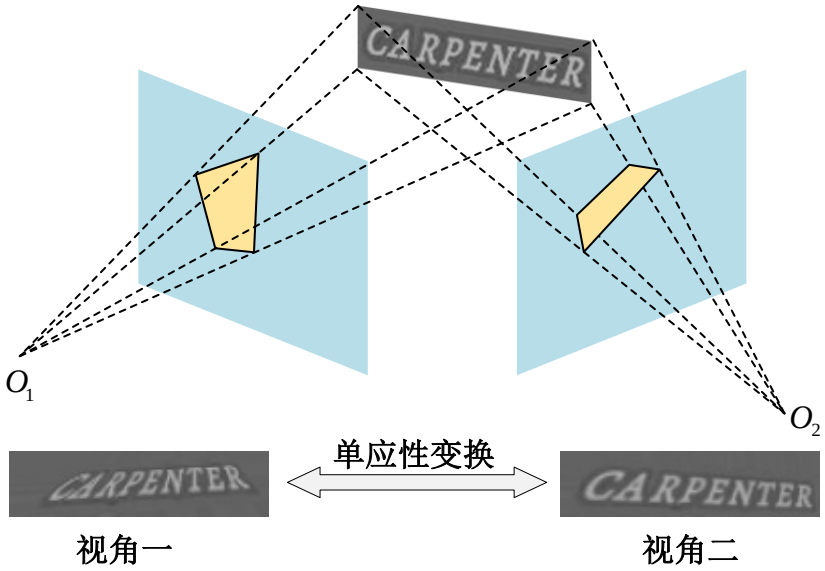


图 3-5　相机成像的平面单应性
Figure 3-5　The camera's planar homography

若文字所在表面为曲面，则上述单应性不成立。要想获得多视角的图像，只能通过具体的内外参矩阵 K 和 T 来计算 $(x_w, y_w, z_w)^T$ 所对应的像素点 $(u, v)^T$。

根据上述分析指出的平面文字与曲面文字的不同特性，本文针对平面文字，提出了基于随机单应性变换的平面文字多视角图像生成方法；针对曲面文字，提出了基于虚拟成像的曲面文字多视角图像构造方法。

3.2.2　平面文字多视角图像生成

根据式(3.14)，我们需要两方面的要素来生成多视角图像：大量的平面文字图像和符合实际情况的单应性矩阵。本节以 SynthText、MJSynth 和 SynthIT 三个数据集为数据源，提出了一种基于随机单应性变换的平面文字多视角图像生成方法，并生成了 PlanarText 数据集。

SynthText、MJSynth 和 SynthIT 均为合成文字识别数据集，共包含 1800 万个文字图像。虽然它们仅提供单个视角的数据，但这两个数据集共包括

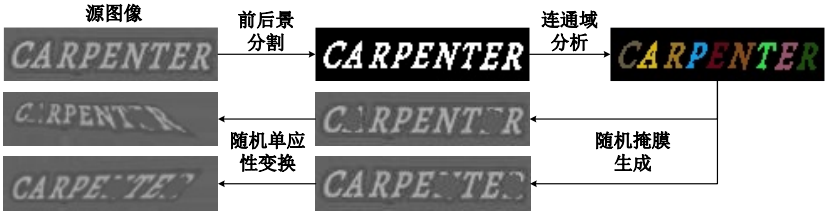
1800 万个文字图像，并且其中文字颜色各异、字体多变、形状多样，故仍然具有较高的数据价值。从生成方式来看，这两者生成的文字图像均可视为平面文字图像，故满足式(3.14)所描述的单应性。

为了批量获得多视角图像，我们须随机地生成符合实际情况的单应性矩阵。若通过直接随机化生成 H 矩阵中的 8 个参数的方式，我们无法控制该变换的畸变程度。因此，本文通过随机化图像顶点偏移的方式来控制单应性变换的畸变程度，并根据(3.14)来求解对应的单应性矩阵。例如，对于图像的左上顶点，有

$$\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} = H \begin{bmatrix} ap_1 \\ bp_2 \\ 1 \end{bmatrix}, \quad p_1, p_2 \sim Be(\alpha, \beta) \tag{3.15}$$

其中 a,b 为与图像尺寸有关的比例系数，p_1, p_2 为遵从贝塔分布的随机数。注意到 H 共有 8 个未知量，而一对点仅能确定两个方程。故本文为图像的 4 个顶点均按该方法生成了对应点，通过方程求解获得了随机的单应性矩阵。

(a) 滑动窗口采样法



(b) 随机掩膜采样法

图 3-6　平面文字多视角图像生成
Figure 3-6　The synthesis of multiple view planar text images

为了进一步体现多视角图像之间的互补性，在进行图像视角变换之前，本文实现了两种图像随机采样方式，使生成的每一个图像仅包含源图像中的部分信息。第一种方法称为滑动窗口采样法，即通过一个固定大小的滑

(a)　　(b)

图 3-7　PlanarText 数据集示例
Figure 3-7　Samples of PlanarText datasets

动窗口随机框选图像中某个区域的采样方法。该采样方式模拟了相机近距离拍摄一较长文本实例的过程，或是相机拍摄一个显示屏上流动文字的过程。第二种方法为随机掩膜采样法，即通过随机遮挡源图像中的某一个或某几个字符的采样方法。该采样方式模拟了相机拍摄受遮挡文本实例的过程。相比于滑动窗口采样法，该方法的实现方式相对复杂，具体做法是：1）基于 KMeans 算法完成前后景分割；2）基于连通域分析完成字符区域分割；3）随机选择若干字符，生成对应位置的掩膜。

结合前述的两种随机采样方式和随机单应性变换，基于随机单应性变换的生成方法能够生成信息互补的多视角图像数据，具体流程如图3-6所示，生成的 PlanarText 数据集的示例如图3-7所示。

3.2.3　曲面文字多视角图像生成

根据3.2.1节的分析，在拍摄曲面物体时图像之间的单应性并不成立。若采用3.2.2所述的平面文字多视角生成方法，则无法反映相机从多个视角拍摄曲面文字的实际情况。为获得较真实的曲面文字图像，本文提出了基于虚拟成像的生成方法：1）通过将文字渲染至三维模型的曲面上以获得文字对象准确的世界坐标；2）基于式(3.13)建立的相机模型，变换相机位置和拍摄角度，批量生成多视角图像。

(1)　曲面模型建立

为建立符合真实情况的曲面模型，我们首先调研了工业场景中常见的曲面文字。如图3-8所示，集装箱表面和冷轧钢胚表面均可视为一种柱面。柱面是动直线沿着一条定曲线平行移动所形成的曲面，其中动直线称为柱面的直母线，定曲线称为柱面的准线。为简化运算，本文采用母线平行于 y 轴的柱面来建立曲面模型，其准线方程可表示为：

$$F(x,z)=0 \tag{3.16}$$

通过改变柱面的准线方程，我们可以获得集装箱表面对应的锯齿状柱面或冷轧钢坯表面对应的圆柱面。

(a) 集装箱侧面编号　　(b) 冷轧钢坯表面序列号

图 3-8　工业场景中的曲面文字
Figure 3-8　Text on curve surface in industrial scenes

曲面模型确定之后，还需生成对应的文字贴图。本文的做法是：1）随机选取一张图像作为背景图；2）生成文字图像，并将其放置到背景图中的某些区域；3）将该背景图作为柱面的展开图。

(2)　虚拟成像

假定相机焦距为 f，分辨率为 $w\times h$，则本文所用相机的内参矩阵为：

$$K=\begin{bmatrix} 0 & f & \frac{h}{2} & 0 \\ -f & 0 & \frac{w}{2} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \tag{3.17}$$

为批量生成多视角图像，还需随机选择相机的位置，确定外参矩阵 T。外参矩阵中共 6 个参数，3 个对应于相机位置，另外 3 个对应相机的拍摄角

度。为了使相机位置在一定范围内，保证相机能够拍摄到清晰的文字图像，本文分两步来确定相机外参。

第一，在固定相机拍摄角度的条件下，先确定相机的位置 $(x_e,y_e,z_e)^T$。具体做法是：1）随机选择曲面上的一个文本序列，并获得其边界的 4 个顶点在世界坐标系上的坐标 $(x_i,y_i,z_i)^T$；2）在相机的相平面上随机选择一个矩形区域，并获得其 4 个顶点在图像坐标系上的坐标 $(u_i,v_i)^T$；3）将 $(x_i,y_i,z_i)^T$ 对应于 $(u_i,v_i)^T$，则有

$$(z_i+z_e)\begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} 0 & f & \frac{h}{2} & 0 \\ -f & 0 & \frac{w}{2} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & x_e \\ 0 & 1 & 0 & y_e \\ 0 & 0 & 1 & z_e \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} \tag{3.18}$$

