

# 电磁泄漏还原图像中的中文文本识别技术研究

吕志强<sup>1,2</sup>, 张磊<sup>1,2</sup>, 夏宇琦<sup>1,2</sup>, 张宁<sup>1</sup>

<sup>1</sup>中国科学院信息工程研究所第四研究室 北京 中国 100093

<sup>2</sup>中国科学院大学网络空间安全学院 北京 中国 100093

**摘要** 现代计算机的显示信号传输过程存在的电磁泄漏, 从电磁泄漏还原得到的图像会受到噪声的严重污染, 使得其中的文本内容难以识别。本文提出了一种新的模型, 利用基于特征强化的神经网络(Feature Enhancement based Neural Network, FENN)对电磁泄漏还原图像中的中文文本进行识别。模型将去噪自编码器(Denoising Autoencoder, DAE)与卷积神经网络(Convolutional Neural Network, CNN)相结合, 对电磁泄漏图像的文本特征进行强化并抑制噪声干扰, 在不损失原始图像信息的情况下将鲁棒特征送入后续的循环神经网络(Recurrent Neural Network, RNN), 最后将连续时间序列分类(Connectionist Temporal Classification Loss, CTC Loss)损失与均方误差损失(Mean Squared Error Loss)结合形成联合损失对模型进行联合训练, 实现无需去噪等常规预处理的中文文本识别。模型在电磁泄漏还原实景数据和公开数据集 RCTW17、CASIA-10k 上进行了测试, 相比于常见的主流识别模型, FENN 在电磁泄漏还原图像中的中文识别率最高提升 5.4%, 体现出明显优势。

**关键词** 电磁泄漏; 去噪自编码器; 特征强化; 中文文本识别; 神经网络

中图分类号 TP183/TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2021.05.14

## Chinese Text Recognition in Electromagnetic Emission Reconstructed Images

LV Zhiqiang<sup>1,2</sup>, ZHANG Lei<sup>1,2</sup>, XIA Yuqi<sup>1,2</sup>, ZHANG Ning<sup>1</sup>

<sup>1</sup>The 4th Laboratory, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

<sup>2</sup>School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100093, China

**Abstract** Electromagnetic emission exists in the process of display signal transmission in modern computers. Therefore, by signal receiving and restoring using eavesdroppers, one can reconstruct the display information emitted from target computer. However, reconstructed images are corrupted by noise, causing difficulty in recognizing its content. In this paper, we propose a new model, using feature-enhancing-based Neural Network (FENN) to recognize Chinese text lines in reconstructed image. The model combines Convolutional Neural Network(CNN) with denoising autoencoder to achieve enhancement of text features and suppress noise interference. Then robust features extracted with image information preserved are feed into the following Recurrent Neural Network(RNN). Finally, with Connectionist temporal classification (CTC) Loss and Mean Squared Error(MSE) loss combined, the model can be trained jointly under a joint loss function, by which the model is able to recognize Chinese text lines in reconstructed images without denoising or any other preprocessing. Experiments were performed on dataset consists of reconstructed images and public datasets including RCTW17 and CASIA-10k. Result shows that our method outperforms common recognition methods by 5.4% at most.

**Key words** electromagnetic emission; denoising autoencoder; feature enhancement; chinese text recognition; neural network

### 1 背景

随着国际信息安全形势的日益严峻, 电子信息设备泄漏信号引起的信息安全和保护问题的危害已引起国内外的广泛关注。现代电子信息设备在运转过程中广泛存在电磁泄漏, 而泄漏的电磁信号中常

常包含大量有用信息, 导致严重的安全性问题<sup>[1-3]</sup>。据研究表明, 现代计算机正常工作时所产生的信息会发生泄漏辐射。电磁泄漏信号包含由信息设备处理的敏感信息, 通过专用信号接收设备和信息设备电磁泄漏发射防护技术(Telecommunications Electronic Material Protected from Emanations Spurious

通讯作者: 张宁, 硕士, 工程师, Email: zhangning@iie.ac.cn。

本课题得到国家重点研发计划课题(No.2018YFF01014303)资助。

收稿日期: 2019-07-10; 修改日期: 2019-10-18; 定稿日期: 2021-03-05

Transmission, 以下简称 TEMPEST)可以截获电磁泄漏信号并恢复其中携带的重要信息。拦截和恢复机密信息会威胁到国家信息的安全。作为该问题的关键技术,电磁泄漏发射防护技术已经成为了全世界主要国家和多个地区的研究热点和研究重点。

TEMPEST 源自美国政府在 20 世纪 60 年代后期开展的一个机密项目,旨在研究如何利用和防范计算机和电信设备通过发射电磁辐射(EMR)导致敏感数据被复现。1985 年,荷兰学者 Van Eck W.第一次公开分析了电磁信息泄漏的机理和显示器电磁辐射造成信息泄漏的实验结果<sup>[1]</sup>。从此 TEMPEST 技术引起了各国政府部门的重视,我国就是从此开始了对 TEMPEST 技术的研究。通过 TEMPEST 技术<sup>[4]</sup>,攻击人员可以突破物理隔离状态下的电子信息设备对其进行攻击,对信息安全造成严重威胁。

对于显示信号而言,其泄漏的严重性远超过其他类型的信号泄漏,因为计算机的显示信号中包含以文本信息为主的大量有用信息。2003 年,Markus G. Kuhn 研究实现了利用 LCD 显示器的电磁泄漏截获还原其显示的内容。日本信息通信研究机构 NICT 在 2004 年公开演示了如何利用高性能测定装置还原显示器中的信息<sup>[5]</sup>。2005 年,日本学者 H Tanaka, O Takizawa 和 A Yamamura 使用电流钳对电源线进行夹持,利用传导泄漏的原理实现了对笔记本电脑显示图像的截获<sup>[6]</sup>。从攻击角度来看,通过对泄漏图像中存在的文本内容进行识别,可以获取目标计算机上存在的敏感信息;从防护角度来看,对泄漏图像中存在的文本内容进行识别可以判断计算机是否存在泄漏敏感信息的可能,进而进行针对性的措施。因此,对显示信号泄漏图像中存在的文本信息进行识别就至关重要。

图片中的文本能够比图片中其他内容提供更加丰富的信息。因此,图像文本识别能够将图像中的文本区域转化成计算机可以读取和编辑的符号。文本识别技术具有广泛用途,例如,文字识别系统常用于图像检索<sup>[7]</sup>,目标定位<sup>[8]</sup>,人机交互<sup>[9]</sup>,辅助导航<sup>[10-11]</sup>。另外,文字识别系统在安防,工业自动化<sup>[12]</sup>等领域还有诸多应用。毫无疑问,文本识别技术对于获取与利用图像中的文本信息至关重要,因此文本识别是计算机视觉领域的热门主题。