整理后，可得一个关于 x_e,y_e,z_e 的方程：

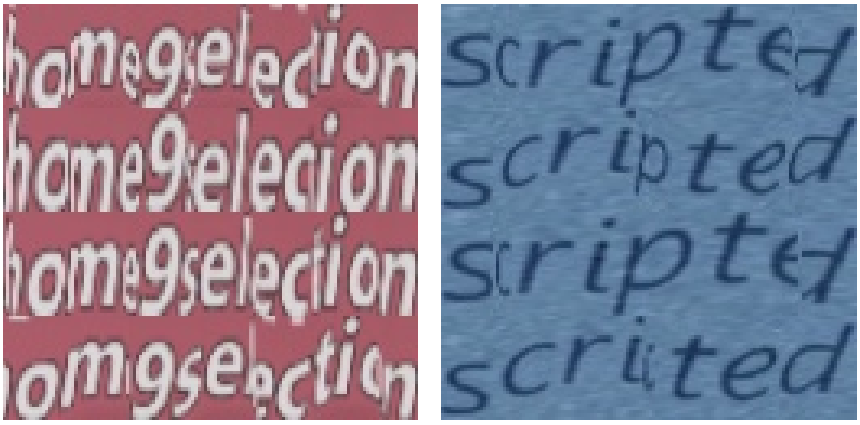
$$\begin{bmatrix} 0 & f & \frac{h}{2}-u \\ -f & 0 & \frac{w}{2}-v \end{bmatrix} \begin{bmatrix} x_e \\ y_e \\ z_e \end{bmatrix} = \begin{bmatrix} z_iu_i - fy_i - \frac{h}{2}z_i \\ z_iv_i + fx_i - \frac{w}{2}z_i \end{bmatrix} \tag{3.19}$$

可以看出，一对点能确定 2 个方程。通过我们随机选择的 4 对点，则能通过线性拟合的方式求解出相机的位置 $(x_e,y_e,z_e)^T$。

第二，将摄像头随机旋转一定的角度。具体做法是利用随机数生成一定范围内的欧拉角 α,β,ϕ，则可确定外参 T 中的旋转矩阵：

$$R = \begin{bmatrix} \cos\theta\cos\phi & \sin\psi\sin\theta\cos\phi - \cos\psi\sin\phi & \cos\psi\sin\theta\cos\phi + \sin\psi\sin\phi \\ \cos\theta\sin\phi & \sin\psi\sin\theta\sin\phi + \cos\psi\cos\phi & \cos\psi\sin\theta\sin\phi - \sin\psi\cos\phi \\ -\sin\theta & \sin\psi\cos\theta & \cos\psi\cos\theta \end{bmatrix} \tag{3.20}$$

在确定了内外参矩阵后，则可通过映射关系，即可计算曲面文字在成像平面上的投影，获得多视角的曲面文字图像。通过上述方法，本文生成了数据集 CylinderText，如图3-9所示。



(a)　　(b)

图 3-9　CylinderText 数据集示例
Figure 3-9　Samples of CylinderText datasets

3.3　实验及分析

本节将介绍数据集及算法实现细节等实验设置、算法各部分作用分析、与现有工作的比较以及局限性分析。

3.3.1　实验设置

(1)　数据集介绍

本次实验使用生成的 PlanarText 和 CylinderText 数据集作为训练集。PlanarText 通过投影变换的方式来生成受遮挡的平面文字的多视角图像，共包含 16 万个数据样本。CylinderText 通过三维建模和虚拟成像的方式来生成曲面文字的多视角图像，共生成 16 万个数据样本。以上两个数据集中的每个样本均包含 4 个视角图像。

为测试算法性能，本文收集和标注了两个真实场景下的多视角数据集 SB-M 和 CIN-M。本文利用智能手机近距离以随机视角拍摄钢铁热轧板坯编号，收集了 137 个样本，每个样本包含 2 至 5 个视角的图像，组成了 SB-M 数据集。CIN-M 数据集来源于工厂监控视频中运动的集装箱编号，共包含 189 个数据样本，每个样本包含 3 至 8 个视角图像。由于拍摄距离较近且集装箱表面凹凸不平，不同视角图像中的字符存在较大的外观差异性。图3-10展示了这两个数据集的几个示例。

(a) 多视角钢铁热轧板坯编号　　(b) 多视角集装箱编号

图 3-10　真实场景多视角数据集示例
Figure 3-10　Samples of real world multiple view text datasets

(2)　算法实现细节

本次实验中采用的模型可接受 4 个视角的图像输入，每个视角的图像均被缩放为宽 128 高 32 的尺寸。测试集 SB-M 和 CIN-M 中的某些样本不足 4 个视角，故本文使用随机白噪声图像补齐缺失的视角图像。

模型的主干网络采用以 ResNet45 网络为子结构的 Siamese 网络结构。ResNet45 输出的单视角特征图的大小为 512×128，即共有 128 个感受野，每个感受野对应的特征维度为 512。相比于 ResNet50，ResNet45 减少了三次下采样操作，提高了输出特征图的分辨率，更符合文字识别任务的需求。自注意力机制模块是一种类似于 Transformer 的多层结构，在实验中层数设置为 2。在预测解码阶段，字符顺序的位置编码 $f_o(t)$ 的形式与式(2.14)一致。在输出端，模型输出的文字序列长度最多为 26，字符类别共 37 类，包括 26 个字母、10 个数字以及一个结束标记符 <eos>。

在训练阶段，模型在 PlanarText 和 CylinderText 数据集上共训练 12 个周期，批次大小设置为 32。本文使用 Adam 算法来优化模型的参数，学习率初始设置为 10^{-4}，在 6 个周期后降为 10^{-5}。另外，本文在训练时采用了随机高斯模糊、运动模糊、颜色偏移等在线数据增强方法以提高模型的泛化性。

3.3.2　实验分析

本小节在 SB-M 和 CIN-M 两个真实场景数据集上，对多视角文字图像生成方式、基于自注意力机制的多视角特征融合方法和基于跨视角注意力机制的预测解码方法进行了消融实验，以验证本章所述方法的有效性。

(1)　多视角文字图像生成方式

为了科学准确地验证3.2节所述多视角文字图像生成方法的有效性，本文多次改变了模型所用训练集，并观察模型在测试集上的准确率。在本次实验中，本文额外构造了一个伪多视角数据集 SynthIT-M 作为空白对照。SynthIT-M 的图像数据源于 SynthIT，每个样本包含 4 张相同的图像。

从表3-1可以看到，当使用 PlanarText 数据集训练时，模型的准确率在两个测试集上分别提升了 18.98% 和 14.29%。当使用 CylinderText 数据集训练时，模型的准确率分别提升了 16.05% 和 11.11%。从提升幅度来看，PlanarText 中的平面多视角数据更符合 SB-M 的场景，而 CylinderText 的曲面多视角数据更符合 CIN-M 的场景。当同时使用 PlanarText 和 CylinderText 数据集时，模型的准确率达到了 50.36% 和 61.36%。

表 3-1　不同训练集下的识别精度测试
Table 3-1　The recognition accuracy with different training set

训练集			识别准确率（%）	
SynthIT-M	PlanarText	CylinderText	SB-M	CIN-M
✓			27.01	43.91
✓	✓		45.99	58.20
✓		✓	43.06	55.02
✓	✓	✓	**50.36**	**61.37**
	✓	✓	**50.36**	60.31

(2)　多视角特征融合

为了验证本文提出的基于自注意力机制的多视角融合方法的有效性，本文实验了若干种常用的融合方法，包括相加、连接和门控单元。它们的具

体运算方式如下：

$$\text{相加：}\ f=\sum_{i=1}^{m} f_i \tag{3.21}$$

$$\text{连接：}\ f=W_a[f1,f2,\cdots,f_m] \tag{3.22}$$

$$\text{门控单元：}\ \begin{cases} g=\text{softmax}(W_g[f1,f2,\cdots,f_m]) \\ f=\sum_{i=1}^{m} g_i \odot f_i \end{cases} \tag{3.23}$$