传统的识别模型大多基于手工设计特征(hand-crafted features)的文字识别<sup>[18-19]</sup>,但是使用人工设计的特征模板去匹配数据中的隐含的细节和模式需要大量的先验知识,而且对于特征本身具有很高的要求,难以满足数据多变的情况,局限性较强。

Weinman 等人<sup>[20]</sup>将字典,相似度和外形信息结合起来组成联合模型,然后使用稀疏置信传播<sup>[21]</sup>计算最可能的字符内容。Mishra 等人<sup>[22]</sup>使用自下而上和从上到下的模型进行文本识别,该模型使用滑动窗检测可能的字符,然后将检测结果作为自下而上的信息。而从下到上的信息来自一个大型词典的统计信息。最后通过条件随机场(CRF)<sup>[23]</sup>将自下而上和从上到下的信息整合入一个统一模型。

近几年,随着计算机性能的不断提升,以及深度学习的蓬勃发展,逐渐出现了基于深度学习的文本识别模型。Tao Wang 等人<sup>[24]</sup>首先提出了基于卷积神经网络的端到端文本识别。利用神经网络的自动特征提取从训练数据中自动学习到隐含的特征,用于代替传统的手工设计特征,为后续的基于神经网络的文本识别研究打下了基础。Bissaco 等人<sup>[25]</sup>使用 HOG 特征<sup>[26]</sup>代替原始像素训练来深度神经网络将每个单词图像过分割成字符(或部分字符),每个部分使用神经网络进行分类,最后使用集束搜索(Beam Search)<sup>[27]</sup>和强 N 元语言模型找到最佳字符序列。Bai 等人的模型将卷积神经网络与循环神经网络结合<sup>[28]</sup>,实现了卓越的文本识别效果。Yin 等人将卷积神经网络直接与 CTC<sup>[16-17]</sup>结合,避免了循环神经网络带来的梯度消失和梯度爆炸问题<sup>[29]</sup>。Liu 等人<sup>[30]</sup>在识别网络中引入残差结构抑制过拟合。

尽管学界在文本识别领域取得了极大进展,但主流模型都只适用于高质量图片中的英文文本识别,对于电磁泄漏还原图像中的中文文本序列识别问题仍未得到有效解决。而且学界也缺乏相关的研究,Faisel G. Mohammed 等人<sup>[31]</sup>利用惯性矩识别噪声字符,但是只能识别简单背景下的单个英文字符识别。Chanda Thapliyal Nautiyal 等人<sup>[32]</sup>利用感知机识别噪声字符,但也只适用于简单背景下的单个英文字符。Sudipta 等人<sup>[33]</sup>将自动编码器与卷积神经网络结合,实现了对带噪声手写数字图片的识别,然而只适用于单个文字的识别,局限性较大不适用于通用情况下的带噪声中文文本序列的识别。任晓文等人<sup>[34]</sup>将噪声抑制与神经网络相结合识别带噪声手写汉字,但同样局限于简单背景下的单个字符。

对于电磁信号泄露图像中的文本,其具有以下特点:文本在图像中所处区域及背景多变,且文本本身字体形状等特点多变。此外,通过电磁泄漏还原的图像质量通常不高,且包含大量随机噪声,其中以高斯噪声和瑞利噪声为主<sup>[13]</sup>。由于识别模型无法做到像人类视觉系统一样,在部分图像信息被破坏或缺失的情况下依然能不受其影响而完成对目标的

识别。这就导致常规的文本识别模型在低质量图像中的识别难度很大<sup>[14]</sup>。此外, 相比于英文文本, 中文字符种类繁多, 字符的外形更加复杂<sup>[15]</sup>, 对中文字符的判别很大程度上依赖于字符间的细微差别, 这些细节极易受噪声干扰影响模型判别。因此, 对于电磁泄漏还原图像中的中文文本识别具有挑战性和实际意义, 也迫切需要对其进行研究。

因此, 针对电磁泄漏还原图像中的中文文本识别问题, 本文提出了一种利用基于特征强化的识别模型(FENN), 实现对电磁泄漏还原的图像中存在的中文文本进行识别。模型将去噪自编码器(Denoising Autoencoder, DAE)与卷积神经网络(Convolutional Neural Network, CNN)相结合形成特征强化模块, 在训练识别模型的同时训练 DAE 重建模型, 实现对电磁泄漏还原图像中鲁棒特征的提取; 利用循环神经网络(Recurrent Neural Network, RNN)学习上下文信息并使用连续时间序列分类(Connectionist temporal classification, CTC)<sup>[16-17]</sup>实现时序序列标注。此外, 本文还提出了一种新的损失函数, 将 CTC 损失与均方误差重建损失的调和平均值作为联合损失函数实现两部分的端到端训练, 当训练使得联合损失最小时两部分同时达到最优, 此时特征强化模块能够提取最佳鲁棒特征, 同时识别模块的识别率最高, 实现针对电磁泄漏重建图像的精确识别。

本文模型的针对性创新点解决了其他主流识别模型无法抵抗噪声干扰的缺点, 同时本模型不会造成图像像素级别的信息损失, 避免了预处理对图像信息的先验修改, 因此能更好地保留原始信息, 提高识别率。实验表明, 相比于其他主流识别模型, 本文的识别模型能够有效地从电磁泄漏还原图像中学习中文文本特征, 提升模型识别率。

本文第 1 节介绍背景与相关工作; 第 2 节介绍模型的原理与构建; 第 3 节说明实验相关的样本数据; 第 4 节给出实验设计和结果分析; 最后第 5 节对本文进行总结。

## 2 模型

本文针对电磁泄漏还原图像中的中文文本识别问题提出了一种新的模型, 模型包括 2 个主要组成部分: 特征强化模块和时序序列标注模块。首先, 带噪图像输入网络后通过特征强化模块得到区分性的二维特征。然后通过多层双向 LSTM 学习时序特征的前后相关性, 最后利用 CTC 对这些时序特征进行序列标注得到中文类别序列。为了最优化该模型, 将特征强化部分的解码器部分与识别模型分离并通

过均方误差损失评判自编码器的重建效果以此评估其学习到的鲁棒特征, 然后将均方误差损失与 CTC 损失相结合组成联合损失函数, 对模型进行联合训练。当联合损失最小时模型达到最优, 此时特征强化模块能够提取最佳鲁棒特征, 同时序列标注模块的准确率最高, 整个模型达到最优, 如图 1 所示。