其中 $W_a \in \mathbb{R}^{e\times em}$ 和 $W_g \in \mathbb{R}^{m\times em}$ 均为可学习的权重矩阵，$\odot$ 表示两个向量的逐元素相乘。从表3-2可以看到，在 SB-M 数据集上自注意力机制相比于其他方式分别提升了 9.49%、6.56% 和 7.29%，在 CIN-M 数据集上则分别提升了 4.75%、5.81% 和 3.70%。这说明了基于自注意力机制的融合方式整体优于其他方式，并且能同时适用于多视角的平面文字和曲面文字，具有较强的泛用性。另外，由于 SB-M 的拍摄视角相差较大，不同视角中字符位置存在较大的偏移。上述三种融合方式均是对应位置特征之间的融合，故融合效果不理想。本文提出的自注意力机制首先根据特征间的相似性完成了多视角之间的特征对齐，有效地解决了不同视角位置偏移的困难，故 ITViewer 在 SB-M 数据集上的提升幅度要高于 CIN-M。

表 3-2　不同融合方式之间的比较
Table 3-2　Comparison between different fusion methods

融合方式	准确率（%）	
	SB-M	CIN-M
相加	40.87	55.56
连接	43.80	54.50
门控单元	43.07	56.61
自注意力机制（本文方法）	**50.36**	**60.31**

(3)　跨视角注意力机制

在预测解码阶段，本文采用了一种跨视角注意力机制。本小节将其与常用的一维注意力机制和二维注意力机制进行了对比。

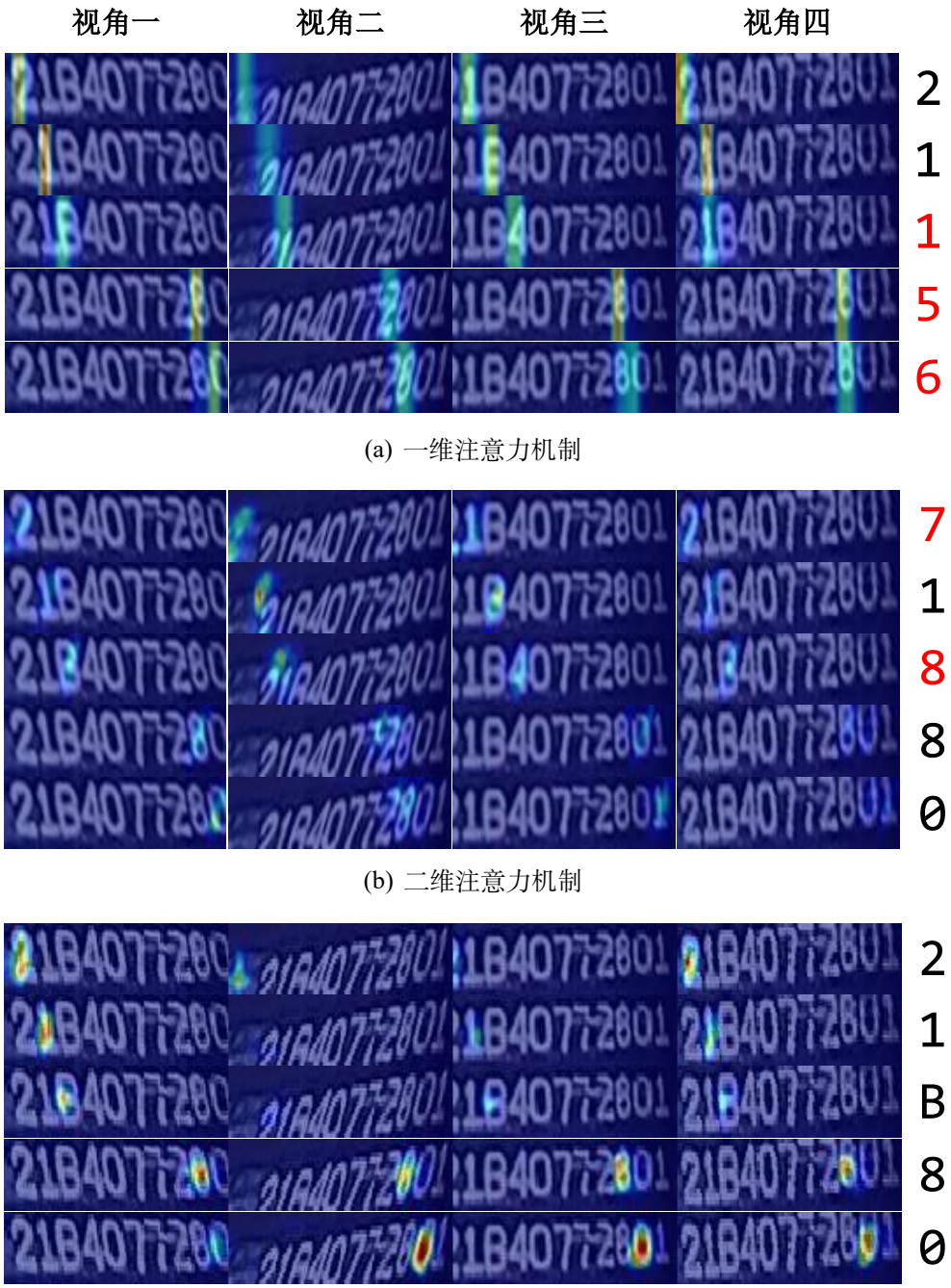
首先我们可视化了不同注意力机制下的注意力图以考察它们对字符的定位精度，如图3-11所示。图3-11中热力图颜色越深代表对应位置的特征越受到模型关注。图右侧表示了模型根据该注意力图所对应的特征得到的字符类别，黑色代表识别正确，红色则代表识别错误。从可视化结果可以看出：

- 一维注意力机制的关注区域呈矩形状，故当识别视角二的倾斜文本时的定位精确度不高，故无法得到十分准确的识别结果。
- 二维注意力机制的关注区域形状任意，能较灵活地定位到单个视角中的字符位置。然而，二维注意力机制无法从其他视角中获得额外的信息，在存在遮挡情况下（如视角三）将出现较严重的注意力偏移现象。
- 从可视化结果来看，本文提出的跨视角注意力机制的关注区域形状较灵活。对于单个字符而言，当其出现在较清晰的视角，其所在区域的热力图越深，反之则越浅。也就是说，其既具有二维注意力机制的二维关注能力，也能够根据多视角之间的区别灵活选择特征，具有很强的多视角识别能力。

其次，我们在测试集上测试了三种注意力机制的识别准确率。如表3-3所示，相比于一维和二维注意力机制，本文提出方法在两个数据集分别提升了 12.4%、6.3% 和 9.49% 和 4.75%，具有较大的优势。这表明本文提出的 ITViewer 能够同时适用于平面文字和曲面文字的多视角识别，具有一定的泛化性。

表 3-3　不同注意力机制的性能对比
Table 3-3　Comparison between different attention methods

注意力机制	SB-M	CIN-M
一维	37.96	54.01
二维	40.87	55.56
跨视角	**50.36**	**60.31**



(a) 一维注意力机制

(b) 二维注意力机制

(c) 跨视角注意力机制（本文方法）

图 3-11　不同注意力机制的注意力图可视化
Figure 3-11　The visualization of different attention mechanisms

3.3.3　与现有方法的对比

为了验证本文提出的 ITViewer 在识别多视角图像中的工业文本上的创新型和有效性，本文测试了现有方法在 CIN-M 和 SB-M 上的识别性能。为了公平比较，这些方法的特征提取部分均使用 ResNet45 网络，并都使用 PlanarText 和 CylinderText 两个合成数据集进行训练。由于大多数现有文字识别算法仅接受单视角的输入，本文在测试时基于 KMP 匹配算法将不同视角的识别结果整合为最终识别结果。表3-4列出了所有方法的准确率，从中我们可以看出：

- 在单视角算法中，由于多视角拍摄导致的文字倾斜变形等特征，CRNN 与 Rosetta 等针对水平文字的方法精度较低，而 SAR 等适用于弯曲文字识别的方法精度相对较高。相比于这些单视角算法，ITViewer 由于其更强的特征融合能力在两个数据集上分别提升了 6.56% 和 7.93%。
- 在 SB-M 数据集上，现有的多视角算法未能体现较大的优势。这是因为它们均未考虑大幅度的视角变化。T^2DAR 方法最初提出于识别视频中的字幕文字，故在 CIN-M 数据集上表现出较高的准确度，但不太适用于 SB-M 的数据分布。相比于现有多视角算法，ITViewer 由于其更强的特征表示能力在两个数据集上分别提升了 8.75% 和 5.28%。

3.3.4　方法适用性分析以及与 ITScanner 的结合

本文提出的 ITViewer 虽然在大部分情况下表现良好，但仍然不适用于一些极端情况。如表3-5所示（错误识别的字符被标为红色），当每个视角图像中的文字均存在缺损或者强烈光照变化时，ITViewer 即使观察到了所有视角的图像特征也无法正确识别。可喜的是，我们可以结合 ITViewer 的跨视角特征融合能力和第二章提出的 ITScanner 的工业语义建模能力共同应对该情况。通过实验，我们发现上述想法是颇为有效的。

3.4　本章小结

从不同视角拍摄同一文字实例，我们将得到具有互补性的多视角图像。基于这一动机，本章提出了基于多视角融合的文字识别方法 ITViewer，包含

表 3-4　与现有方法的对比
Table 3-4　Comparison with state-of-the-arts

方法		SB-M	CIN-M
单视角	CRNN[6]	35.77	47.61
	Rosetta[13]	37.96	46.56
	SAR[25]	41.61	52.38
	SRN[10]	40.87	50.26
	ABINet[12]	43.80	49.21
	VisionLAN[11]	40.87	51.32
多视角	T^2DAR[52]	40.87	55.03
	Zbigniew 等人[61]	41.61	52.91
	ITViewer（本文方法）	**50.36**	**60.31**

基于自注意力的特征融合和基于跨视角注意力机制的预测解码。本章还提出了一种多视角文字图像生成方法，并生成了多视角平面文字数据集 PlanarText 和曲面文字数据集 CylinderText。另外，本章还收集了两个真实多视角文字数据集 SB-M 和 CIN-N 作为测试基准。实验表明 ITViewer 在两个测试集上的识别准确率均高于现有方法，并且可以同时处理平面文字和曲面文字识别，具有一定的鲁棒性。

表 3-5　失败案例分析以及与 ITScanner 的结合
Table 3-5　The failure cases and the combination with ITScanner

多视角图像		
ITViewer	HTIU0881611	21B40771502
ITViewer + ITScanner	HTTU8881611	21B40771802


20003506

第四章　工业现场监控视频下的文字识别系统设计

本章以 ITScanner 和 ITViewer 两个工业文字识别方法为核心算法，设计了一种面向工业现场监控视频的文字识别系统。

4.1　系统需求分析

工业现场中存在大量的文字识别任务。以往的工业文字识别应用一般部署于移动端设备上，仍需要人工完成文字区域的准确定位，距离理想的工业自动化、智能化还有一定差距。相较于前者，部署于工业现场的视频监控系统中的文字识别应用具有以下优势：1）全自动性，即完全通过机器视觉完成文字定位和识别，无需额外的人工协助；2）云计算特性，即能够利用云计算的高性能优势集中处理文字识别任务，节省了边缘设备的成本投入；3）实时性，即文字识别过程与工业现场的图像采集过程同步，能够实时反映当前生产状况。

为此，面向工业现场监控视频的文字识别系统需实现以下需求：1）准确检测监控视频中文字出现的时间和位置；2）准确识别监控视频中的文字内容；3）较小的时间复杂度，保证算法的实时运行。相比于已有的工业文字识别应用，面向工业现场监控视频的文字识别系统面临多方面的挑战，也具有更加强大的功能，具有较大的研究价值。

4.2　系统整体设计

本节首先介绍了面向工业现场监控视频的文字识别系统的整体框架，并在该框架逐一介绍了各个组成部分的具体结构和算法流程。针对工业现场监控视频的文字识别系统的准确性和实时性的需求，在常见的文字识别系统的基础上，添加了文字跟踪和基于轨迹的再识别步骤。文字跟踪有利于文字目标的精确定位，而基于轨迹的再识别步骤运用了第二、三章提出了文字识别方法，极大地提高了识别精度。

4.2.1　整体框架

视频可看作一连串的图像帧。常见的文字识别系统通常包含逐帧文字检测和文字识别两个步骤。本文设计的文字识别系统在此基础上添加了文字跟踪和基于轨迹的再识别步骤。文字跟踪步骤通过利用视频相邻图像之间的关联性进一步提升了文字检测的定位精度，获得了每个文字目标在视频中的运动轨迹，为后续的精确识别提供了大量有价值的信息。基于轨迹的再识别步骤首先将文字轨迹转化为多视角文字图像，随后调用第二、三章设计的两种工业文字识别方法 ITScanner 和 ITViewer 获得更加准确的识别结果。

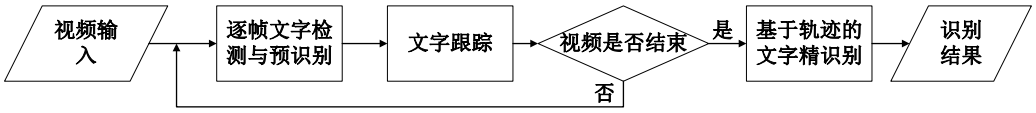


图 4-1　文字识别系统整体框架
Figure 4-1　The framework of the text recognition system

4.2.2　文字检测与预识别

该模块是视频数据进入系统后遇到的第一个运算模块，其主要作用是确定视频中何时何处出现了文字, 并初步识别相应区域中的文字内容，为后续流程提供一定的信息指导。

在监控视频中文字区域的占比通常较小，若是直接输入到文字识别方法中，将引入大量的背景信息的干扰，导致不理想的识别结果。因此在识别之前，一般会采用文字检测模型准确定位原始图像中的文字区域。随后，我们裁剪出这些区域后再输入到文字识别方法中，分别获得每个区域中存在的文字内容。考虑到工业现场中文字形状方向多变，该模块应能识别任意形状的文字区域。另外，系统的实时性也是模型性能的重要组成部分，该模块作为系统的运行次数最频繁的一部分，应选用推理速度较快的算法。基于上述的分析，本文选择了 DBNet[62]和 CRNN[6]两个方法。

DBNet 全称可微分二值化网络。作为一种基于图像分割的文字检测方法，DBNet 能够更加精确地定位各种形状的文字，如曲形文字和斜向文字，十分符合本文场景特点。另外，DBNet 提出了一种可微分二值化模块，使得

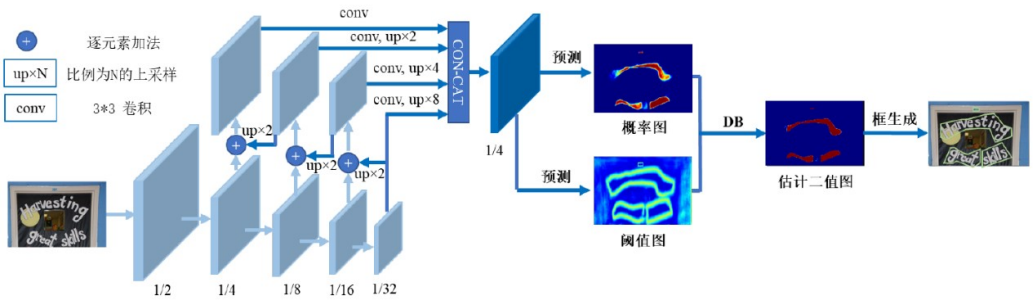


图 4-2　DBNet 网络结构[62]

Figure 4-2　The network structure of DBNet

后处理的二值化步骤和分割网络能够联合优化，较大程度地简化了后处理步骤，推理时能够一秒能处理 60 张图像，符合实际应用中的实时性要求。

CRNN 算法是一种基于 CTC 的文字识别算法。该方法遵循第1.2.1节介绍的文字识别范式，包括基于 CNN 的特征提取、基于 RNN 的序列建模以及基于 CTC 的预测解码。CRNN 具有识别效率高、易部署等优点，在推理时能够一秒处理 200 多张输入图像，是工业界广泛使用的一种文字识别算法。

在实际运行中，系统首先将视频画面逐帧输入到 DBNet 中定位文字区域，再将每个区域都缩放为宽 128、高 32 的文本行图像输入到 CRNN 中，得到该区域中的文字内容。经过文字检测和预识别后，我们将在后续流程中使用 $O=(x,y,w,h,\gamma,p)$ 来表示一个文字区域，其中 x,y,w,h 表示该区域的中心点坐标和长宽，γ 表示文字内容，p 代表该区域的置信度。