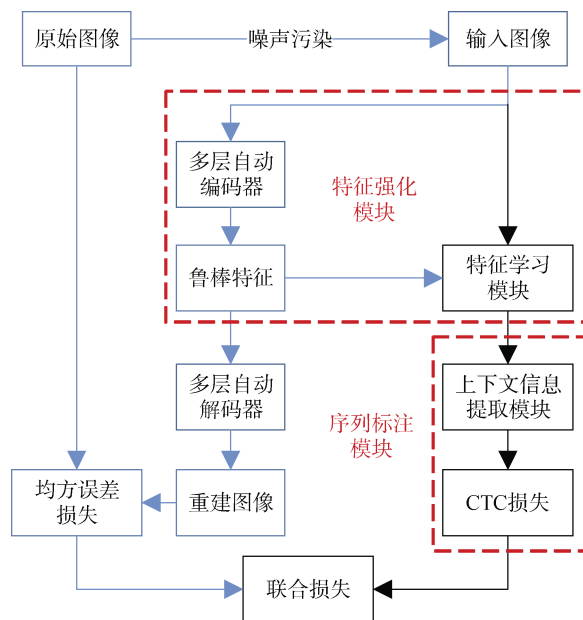


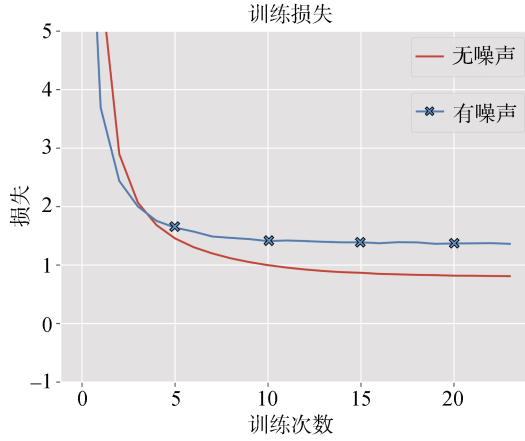
图 1 模型完整架构图

Figure 1 The overall framework of the model

### 2.1 特征强化模块

对于文本识别问题, 主流的模型通常采用卷积神经网络提取图像的二维特征, 这种模型对于高质量的图像是适用的, 但是对于电磁泄漏还原图像这种被噪声污染的低质量图像, 仅使用常规卷积神经网络是不够的。尽管卷积神经网络能够通过大量数据学习其内在特征, 因此对一定程度的数据变化具有适应性, 但需要强调的是这些数据变化不能对区分性的特征信息造成破坏。例如平移、旋转等变化确实从一定程度上丰富了数据量, 但由于没有对目标本身的区分性特征造成破坏, 因此这种数据变化是有益的。然而, 噪声污染对图像中的区分性特征是具有破坏性的, 噪声污染从像素级别改变了图像本身的信息, 使得区分性特征变得不完整。进而增加了卷积神经网络对有效特征的学习难度, 最终影响识别效果。为了直观体现上述问题, 采用对比实验在手写数字集 MNIST 上进行验证, 结果如图 2 所示, 其中红色曲线为卷积神经网络模型在原始 MNIST 数据集上的训练损失, 蓝色曲线为相同模型在被均值为 0 方差为 1 的随机高斯噪声污染的 MNIST 上的训

训练损失。可以看出, 对于被噪声污染的图像, 模型的训练损失明显高于无噪声的情况, 对识别效果造成很大影响。



(注: 损失越低说明模型学习到的特征越有效, 最终识别效果越好)

图2 卷积神经网络在有噪声污染下的训练损失图  
Figure 2 CNN training loss with and without noise

因此, 对电磁泄漏还原图像而言, 最重要的是从中学习到有效特征。实验表明, 利用无监督方式对模型进行预先优化可以最终提升模型的识别性能<sup>[52]</sup>。因此 FENN 引入去噪自编码器, 在使用噪声污染图像和无噪声污染的原始图像训练去噪自编码器的过程中, 以无监督学习的方式最小化重建误差实现对原始无噪声污染图像  $X$  和有效特征  $Y$  之间互信息的最大化, 此时可以认为去噪自编码器学习到了噪声污染图像中的有效特征。然而, 仅依靠去噪自编码器学习的特征是不够的, 去噪自编码器的局限在于其仅对训练过程中出现过的噪声分布表现良好, 这也就意味着, 对于陌生的噪声污染图像, 去噪自编码器不一定能够得到最有效的特征。考虑到上述原因, FENN 将预训练的去噪自编码器和卷积神经网络的特征学习部分相结合, 将二者的损失函数组成联合损失函数并通过优化联合损失函数使得模型能够在鲁棒特征的同时避免原图信息的流失, 以此强化文本特征, 弱化噪声的干扰而非仅利用原始图像特征或仅利用鲁棒特征, 最终有效地从电磁泄漏还原图像中学习文本特征, 提升模型识别率。特征强化模块完整结构如图3所示。

### 2.1.1 特征学习

特征学习部分通过卷积层将输入文本图像转化为具有区分性的二维特征表示, 如图4所示。卷积神经网络是一种层级式神经网络, 通过序列式堆叠多个二维卷积层, 实现对二维图形特征表达能力, 能够拟合复杂和抽象的二维数据模式, 在文字识别,