4.2.3　文字跟踪

相比于静态图片，视频中相邻画面存在紧密的联系。合理运用这种联系能够极大程度地提升文字识别的准确率。因此，本文通过多文字目标跟踪来确定不同帧中相同文本的位置。具体地，我们考虑了一种基于检测的跟踪方法，该方法通过将连续帧中的文本检测和识别结果相关联来实现。

(1)　轨迹关联规则

在文字跟踪中，当前帧 f_t 中出现的文字目标和前一帧 f_{t-1} 的相互匹配至关重要，关键在于度量不同文字区域之间的相似性。本文根据文字目标的特点，设计了两方面的相似性度量规则：

- 位置相似度。给定当前帧 f_t 中的文字区域 a 和前一帧 f_{t-1} 中的文字区域 b，如果两区域表示的为同一个文字目标，则它们在空间上应该相邻较近。故本文使用两区域之间的交并比来计算两区域间的位置相似度，即：

$$S_1(a,b) = \frac{\text{Area}(a \cap b)}{\text{Area}(a \cup b)} \tag{4.1}$$

- 内容相似度。如果两个文字区域表示的为同一个文字目标，则它们的文字内容应该相同。故本文使用前一步中获得的预识别结果之间的编辑距离来计算两文字区域之间的内容相似度，即：

$$S_2(a,b) = \frac{2 * \text{EditDis}(\gamma_a, \gamma_b)}{\text{len}(\gamma_a) + \text{len}(\gamma_b))} \tag{4.2}$$

基于以上定义，本文在文字跟踪时使用的相似度规则为：

$$S(a,b) = k_1 S_1(a,b) + k_2 S_2(a,b) \tag{4.3}$$

其中 k_1, k_2 是两个权重系数，用于平衡两种相似度规则的影响。当两区域之间的相似度小于一定阈值 S_T 时，本方法则认为这两个区域中为同一个文字。

在实际运行中时，f_t 中的某些区域可能与 f_{t-1} 中的多个区域都能匹配成功。为了处理这种情况并获得最优的匹配结果，本文采用了匈牙利算法[63]来求解该匹配问题。

(2)　轨迹的创建与结束

当新的文字目标进入或离开视频范围时，轨迹也应相应地被创建或者结束。当存在一个文字区域与其他任何区域之间的相似度均小于阈值 S_T 时，我们认为该文字区域对应于一个新的文字目标。为了避免可能的误检测，在接下来的若干帧中，若该区域成功地完成了一次匹配，则证实了该文字目标的存在性，我们在该时间点创建新的轨迹。当一个已有轨迹对应的区域经过 t 帧图像后，仍然无法完成新的匹配，这说明该区域对应的文字已经离开了视频范围。为了防止轨迹数量无限制地增加，该类失活轨迹应被及时地终止。

通过文字跟踪步骤后，我们将使用轨迹来描述一个文字目标，即：

$$T = \{O_{t_s}, \cdots, O_{t_e}\}, \quad O_i = (x, y, w, h, \gamma, p)$$

其中 t_s, t_e 分别为该目标出现和消失的时间点。

4.2.4　基于轨迹的文字精识别

文字轨迹包含了大量有价值的信息，它既包含了每一帧中该文字的位置信息和大致内容，还描述了该文字在视频中的运动情况。在本阶段，系统将综合轨迹中的信息，结合第三章提出的 ITViewer 的跨视角特征融合能力和第二章提出的 ITScanner 的工业语义建模能力完成更准确的工业文字识别。

文字轨迹中包含了一系列的文字区域，可视为对同一目标的多角度拍摄结果。由于运动模糊、对焦不准等原因，摄像头在拍摄运动物体时所得图像质量不稳定。为了筛选出较清晰的文字图像，系统先按照轨迹中的区域置信度 p 进行排序，并选择前 N 个图像输入到 ITViewer 中。ITViewer 是一种基于多视角特征融合的文字识别方法，能结合多个输入图像中的特征信息完成准确的识别任务。首先，ITViewer 基于自注意力机制寻找不同视角图像中的相似区域，并通过加权平均的方式融合来自不同视角的特征，获得跨视角特征图。随后，ITViewer 采用一种跨视角注意力机制从跨视角特征图中分别寻找每个字符对应的特征表示，最终完成字符分类。

虽然前文已通过实验验证了 ITViewer 的精度，该方法仍有其局限性。为了进一步提升整个系统的文字识别准确率，本文将第二章提出的 ITScanner 的语言模型部分作为该系统的文本纠错模型，利用其学习到的工业语义模式判断识别结果中是否存在错误，并加以修正。

4.3　实验及分析

4.3.1　测试数据集与评价指标

(1)　测试数据集

为了真实地体现该系统在工业现场的实际效果，本文以某钢铁厂的物

料入口为具体场景，以该场景的监控视频为本文实验的测试数据。这些视频片段具有如下特点：（1）夜晚拍摄，场景中存在较多的运动或闪烁的光源干扰；（2）运动文本，所需识别的文本均喷印于运动货车架上；（3）单个视频中包含的目标文本数量不确定。图4-3展示了这些视频片段的若干截图。根据实际需求，本文仅关注视频中的集装箱编号文本。

(a)　　(b)

图 4-3　测试所用的监控视频
Figure 4-3　The surveillance video in our experiments

(2)　评价指标

本文设计的文字识别系统包含文字检测、跟踪、识别三个层面的工作。为了全面地测试该系统的性能，本文参考 Yin 等人的工作[64]，分别选择了如下符合实际需求的指标。

在检测方面，由于视频中存在大量的无关文本，而本文仅关注视频中的集装箱编号文本，故本文选择召回率（Recall Rate）作为文字检测的指标，以度量文字检测算法在检测目标文本时的精确程度。

在跟踪方面，本文着重关注所得轨迹与真实轨迹的重合程度，故选择了多目标跟踪领域中的两个评价指标 MOTP 和 MOTA。MOTP 主要用于衡量跟踪的位置误差，该值越高，跟踪的误差越小。MOTA 统计了跟踪过程中系统的漏检、误检等情况，该值越高，这些情况发生的次数越少，跟踪精度越高。

在识别方面，本文使用严格的单词级别准确率来衡量识别情况。单词级别准确率要求识别结果与真实标签完全一致，无任何漏识、错识、多识等情况。

4.3.2　各模块实验分析

本小节在测试数据集上分别对系统各模块的性能进行了全面的评估，并讨论了系统中各类超参数的影响。

(1)　文字检测

文字检测是本系统的重要先导步骤，其定位精度与文字识别的准确度息息相关。另外，由于文字检测需处理视频中的每一帧画面，是该系统中运行最频繁的部分，故其运行速度同样十分关键。本文比较了若干常用的文字检测方法的定位精度和运算速度。

表 4-1　不同文字检测方法的比较
Table 4-1　Comparison of different text detection methods

方法	骨干网络	召回率 (%)	运算速度（帧/秒）
CTPN[65]	VGG-16	72.1	7.1
EAST[66]	PVANet	83.3	13.2
TextBoxes++[67]	VGG-16	76.7	2.3
PSENet[68]	ResNet-18	**84.5**	3.9
CRAFT[69]	VGG-16	84.3	8.6
DBNet[62]	ResNet-18	78.4	**50**

从表4-1可以看到，与其他方法相比，DBNet 具有极快的运算速度。由于在实际应用中需处理的视频帧率大多为 30 帧/秒，故仅有 DBNet 方法的速度达到了实时性要求。在定位精度上，DBNet 仅比 PSENet 略低 6.1%，仍然保持在一个较高的水平，能满足实际的检测需求。因此，本文选择 DBNet 方法作为系统的文字检测方法。

(2)　文字跟踪

在进行文字跟踪时，本文提出了位置相似度和内容相似度来共同判断视频前后的两个文字区域是否为同一文字区域。本小节设计了相关实验来验证这两种相似度的作用，并讨论了式(4.3)中 k_1, k_2 的取值对跟踪精度的影响。

从表4-2可以看到，当仅使用位置相似度（即 $k_1 = 1, k_2 = 0$）时，系统

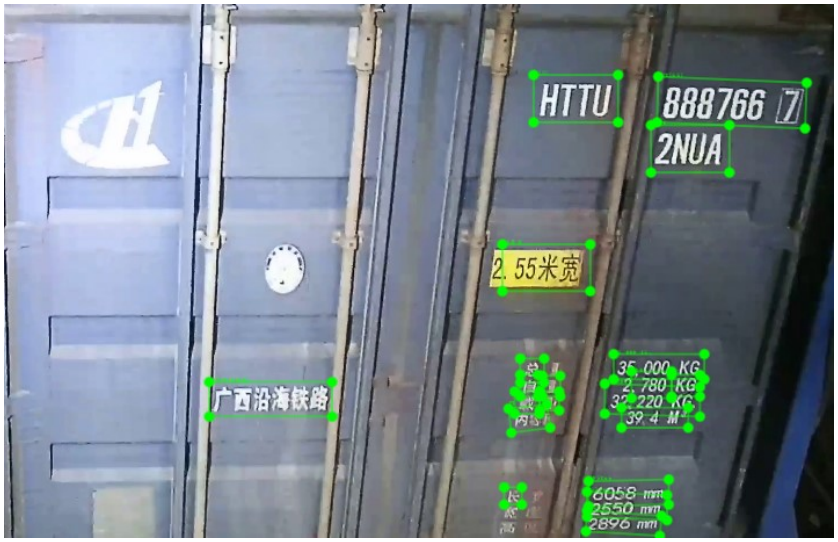


图 4-4　DBNet 文字检测结果示例
Figure 4-4　The detection results of DBNet

的 MOTA 和 MOTP 分别下降了 4.31% 和 3.36%。当仅使用内容相似度（即 $k_1=0,k_2=1$）时，系统的 MOTA 和 MOTP 分别下降了 4.95% 和 4.70%。因此，这两种相似度是相辅相成的，均对文字的跟踪有突出作用。通过不断改变 k_1,k_2 取值，本文的实验结果显示当 $k_1=1,k_2=5$ 时系统的跟踪准确率达到了最佳水平。因此，在后续实验中本文均使用该取值。

表 4-2　不同参数下的跟踪精度测试
Table 4-2　The tracking accuracy with different parameters

k_1	k_2	MOTA(%)	MOTP(%)
1	0	56.61	65.07
0	1	55.97	63.73
1	1	56.74	66.58
1	2	58.29	67.05
1	3	60.40	67.94
1	5	60.92	**68.43**
1	10	**61.32**	67.09
1	20	57.63	67.10

(3)　文字识别

文字识别是系统的核心步骤。本文结合了 ITViewer 和 ITScanner 两种

方法的优势，并基于跟踪轨迹中的图像信息完成了更加准确的文字识别。为验证该方法的有效性和必要性，本文进行了相关消融实验。

表4-3中展示了实验所用四种文字识别策略。法一不进行额外的识别工作，仅从轨迹中挑选置信度最高的预识别结果作为该轨迹对应文本的识别结果；法二在法一的基础上引入了 ITScanner 的文字纠错功能；法三仅使用 ITViewer 完成文字精识别；法四在法三的基础上引入了 ITScanner 的文字纠错功能。从表4-3可以看出，仅使用预识别的结果将使得系统准确率较低。无论是采用 ITViewer 进行文字精识别，或是引入 ITScanner 的文字纠错能力均能使系统的识别准确率大幅提高。

表 4-3　不同策略的文字识别准确率比较

Table 4-3　The comparison of text recognition under different policies

	预识别	ITViewer	ITScanner	识别准确率
法一	✓			48.72
法二	✓		✓	54.65
法三		✓		53.96
法四		✓	✓	**57.32**

表 4-4　与现有方法的对比

Table 4-4　Comparison with state-of-the-arts methods

方法	检测召回率（%）	跟踪 MOTP（%）	识别准确率（%）	运算速度（帧/秒）
Wang 等人[50]	76.7	**72.87**	54.26	13.4
Lyu 等人[20]	82.6	65.13	53.41	6.9
Cheng 等人[51]	**84.1**	69.38	55.14	11.9
本文方法	78.4	68.43	**57.32**	37.2

4.3.3　与现有方法的对比

本小节在测试数据集上对比了若干现有方法的检测召回率、跟踪精度 MOTP、识别准确率以及运算速度。从表4-4可以看出，为了考虑系统的实

时性，本文算法牺牲了一部分的检测和跟踪精度，保证了较高的运算速度（37.2 帧/秒）。但在识别方面，由于系统采用的 ITViewer 和 ITScanner 在识别工业文本上具有优势，故系统的识别准确率仍然是最高的水平（57.32%）。

4.4　本章总结

本章立足于实际工业应用场景，以本文设计的工业文字识别算法 ITScanner 和 ITViewer 为核心算法，设计了一种集文字检测、跟踪、高精度识别的工业文字识别系统。该系统包含基于 DBNet 和 CRNN 的实时文字检测和预识别、基于文字位置和文字内容特点的文字跟踪、以及基于 ITScanner 和 ITViewer 的文字精识别。实验证明该系统能够较为精确地检测和识别工业现场监控视频中的文字，并且运算速度能达到实时性要求。

第五章 总结与展望

5.1 工作总结

本课题主要关注工业场景中的文字识别任务。针对工业现场环境多变、干扰较强的挑战，本课题提出在文字识别任务中引入多源信息融合的思想，利用多源输入的共识性和互补性的特点，构建鲁棒的工业文字识别系统。

通过引入文字语义信息，本文提出了一种基于工业语义与视觉融合的文字识别方法（ITScanner）。为了准确地描述工业文字中的语义，本文进行了大量的调查研究，并总结了工业文本中的三大语义模式——分段式语义、段间独立性和段内关联性。以此为理论基础，本文基于卷积神经网络、自注意力机制模块、多层感知单元等网络结构实现了工业语义的建模，并提出了基于基于工业语义与视觉融合的文字识别方法。该方法在真实工业场景数据集上相比于基准方法有较大的精度提升，体现了语义信息对于工业文字识别的重要作用。

通过引入多视角图像信息，本文提出了一种基于多视角特征融合的文字识别方法（ITViewer）。该方法能够接受多视角的图像输入，通过自注意力机制的全局关联性融合不同视角中的文字特征以获得更鲁棒的特征表示。考虑到多视角文字图像的收集与标注十分耗时耗力，本文基于相机成像原理，提出了一种多视角文字图像的生成方法，并生成了 PlanarText 和 CylinderText 两个数据集。实验证明 ITViewer 具有较强的特征融合能力，能够较准确地遮挡或模糊图像中的工业文字，精度大大超越了基准方法。

最后，本文以 ITScanner 和 ITViewer 为核心算法，构建了一种集文字检测、跟踪、高精度识别的工业文字识别系统。该系统分为三阶段流程：实时文字检测和预识别、基于文字位置和文字内容特点的文字跟踪以及基于 ITScanner 和 ITViewer 的文字精识别。实验证明该系统能够较为精确地检测和识别工业现场监控视频中的文字，并且运算速度能达到实时性要求。

5.2 课题研究展望

尽管本文基于多源信息融合的思想提出了若干抗干扰能力强的工业文字识别方法，但是仍有一些可以去挖掘和深入的研究方向：

(1) 高层次工业语义建模

本文提出了一种基于工业语义和视觉融合的文字识别算法（ITScanner），能够较鲁棒地识别低质量图像中的工业文本。然而，本文提出的工业语义建模方法仅考虑了单个文字区域中的语义信息，属于字符级别的低层次语义。工业现场拍摄的图像中往往存在多个文字区域。这些文字出现在同一时间、同一地点，所携带的语义信息之间存在千丝万缕的关联性。不同文字区域之间关联性的描述与建模，属于语句级、段落级等高层次的语义建模，需要更多自然语言处理领域的相关技术。该问题可以从以下两个方面去探索。一是探索如何融合不同文字区域的特征，完成多文字区域的并行预测解码。二是参考借鉴 BERT[70]等相关工作的语义建模思想。由于该类模型的参数量较大，在文字识别应用中需着重考虑模型的轻量化。