视觉目标识别等领域获得了极大成功<sup>[35-37]</sup>。

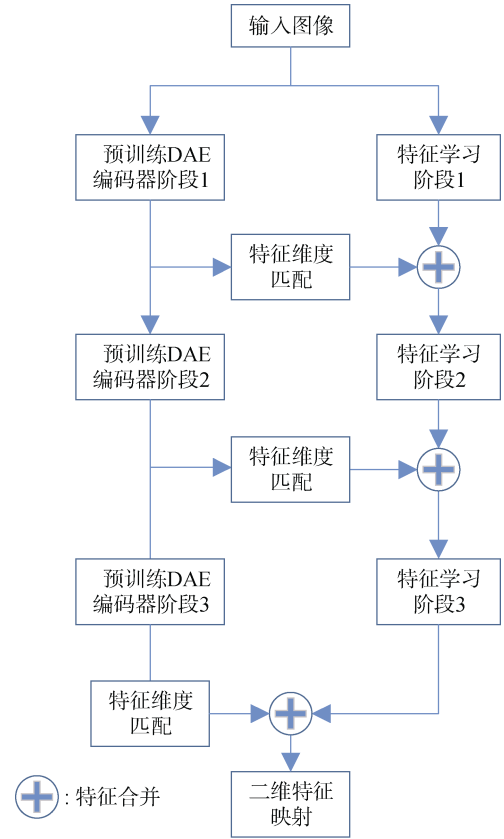


图3 特征强化模块图  
Figure 3 2D feature extracting module

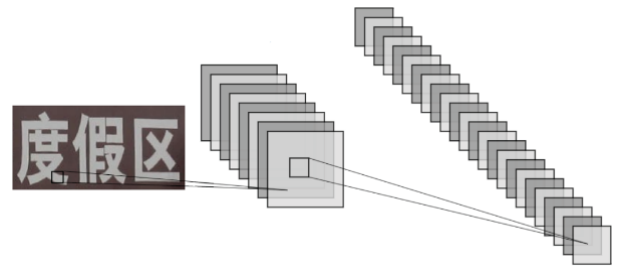


图4 文本图像的二维特征示意图  
Figure 4 2D feature maps of a text image

**定义1.** 卷积层(CONV layer)。对于长度为  $n$  的输入向量  $x = (x_0, x_1, \dots, x_{n-1}) \in \mathbb{R}^n$ , 卷积层对输入向量进行卷积操作得到相同长度的输出向量  $y = (y_0, y_1, \dots, y_{n-1}) \in \mathbb{R}^n$ , 如图5所示。即输出向量的每一维  $y_i$  是对应输入  $x_i$  及其相邻元素的加权和:

$$y_i = \sum_{j \in N_i} \omega_j x_i$$

其中  $N_i$  是  $x_i$  的邻域内全体元素的索引集合,  $\omega_j$  是邻域内元素的权重。



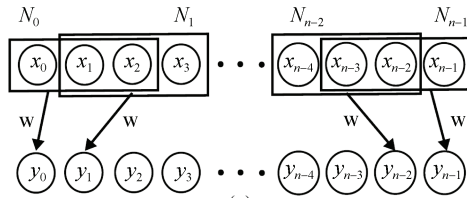


图 5 卷积操作示意图

Figure 5 Concept of convolution

将卷积操作记作“ $*$ ”，并将其扩展至二维空间，则可得到表述二维卷积层的公式：

$$f(x, y) * g(x, y) = \sum_{n_1=-\infty}^{\infty} \sum_{n_2=-\infty}^{\infty} f[n_1, n_2] \cdot g[x - n_1, y - n_2]$$

为应对中文文本识别问题具有字符种类繁多，字体多变且存在大量外形相近的字符等特点。特征学习部分采用较深层的卷积层，以最大限度地学习高层次的抽象特征。特征学习部分的卷积核尺寸均为  $3 \times 3$ ，卷积核数量从浅层的 64 个逐渐增加至深层的 512 个，最终的特征图数量为 512。模型引入池化层 (POOL) 对特征进行下采样以减小维度，同时有助于实现卷积神经网络的局部不变性。

对于噪声污染的图像，模型在判别过程中会由于噪声特征的作用而产生过拟合，使得模型在陌生数据下表现不佳，因此模型在设计时需要尽可能减轻过拟合。模型的特征学习采用序列式堆叠的相同尺寸的卷积层<sup>[38]</sup>，可以有效降低参数总量从而减轻过拟合。例如，连续两层  $3 \times 3$  的卷积核等价于单层  $5 \times 5$  卷积核的覆盖范围，而一个  $5 \times 5$  的卷积核参数总量为 25，而与其等效的两个  $3 \times 3$  的卷积核参数总量只有 18，参数同比减少了 28%。二维特征学习模块部分参数详见表 1，其中 CONV 为卷积层， $n$  为卷

表 1 二维特征学习模块部分参数

Table 1 Configurations of 2D feature learning

模块	类型
输入	图像, 尺寸: $128 \times 32$
特征学习 阶段 1	CONV, $n: 64, k: 3 \times 3$
	POOL, $k: 2 \times 2$
	CONV, $n: 128, k: 3 \times 3$
	CONV, $n: 128, k: 3 \times 3$
特征学习 阶段 2	POOL, $k: 2 \times 2$
	CONV, $n: 256, k: 3 \times 3$
	CONV, $n: 256, k: 3 \times 3$
	CONV, $n: 256, k: 3 \times 3$
特征学习 阶段 3	POOL, $k: 1 \times 2$
	CONV, $n: 512, k: 3 \times 3$
	CONV, $n: 512, k: 3 \times 3$
	CONV, $n: 512, k: 3 \times 3$
特征学习 阶段 3	POOL, $k: 1 \times 2$

积核个数， $k$  为卷积核尺寸，POOL 为池化层，用于对特征图的下采样以降低其维度并实现平移、旋转不变性等特性。

### 2.1.2 去噪自编码器

定义 2. 自动编码器<sup>[53]</sup>。自动编码器分为两部分，编码器将输入向量  $x = (x_0, x_1, \dots, x_{n-1}) \in \mathbb{R}^n$  通过一种映射方式  $y = f_{\theta}(x) = s(Wx + b)$ ，其中  $\theta = \{W, b\}$  映射到一个隐藏层，得到表征向量  $y = (y_0, y_1, \dots, y_{n-1}) \in \mathbb{R}^n$ ，解码器将表征向量通过映射  $z = g_{\theta'}(y) = s(W'y + b')$ ，其中  $\theta' = \{W', b'\}$  映射为重建向量  $z = (z_0, z_1, \dots, z_{n-1}) \in \mathbb{R}^n$ 。模型通过最小化平均重建误差进行最优优化：

$$\theta^*, \theta'^* = \arg \min_{\theta, \theta'} \frac{1}{n} \sum_{i=1}^n L(x^{(i)}, z^{(i)})$$

其中  $L$  是损失函数如均方误差等。对于自动编码器，可以通过有目的地训练可以使其具有图像去噪与重建的功能，即模型在输入图像部分信息被破坏的情况下尽可能使输出图像与未被破坏的原始图像接近，也就是去噪自编码器 (DAE)，其结构如图 6 所示。

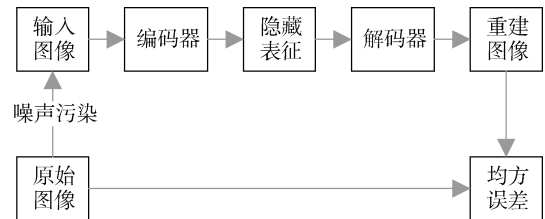


图 6 去噪自编码器结构图

Figure 6 The architecture of Denoising Autoencoder (DAE)