(2) 强泛化性的多视角文字识别

本文涉及的多视角图像均为同一相机通过多次拍摄所得。在实际应用中，由于工业现场监控将存在较多的交叉区域，我们能通过多个相机同时拍摄同一文字区域获得多视角文字图像。虽然这使得多视角图像的收集变得极为便利，但由于不同相机的内在参数（如焦距、像素大小等）和拍摄条件（如曝光时间、光照强度等）均不一致，不同视角图像之间的特征对齐和融合将受到较大干扰。这对多视角文字识别算法的泛化性提出了更高的要求。本文认为该问题可以从数据和算法两方面进行深入研究。第一，应增多数据集的覆盖场景和所用的相机型号。考虑到数据集的采集成本，笔者认为可以利用商用游戏引擎的强大渲染能力，构建虚拟场景和不同配置的虚拟摄像头，自动生成逼真的工业文字数据集。第二，可以探索数据层面、特征层面、决策层面等多层面的混合融合方式以提升模型的泛化能力。

(3) 基于多源信息融合的端到端视频文字识别

本文提出的工业文字识别系统虽然涵盖了文字检测、跟踪和识别三个阶段，但多源信息融合的思想仅体现在文字识别阶段中。为进一步提升该系统的识别精度，可以从以下两个方面去探索。其一，在文字检测与跟踪任务中引入多源信息融合的思想，借助多源信息的共识性和互补性提升文字检测与跟踪精度。其二，构建端到端的视频文字识别算法，将文字检测、跟踪和识别三个阶段整合到统一框架中，并通过端到端的训练学习到最优的模型参数。

参考文献

[1] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks[J]. Advances in Neural Information Processing Systems, 2015, 28.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv:1409.1556, 2014.

[3] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.

[4] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. 2019: 6105-6114.

[5] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. ArXiv:2010.11929, 2020.

[6] Shi B, Bai X, Yao C. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(11): 2298-2304.

[7] Shi B, Wang X, Lyu P, et al. Robust scene text recognition with automatic rectification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 4168-4176.

[8] Cheng Z, Bai F, Xu Y, et al. Focusing attention: Towards accurate text recognition in natural images[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 5076-5084.

[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems. 2017: 5998-6008.

[10] Yu D, Li X, Zhang C, et al. Towards accurate scene text recognition with semantic reasoning networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2020: 12113-12122.

[11] Wang Y, Xie H, Fang S, et al. From Two to One: A New Scene Text Recognizer with Visual Language Modeling Network[C]//Proceedings of the IEEE International Conference on Computer Vision. 2021: 14194-14203.

[12] Fang S, Xie H, Wang Y, et al. Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2021: 7098-7107.

[13] Borisyuk F, Gordo A, Sivakumar V. Rosetta: Large scale system for text detection and recognition in images[C]//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 71-79.

[14] Baek J, Kim G, Lee J, et al. What is wrong with scene text recognition model comparisons? dataset and model analysis[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 4715-4723.

[15] Graves A, Fernández S, Gomez F, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C]//Proceedings of the 23rd International Conference on Machine Learning. 2006: 369-376.

[16] Hu W, Cai X, Hou J, et al. Gtc: Guided training of ctc towards efficient and accurate scene text recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence: vol. 34: 07. 2020: 11005-11012.

[17] Du Y, Chen Z, Jia C, et al. SVTR: Scene Text Recognition with a Single Visual Model[J]. ArXiv:2205.00159, 2022.

[18] Liao M, Zhang J, Wan Z, et al. Scene text recognition from two-dimensional perspective[C]//Proceedings of the AAAI conference on Artificial Intelligence: vol. 33: 01. 2019: 8714-8721.

[19] Wan Z, He M, Chen H, et al. Textscanner: Reading characters in order for robust scene text recognition[C]//Proceedings of the AAAI conference on Artificial Intelligence: vol. 34: 07. 2020: 12120-12127.

[20] Lyu P, Liao M, Yao C, et al. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 67-83.

[21] 廖明辉. 自然场景端到端文字识别方法研究[D]. 武汉：华中科技大学, 2021.

[22] Zhan F, Lu S. Esir: End-to-end scene text recognition via iterative image rectification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2059-2068.

[23] Shi B, Yang M, Wang X, et al. Aster: An attentional scene text recognizer with flexible rectification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(9): 2035-2048.

[24] Yang X, He D, Zhou Z, et al. Learning to Read Irregular Text with Attention Mechanisms[C]//Proceedings of the International Joint Conference on Artificial Intelligence. 2017: 3280-3286.

[25] Li H, Wang P, Shen C, et al. Show, attend and read: A simple and strong baseline for irregular text recognition[C]//Proceedings of the AAAI conference on Artificial Intelligence: vol. 33: 01. 2019: 8610-8617.

[26] 甘吉. 手写文字识别及相关问题算法研究[D]. 北京：中国科学院大学, 2021.

[27] 金典. 基于时空上下文的视频文字识别算法研究[D]. 武汉：华中科技大学, 2021.

[28] Yue X, Kuang Z, Lin C, et al. Robustscanner: Dynamically enhancing positional clues for robust text recognition[C]//European Conference on Computer Vision. 2020: 135-151.

[29] Lyu P, Yang Z, Leng X, et al. 2d attentional irregular scene text recognizer[J]. ArXiv:1906.05708, 2019.

[30] Yan R, Peng L, Xiao S, et al. Primitive representation learning for scene text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 284-293.

[31] Sheng F, Chen Z, Xu B. NRTR: A no-recurrence sequence-to-sequence model for scene text recognition[C]//2019 International conference on document analysis and recognition (ICDAR). 2019: 781-786.

[32] Yang L, Wang P, Li H, et al. A holistic representation guided attention network for scene text recognition[J]. Neurocomputing, 2020, 414: 67-75.

[33] Xie X, Fu L, Zhang Z, et al. Toward Understanding WordArt: Corner-Guided Transformer for Scene Text Recognition[C]//European Conference on Computer Vision. 2022: 303-321.

[34] Qiao Z, Zhou Y, Wei J, et al. PIMNet: a parallel, iterative and mimicking network for scene text recognition[C]//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 2046-2055.

[35] 李洋, 赵鸣, 徐梦瑶, 刘云飞, 钱雨辰. 多源信息融合技术研究综述[J]. 智能计算机与应用, 2019, 9(05): 186-189.

[36] Nagai A. On the improvement of recognizing single-line strings of japanese historical cursive[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). 2019: 621-628.

[37] Wei J, Chen K, He J, et al. A new approach for integrated recognition and correction of texts from images[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). 2019: 615-620.

[38] Qiao Z, Zhou Y, Yang D, et al. Seed: Semantics enhanced encoder-decoder framework for scene text recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13528-13537.

[39] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information[J]. Transactions of the Association for Computational Linguistics, 2017, 5: 135-146.

[40] 程昌旭. 基于文字语义与先验分布的图像文本识别方法研究[D]. 武汉: 华中科技大学, 2021.

[41] Bhunia A K, Sain A, Kumar A, et al. Joint visual semantic reasoning: Multi-stage decoder for text recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 14940-14949.

[42] Wang W, Liu X, Ji X, et al. Ae textspotter: Learning visual and linguistic representation for ambiguous text spotting[C]//European Conference on Computer Vision. 2020: 457-473.

[43] Zhong D, Lyu S, Shivakumara P, et al. SGBANet: Semantic GAN and Balanced Attention Network for Arbitrarily Oriented Scene Text Recognition[C]//European Conference on Computer Vision. 2022: 464-480.

[44] Wang P, Da C, Yao C. Multi-granularity Prediction for Scene Text Recognition[C]//European Conference on Computer Vision. 2022: 339-355.

[45] Zhang X, Zhu B, Yao X, et al. Context-based Contrastive Learning for Scene Text Recognition[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2022: 3353-3361.

[46] Feng Y, Zhang Z, Zhao X, et al. Gvcnn: Group-view convolutional neural networks for 3d shape recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 264-272.

[47] Qiu H, Wang C, Wang J, et al. Cross view fusion for 3d human pose estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4342-4351.

[48] Zhang Y, Liao Q, Ding L, et al. Bridging 2D and 3D Segmentation Networks for Computation-Efficient Volumetric Medical Image Segmentation: An Empirical Study of 2.5 D Solutions[J]. Computerized Medical Imaging and Graphics, 2022: 102088.

[49] Chen C, Biffi C, Tarroni G, et al. Learning shape priors for robust cardiac MR segmentation from multi-view images[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. 2019: 523-531.

[50] Wang X, Jiang Y, Yang S, et al. End-to-end scene text recognition in videos based on multi frame tracking[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR): vol. 1. 2017: 1255-1260.