去噪自编码器实现过程如下，假设现有输入图像  $x$ ， $x$  被噪声污染后得到  $\tilde{x}$ ，即  $x$  中的对应信息在  $\tilde{x}$  中被修改，自编码器需要在训练过程中自动将这些被修改的信息还原。这样， $\tilde{x}$  通过自编码器的映射得到隐藏层表征  $y = f_{\theta}(\tilde{x}) = s(W\tilde{x} + b)$ ，通过该表征可以得到重建图像  $z = g_{\theta'}(y) = s(W'y + b')$ 。

特征强化部分中所用的去噪自编码器由编码器和解码器两部分构成，其中编码器部分包含三个阶段，这三个阶段得到的特征与特征学习部分对应阶段的输出特征进行合并。具体参数见表 2，其中  $n$  为卷积核个数， $k$  为模板尺寸。

表 2 去噪自编码器参数  
Table 2 Configurations of DAE

模块	类型
输入	图像, 尺寸:128×32
编码器 阶段 1	CONV, $n:64, k:3 \times 3$
	CONV, $n:64, k:3 \times 3$
	POOL, $k:2 \times 2$
编码器 阶段 2	CONV, $n:128, k:3 \times 3$
	CONV, $n:128, k:3 \times 3$
	POOL, $k:2 \times 2$
编码器 阶段 3	CONV, $n:256, k:3 \times 3$
	CONV, $n:256, k:3 \times 3$
	CONV, $n:256, k:3 \times 3$
	POOL, $k:2 \times 2$
解码器	UP-CONV, $n:256, k:3 \times 3$
	CONV, $n:256, k:3 \times 3$
	CONV, $n:256, k:3 \times 3$
	CONV, $n:256, k:3 \times 3$
	UP-CONV, $n:256, k:3 \times 3$
	CONV, $n:128, k:3 \times 3$
	CONV, $n:128, k:3 \times 3$
	UP-CONV, $n:256, k:3 \times 3$
	CONV, $n:64, k:3 \times 3$
	CONV, $n:64, k:3 \times 3$

UP-CONV 为上采样层, 卷积核通过设置不同填充和卷积步长实现将小尺寸特征恢复成原始尺寸, 解码器将上采样层与卷积层结合, 将编码器的输出特征还原无噪声污染的原始图像, 如图 7 所示。

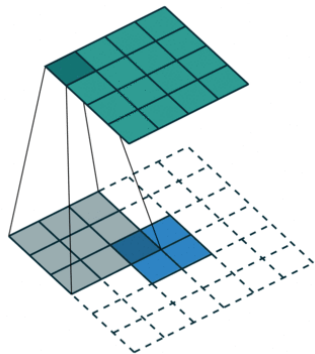


图 7 UP-CONV 层上采样操作示意图  
Figure7 Upsampling of UP-CONV layer

其中, 蓝色像素为编码器部分得到的特征, 绿色像素为上采样结果。将特征像素之间分隔开, 使得卷积核的步长得以增加以实现上采样的效果。当去噪自编码器训练完成后, 可以将其编码部分单独取出作为特征提取模块, 相比于常规卷积神经网络, 去噪自编码器从噪声污染图像提取的特征更鲁棒, 受信息破坏的影响更小。因此, FENN 将去噪自编码器的编码部分融入识别模型, 在保留输入图像全部信息的同时尽可能提取文本本身的特征而抵御噪声

的影响。  
2.1.3 特征强化

将特征学习部分和去噪自编码器结合得到完整的特征强化模块。特征强化模块背后的逻辑在于, 电磁泄漏还原的文本图像中既包含文本信息同时也包含噪声信息, 文本内容与噪声特征是无关系的。如果仅使用常规卷积神经网络提取二维特征, 则模型在训练过程中会同时考虑文本和噪声信息, 导致文本和噪声的特征混杂, 阻碍模型学习到区分性的文本特征, 影响识别精度。实际测试中发现, 对于噪声污染的图像, 模型极易出现过拟合, 原因是模型在特征学习过程中将与文本信息无关的噪声特征也作为判别的依据, 导致卷积特征不够鲁棒。因此, 模型需要尽可能抵御噪声对特征的影响, 提取更加鲁棒, 更加具有区分性的文本特征。

然而, 对于常规的基于卷积神经网络的识别模型而言, 上述功能显然是难以控制的。因此, 本文不只依靠卷积神经网络自身学习图像特征, 而是通过在模型上增加去噪自编码器实现鲁棒特征提取。通过引入特征强化部分, 在模型内部强化文本特征, 使得识别模型在判别时能够不受噪声影响。

从图 3 结构上看, 特征学习仍包含原始图像中的完整特征信息, 不会导致原始特征信息的更改或丢失, 特征强化部分包含抑制噪声后的鲁棒文本特征信息。因此, 完整的识别模型同时根据特征强化部分中的鲁棒特征和卷积主干中的原始特征进行判别, 使得完整模型能够更加充分有效地从噪声污染的图像中学习到文本的区分性特征, 提升识别精度。此外, 将特征强化部分融入识别模型内部能够提升模型的整体性, 识别过程是端到端的, 无需经过任何单独的去噪等预处理, 完全由模型自主学习全部特征。

特征强化部分由去噪自编码器的编码模块组成。输入图像共经过序列式 3 级去噪自编码器。特征强化部分与特征学习部分的连接方式采用简化的密集连接<sup>[39]</sup>, 使得去噪自编码器中各阶段的鲁棒特征与特征学习模块各阶段的常规特征产生联系, 有助于避免梯度消失问题。预训练去噪自编码器各阶段的输出特征经过后续的特征维度匹配得到与特征学习部分的输出特征相同维度的特征, 然后沿通道方向进行特征合并之后, 特征图组中包含特征强化部分各级输出的多级别鲁棒特征, 增加后续层输入数据的多样性, 提升模型效率。其中浅层去噪自编码器的输出特征对原图中的信息保留更多, 在卷积部分中会流经更深的卷积层, 最终的得到特征抽象程度高, 在判别时的比重高。特征强化模块的完整数据

流如下, 图像经过去噪自编码器和下采样后的输出分为两路, 一路经过特征维度匹配与特征学习部分的特征合并, 然后进行后续的特征学习; 另一路送入后续的去噪自编码器进一步得到更加鲁棒的二维特征, 然后与特征学习部分中的特征学习模块的输出特征沿通道维度进行合并(channel-wise concat), 如图 8 所示。

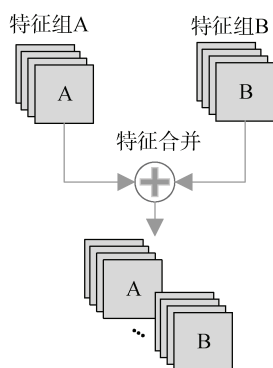


图 8 特征合并示意图

Figure 8 Channel-wise concatenation of features

#### 2.1.4 二维特征映射

由于 RNN 无法直接处理二维特征, 因此卷积特征在送入上下文信息提取模块之前需要映射成一维的特征序列。图像局部区域的特征提取由卷积模块、下采样模块和激活函数共同实现, 特征学习部分最终的特征图具有平移不变性, 因此可以将卷积层得到的这些特征图组进行维度重置。特征图组中的每一列及其在通道维度的延伸对应原始输入图像中的一个矩形区域, 这些矩形区域的排列方式和特征图组中对应列的排列方式是一致的, 即从左向右依次排列。映射操作沿卷积特征图组的通道维度, 将每张特征图的第  $t$  列取出后组成一个二维特征向量并展开, 然后以首尾相接的形式组成一维特征序列送入 RNN, 每个特征向量都与一个感受野相联系, 可以描述对应区域的二维特征, 如图 9 所示。

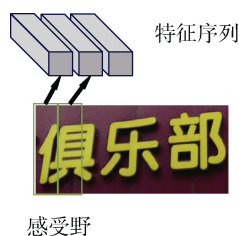


图 9 特征序列与感受野示意图

Figure 9 Feature sequence and receptive field

特征学习部分最后一个卷积层输出特征图组的

维度为  $H \times W \times D$ 。其中,  $H, W, D$  分别为高度、宽度和深度。映射后得到  $W$  个特征序列对应原特征图的宽度  $W$ , 作为上下文信息序列  $c = (c_1, c_2, \dots, c_W)$  送入循环部分。

## 2.2 序列标注模块

输入图像经过特征强化模块提取特征得到了时序特征序列, 于是文本图像识别问题就转化为了序列标注问题。与英文不同, 中文文本没有单词级别的显式分词, 导致检测阶段得到的文本行普遍较长, 且前后信息关系紧密。因此, 需要使用 RNN 学习时序特征序列中的上下文信息。RNN 通过内部自连接的隐藏层保留过往信息, 有效地提升了中文文本识别的准确率。

### 2.2.1 LSTM

模型使用长-短时记忆体(Long-Short Term Memory, LSTM)实现上下文特征提取。

LSTM<sup>[40]</sup>是 RNN 的一个变种, 通过清除或添加内部存储单元中的信息, 避免模型过分参考小范围内的上下文信息, 使其充分学习到序列中的长期依赖关系<sup>[41-42]</sup>。LSTM 具有方向性, 单向的 LSTM 只能利用一个方向的信息, 但是对于电磁泄漏还原图像中的中文文本, 上下文两个方向的信息都有作用, 需要同时考虑。因此, 将两个方向相反的 LSTM 结合在一起形成双向结构, 如图 11 所示。正向结构学习下文信息, 反向结构学习上文信息, 这种结构非常适合于序列识别任务<sup>[43]</sup>。

对于被噪声污染而难以从二维特征判别的字符, 考虑其上下文能够辅助判别。这里的上下文既包括像素级别的上下文, 也包括字符级别的上下文。从像

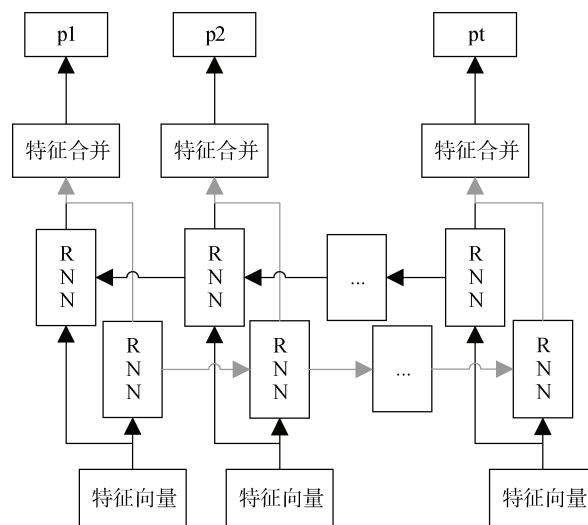


图 10 双向 RNN 结构图

Figure 10 Bilateral RNN



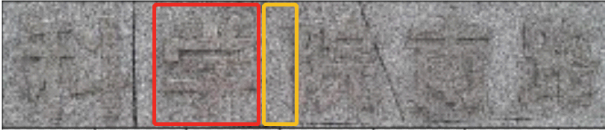


图 11 噪声污染文本图像中的分割情况示意图

Figure 11 Segmentation in noisy text image

素级别来看,考虑到中文字符的形状多变,尺寸不固定的特定,通常需结合多个连续像素列的信息才能充分描述。因此,当字符的某一部分因为噪声污染而难以识别时,可以根据字符结构的上下文特征进行判断。从语义级别来看,利用前后字符之间包含的上下文信息同样可以帮助判别。例如,对于“中央\*视台”(“\*”代表被噪声污染而无法识别的字符),如果考虑上下文信息则该字符有较高概率被预测为“电”。相比于每个字符单独识别,这种考虑序列上下文的方式更适合于文本序列中出现字符因噪声污染而难以识别的情况。

### 2.2.2 时序对齐

在电磁泄漏还原图像中,各字符的间隔区域不明显,导致字符对齐困难,如图 12 所示,其中红色框为字符区域,黄色框为字符间隙区域。因此,翻译模块采用 CTC 解码器,以无需对齐的方式实现对文本序列的直接判别,避免了噪声对分割过程的干扰。CTC<sup>[44]</sup>将特征输出转化为所有可能的类别序列的概率分布,然后最大化正确类别标注的概率。CTC 在编解码过程中引入占位符,使得模型无需对文本序列进行分割,也不限定文本序列在图像中的具体位置,避免了由于噪声干扰导致的字符分割困难。对于宽度为  $W$  的卷积特征图,特征序列  $\mathbf{x} = x_1, \dots, x_W$  共有  $W$  个时刻,每一个时刻  $t$  对应一个 softmax 输出的类别概率分布。令  $C$  为所有字符类别集合,则 softmax 的输出类别集合为  $C' = C \cup \{\text{占位符}\}$ 。对于所有时刻的特征输出矩阵中,其中横轴  $t$  为时序方向,纵轴  $s$