[51] Cheng Z, Lu J, Niu Y, et al. You only recognize once: Towards fast video text spotting[C]//Proceedings of the 27th ACM International Conference on Multimedia. 2019: 855-863.

[52] Tian S, Yin X C, Su Y, et al. A unified framework for tracking based text detection and recognition from web videos[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(3): 542-554.

[53] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2117-2125.

[54] Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition[C]//Workshop on Deep Learning, Advances in Neural Information Processing Systems. 2014.

[55] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2016: 2315-2324.

[56] Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 robust reading competition[C]//2013 12th international conference on document analysis and recognition. 2013: 1484-1493.

[57] Mishra A, Alahari K, Jawahar C. Top-down and bottom-up cues for scene text recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2012: 2687-2694.

[58] Wang K, Babenko B, Belongie S. End-to-end scene text recognition[C]//Proceedings of the IEEE International Conference on Computer Vision. 2011: 1457-1464.

[59] Chen J, Li B, Xue X. Scene text telescope: Text-focused scene image super-resolution[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 12026-12035.

[60] Hartley R, Zisserman A. Multiple view geometry in computer vision[M]. Cambridge university press, 2003.

[61] Wojna Z, Gorban A N, Lee D S, et al. Attention-based extraction of structured information from street view imagery[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR): vol. 1. 2017: 844-850.

[62] Liao M, Wan Z, Yao C, et al. Real-time scene text detection with differentiable binarization[C]//Proceedings of the AAAI conference on artificial intelligence: vol. 34: 07. 2020: 11474-11481.

[63] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval research logistics quarterly, 1955, 2(1-2): 83-97.

[64] Yin X C, Zuo Z Y, Tian S, et al. Text detection, tracking and recognition in video: a comprehensive survey[J]. IEEE Transactions on Image Processing, 2016, 25(6): 2752-2773.

[65] Tian Z, Huang W, He T, et al. Detecting text in natural image with connectionist text proposal network[C]//European conference on computer vision. 2016: 56-72.

[66] Zhou X, Yao C, Wen H, et al. EAST: an efficient and accurate scene text detector[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[67] Liao M, Shi B, Bai X. TextBoxes++: A single-shot oriented scene text detector[J]. IEEE Transactions on Image Processing, 2018, 27(8): 3676-3690.

[68] Wang W, Xie E, Li X, et al. Shape robust text detection with progressive scale expansion network[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[69] Baek Y, Lee B, Han D, et al. Character region awareness for text detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9365-9374.

[70] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. ArXiv:1810.04805, 2018.

攻读学位期间学术论文和科研成果目录

[1] Guowei Deng, Jingzheng Tu, Cailian Chen*, Jianping He, and Xinyi Le. Knowledge-Based Scene Text Recognition for Industrial Applications[C]// The 23rd IEEE International Conference on Industrial Technology (ICIT'22). IEEE, Shanghai, 2022.（获"最佳学生论文"）

[2] 邓国伟、陈彩莲、涂静正、关新平、杨博，"一种基于联合检测与表征提取的多目标跟踪系统和方法"，专利申请号：CN202011510839.1，公开日期：2021 年 4 月 30 日


20003506

致 谢

三年的研究生生活即将步入尾声。回首过往的学习经历，既有曲折坎坷的艰苦探索，亦有柳暗花明的宝贵瞬间，两者均令我收获颇丰。值此论文完成之际，我要向所有关心和支持我的家人、老师、同学表达衷心的感谢。

首先，我要感谢我的家人。感谢父母无微不至的关心和谆谆教诲，令我无论何时何地都有坚实的依靠。感谢杨长发舅舅，作为我在上海唯一的亲人，您的热情和关爱给了我许多家的温暖，您丰富的阅历和见识使我开阔了视野。感谢外公外婆多年以来的照顾，您们殷切的期盼一直都是我奋斗的动力。

其次，我要感谢陈彩莲老师。从本科开始，陈老师敏锐的学术思维和精益求精的品格一直是我学习的榜样。进入实验室后，陈老师一方面提供了良好的科研环境和先进的实验设备，另一方面耐心地教导我深入探索科研问题，激励我持续提高思考深度，向更高质量的科研成果迈进。在陈老师的指导下，我完成了系统化的科研训练，并取得了一系列的科研成果。

非常感谢实验室的何建平老师、乐心怡老师和许齐敏老师给我的督促和指导。感谢各位老师在组会上提出的宝贵意见以及在论文写作上给我的无私帮助。

感谢涂静正师姐。在科研上，静正师姐的自律精神和学术思维一直是我学习道路上的榜样。在生活上，静正师姐乐观向上、虚心待人的品格一直感染着我。也非常感谢实验室的各位同学们，你们的陪伴和友谊是我人生中宝贵的财富。

感谢我的两位室友蔡建伟和吴昊。我们是本科以来就认识的朋友。在研究生阶段，我们朝夕相处、患难与共，度过了一段难忘的时光。小到生活琐事、大到学术巅峰和人生哲理，我们无话不谈、毫无保留，和你们的友谊我将一生珍重。

最后，我要感谢我自己。三年以来，我努力地提升自己各方面的能力，勇敢地走出了自己的舒适区，收获了很多人生经验。在这个过程中，我遇到了不少的困难，有过不少沮丧的瞬间，感谢我能够一路坚持，没有放弃。希望在以后的日子里，我能够再接再厉，继续朝着理想中的自己而努力。

上海交通大学
学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全知晓本声明的法律后果由本人承担。

学位论文作者签名：邓国伟
日期：2023年 2月 16日

上海交通大学
学位论文使用授权书

本人同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于：

☑公开论文

□ 内部论文，保密 □1 年/□2 年/□3 年，过保密期后适用本授权书。

□ 秘密论文，保密 ___ 年（不超过 10 年），过保密期后适用本授权书。

□ 机密论文，保密 ___ 年（不超过 20 年），过保密期后适用本授权书。

（请在以上方框内选择打 “√”）

学位论文作者签名：邓国伟　指导教师签名：陈彩莲
日期：2023年 2月 16日　日期：2023年 2月 16日

20003506

120032910121

上海交通大学硕士学位论文答辩决议书

姓　名	邓国伟	学号	120032910121	所在学科	电子信息
指导教师	陈彩莲	答辩日期	2023-02-08	答辩地点	电信群楼2-410会议室
论文题目	基于多模态融合的工业文字识别方法及应用				

投票表决结果：5 /5 /5（同意票数/实到委员数/应到委员数） 答辩结论：☑通过 ☐未通过

评语和决议：

邓国伟同学的硕士论文《基于多模态融合的工业文字识别方法及应用》围绕环境多变、干扰较强的工业场景中的文字识别问题进行了深入研究，选题具有较好的理论意义和应用价值，论文的主要创新成果如下：

1）设计了一种基于工业语义与视觉融合的文字识别模型，提出了基于并行注意力机制的视觉模型，建立了基于分段语义注意力机制的语言模型，解决了工业文本词库量大，信息密度高导致的语义建模难的问题，提高了文字识别准确率。

2）设计了一种基于多视角特征融合的文字识别模型，提出了基于自注意力的特征融合和基于跨视角注意力机制的预测解码，构建了一种多视角文字图像数据集生成方案，解决了遮挡条件下的文字识别问题。

3）设计了一套面向工业现场监控视频的文字识别系统，提出了一种集文字检测、跟踪、高精度识别的工业文字识别方法，能够较为精确地检测和识别工业现场监控视频中的文字，并且运算速度能满足实时性要求。

论文条理清楚，内容充实，分析合理，结果可靠。论文工作表明作者已经掌握本学科的理论知识和实践能力，具备独立科研与创新能力。答辩中邓国伟同学思路清晰，叙述有条不紊，回答问题准确。经答辩委员会认真讨论并无记名投票，同意通过邓国伟同学工程硕士学位论文答辩，并建议授予其工程硕士学位。

2023 年 2 月 8 日

答辩委员会成员签名	职务	姓名	职称	单位	签名
	主席	杨博	教授	上海交通大学	杨博
	委员	戴文斌	教授	上海交通大学	戴文斌
	委员	殷翔	副教授	上海交通大学	殷翔
	委员	郭逸	助理研究员	上海交通大学	郭逸
	委员	秦凯运	高级工程师	上海宝信软件股份有限公司	秦凯运
	秘书	吴沂红		上海交通大学	吴沂红