为每一帧的类别分布,包括占位符。特征序列和目标序列之间的对齐方式不唯一,CTC 会尝试所有可能的对齐方式,然后将每种情况的概率相加并选择概率最大的路径。

在得到输出路径后,CTC 通过一种多对一的映射方式  $\gamma$  将逐帧序列解码为最终的类别序列。解码时  $\gamma$  合并无占位符分割的重复字符并去除占位符。那么令  $A$  为  $\gamma$  中包含所有正确解码路径的子集,则输出目标序列  $Y$  的概率为  $A$  中所有路径的概率之和:

$$p(Y|x) = \sum_{\forall \pi \in A} p(\pi|x)$$

对于  $A$  中的每条路径  $\pi$ ,其概率计算如下:令  $y$  为 RNN 的输出序列也就是 CTC 的输入序列,则  $p(t, \pi_t)$  为  $y$  的第  $t$  帧判别为路径  $\pi$  中第  $t$  项的概率,则路径  $\pi$  的概率为:

$$p(\pi|x) = \prod_{t=1}^W p(t, \pi_t), \forall \pi \in C'$$

其中  $\pi_t$  为路径  $\pi$  第  $t$  帧的类别。对于不包含候选词典的中文文本识别任务。根据上述公式,CTC 在寻找概率最大路径时可以使用贪婪算法<sup>[45]</sup>,即认为每一个时刻最大概率的类别组成的路径即为最优路径:

$$\beta(x) \approx \gamma(\pi')$$

其中,  $\pi' = \arg \max_{\pi} p(\pi|x)$

## 2.3 模型训练

### 模型最优化

模型最优化过程就是要找到一组参数使得对于给定的输入样本,模型输出正确标注序列的概率最大,因此可以采用极大似然参数估计方法将模型的最优化问题转化为最小化损失函数。由于模型的文本识别模块和去噪自编码器采用不同的损失函数进行评判,因此联合训练时需要将二者结合。对于识别模块,沿用 2.2.2 小节中的符号,CTC 损失的定义为样本的正确标注的负对数概率:

$$L_1(x, z) = -\ln p(z|x)$$

对于去噪自编码器,采用均方误差损失函数:

$$L_2(x, z) = \frac{1}{n} \sum_{i=1}^n (x_i - z_i)^2$$

为了解决损失函数一致性的问题,对上述两损失函数求调和平均得到联合损失,因此模型训练的目标就是最小化下述联合损失函数:

$$L(x, z) = \frac{1}{\frac{1}{L_1(x, z)} + \frac{1}{L_2(x, z)}}$$



图 12 RCTW17 样本图像

Figure 12 Samples from RCTW17



在训练中优化联合损失的好处在于, 特征强化模块的编码器部分将特征输入识别模块, 因此识别模块的 CTC 损失函数会从时序序列标注的角度对特征强化模块得到的鲁棒特征进行间接评估, 这样相比于仅通过均方误差损失朴素地优化去噪自编码器更能适用于文本识别问题。在反向传播过程中, 特征强化模块的编码器同时考虑均方误差重建损失和 CTC 损失, 使得模型的优化方向更加明确, 鲁棒特征的针对性也更强, 而解码器部分不参与 CTC 损失评判, 从一定程度上降低了过拟合。当且仅当联合损失最小时两部分同时达到最优, 此时特征强化模块能够提取最佳鲁棒特征, 同时识别模块的识别率最高, 整个模型达到最优。



图 13 CASIA-10k 样本图像

Figure13 sample images from CASIA-10k dataset

### 3 样本数据

#### 3.1 噪声模型

电磁泄漏还原图像中的噪声以高斯噪声和瑞利噪声占主导性。其他类型噪声由于分布较少且对图像影响甚微, 这里不做讨论。因此, 实验选择加性高斯白噪声和瑞利噪声作为噪声源代表。

高斯噪声常见于信号放大原件或信号探测设备, 由原子的热运动和物体的热辐射导致<sup>[7]</sup>。高斯噪声的概率密度函数公式如下。

$$P(g) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{(g-\mu)^2}{2\sigma^2}}$$

其中,  $g$  是像素值,  $\sigma$  是标准差,  $\mu$  是均值。概率密度分布瑞利噪声通常在电磁信道传输的图像中出现, 如雷达探测图像等<sup>[7]</sup>。其概率密度函数公式如下。

$$P(g) = \begin{cases} \frac{2}{b}(g-a)e^{-\frac{(g-a)^2}{b}}, & g \geq a \\ 0, & g < a \end{cases}$$

其中,  $g$  为像素值, 均值  $\mu = a + \sqrt{\frac{\pi b}{4}}$ , 方差  $\sigma^2 = \frac{b(4-\pi)}{4}$ 。

#### 3.2 数据集

为弥补公开数据集样本数量的不足, 构建合成中文图像数据集用于模型预训练阶段, 然后从标准数据集 RCTW17 和 CASIA-10k 的训练集以及电磁泄漏还原图像实景数据集中截取样本图像中的文本区域并添加噪声作为微调训练阶段所需的数据集。实验验证时, 从 RCTW17 和 CASIA-10k 的测试集中根据标注数据截取样本图像中的文本区域并添加随机噪声与电磁泄漏还原图像实景数据共同作为测试集。

##### 3.2.1 标准数据集

**RCTW17.** 该数据集来自 ICDAR 2017 Competition on Reading Chinese in the Wild 竞赛<sup>[46]</sup>, 是一个大规模自然场景数据集, 包括街景, 海报, 菜单, 室内场景, 截屏等, 数据集共包含 12 263 张的中文文本图像, 其中 8346 张为训练图像, 4229 张为测试图像。图像经过详细标注, 标注内容包括文本区域的 4 个顶点坐标以及区域中的文本内容, 每张图像都至少包含一行中文文本, 评估阶段以文本行为单位进行识别, 如图 14 所示。

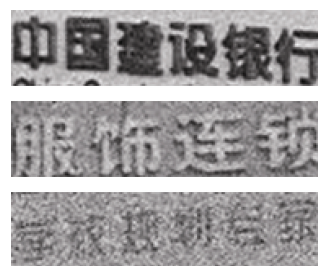


图 14 低中高 3 种噪声强度下的样本图像

Figure14 Noisy images corrupted by low, medium and high level noise

**CASIA-10K.** 该数据集为中科院自动化所 PAL 团队提出的中文场景文本数据集<sup>[9,47]</sup>, 包含 10000 张不同场景下的图像, 其中 7000 张为训练样本, 3000 张为测试样本。每张图片的标注内容为文本区域的 4 个顶点坐标以及文本区域的内容。评估阶段以文本行为单位进行识别, 如图 15 所示。

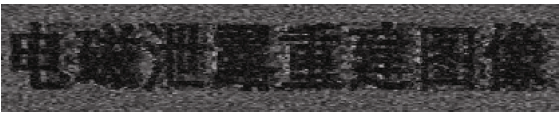


图 15 电磁泄漏还原图像  
Figure 15 Reconstructed image from electromagnetic emission

3.2.2 合成数据集

由于公开数据集中的样本数量有限, 不足以训练出可用的中文识别模型。因此, 首先使用文本图像合成引擎<sup>[48]</sup>构建中文文本合成图像用于预训练阶段, 待模型在合成数据集上收敛之后, 再迁移到目标数据集进行微调训练(fine-tuning)。合成数据集包含 100 万个样本, 每张样本图像包含一个文本行, 文本行长度从 1 字符到 10 字符不等, 文本内容来自人民日报 2014 年语料库。每张图像中文本所使用的字体从 46 种中文字体中随机选择。字符种类为 5020, 其中包括 10 个数字, 10 个常用标点符号, 以及 5000 个常用汉字。常用字的选择根据现代汉语语料库汉字频率表从高到低排列, 到第 5000 个常用字为止其累计使用频率为 99.95, 足以涵盖日常场景中的绝大部分中文字符<sup>[49]</sup>。

3.2.3 样本加噪

构建图像数据集时, 首先根据标注文件中的文本区坐标从公开数据集 RCTW17 和 CASIA-10k 中截取水平方向的文本区域并进行灰度化处理, 然后添加高斯噪声和瑞利噪声对图像进行污染以模拟电磁泄漏还原的情况, 加噪后的样本分布情况见表 3, 加噪后的样本如图 16 所示。对于加噪后的样本, 使用峰值信噪比(psnr)和结构相似性(ssim)进行评价。

表 3 加噪训练样本分布表  
Table3 Distribution of noisy training samples

	Noise ranges			
	Gaussian		Rayleigh	
	psnr(dB)	ssim	psnr(dB)	ssim
Max	14.76	0.28	13.8	0.27
Min	22.03	0.81	21.40	0.79

3.2.4 实景测试集

针对显示信号电磁泄漏还原实战场景, 本文采用如下攻击场景, 从 RCTW17 和 CASIA-10K 数据集的训练集和测试集中挑选中带有文本的图像信息在目标计算机上显示, 目标计算机放置在离地 1m 的平台上。将电磁泄漏信号接收天线放置于离地高度 1m, 距离目标计算机从最近 0.5m 到最远 3m 的

位置, 每个位置间距长度为 0.5m, 在每个距离上采集 50 张不同的还原图像, 如图 17 所示。共采集 300 张作为实景数据集并计算平均峰值信噪比与结构相似度, 详细参数见表 4, 其中 70%作为训练集, 30%作为测试集。

表 4 电磁还原图像样本分布表  
Table 4 Distribution of reconstructed image samples

Range	Average PSNR	Average SSIM
0.5	21.0	0.74
1	19.1	0.66
1.5	17.3	0.55
2	15.4	0.46
2.5	13.6	0.36
3	11.7	0.27

4 实验

实验流程如下, 首先, 使用原始未加噪图像集与加噪图像集对去噪自编码器进行预训练, 然后在构建识别模型时使用预训练的去噪自编码器中的编码器模块构建特征强化部分。接着将识别模型在加噪合成样本训练集上进行预训练, 待收敛后迁移至目标数据集进行微调(fine-tuning)至再次收敛, 然后在测试集上进行测试, 对比不同模型的结果。对照组模型与本文模型的训练过程保持一致。

4.1 实验细节

实验环境为 Ubuntu 14.04 系统工作站, 处理器为 Intel Xeon E5-2620 v4 @ 2.10GHz, 内存容量为 64GB, 显卡型号为 NVIDIA Tesla K40c。实验所用的程序均使用 Python 语言在 TensorFlow 环境下实现。对于去噪自编码器部分, 训练时, 模型首先在 100 万张合成训练集上进行训练, 待模型收敛后再迁移到目标训练集进行微调(fine-tuning)直到再次收敛。然后在训练完整识别模型时使用上述预训练去噪自编码器的编码器部分, 同样在先 100 万张合成训练集上进行训练, 待模型收敛后再迁移到目标数据集进行微调(fine-tuning)直到再次收敛。收敛判断标准为精确率在连续 5 个训练周期内增长不超过 0.01%。训练过程共耗时约 50 h。为了加速训练, 训练过程中所有样本均缩放到 128×32, 反向传播过程使用 Adam 优化器<sup>[50]</sup>, 学习率设置为 0.0001。模型中的所有层选择“Xavier”初始化<sup>[51]</sup>, 批大小为 128。完整网络的训练环境等设置与去噪自编码器部分保持一致, 训练耗时约 80 h。

4.2 实验结果

实验结果使用精确率和编辑距离对模型进行

评估。

**定义 3. 精确率。**指模型正确判别的样本在全部测试样本中所占的比例。

**定义 4. 编辑距离。**对于字符串  $s1, s2$ , 他们的编辑距离是将  $s1$  转换成  $s2$  所需的最小编辑操作数。编辑操作包括: 插入字符, 删除字符和替换字符。

**算法 1.** 编辑距离( $s1, s2$ )

输入:

( $s1, s2$ ): 用于计算编辑距离的两字符串

输出:

$m[|s1|, |s2|]$ : 字符串  $s1$  和  $s2$  的编辑距离

```
1 int  $m[i, j] = 0$ 
2 FOR  $i \leftarrow 1$  TO  $|s1|$ 
3 DO  $m[i, 0] = i$ 
4 FOR  $j \leftarrow 1$  TO  $|s2|$ 
5 DO  $m[0, j] = j$ 
6 FOR  $i \leftarrow 1$  TO  $|s1|$ 
7 DO FOR  $j \leftarrow 1$  TO  $|s2|$ 
8 DO IF  $s1[i] == s2[j]$ 
9 THEN  $m[i, j] = \min\{$ 
10  $m[i-1, j]+1,$ 
11  $m[i, j-1]+1,$ 
12  $m[i-1, j-1]\}$ 
13 ELSE  $m[i, j] = \min\{$ 
14  $m[i-1, j]+1,$ 
15  $m[i, j-1]+1,$ 
16  $m[i-1, j-1]+1\}$ 
17 RETURN  $m[|s1|, |s2|]$ 
```

### 结果对比

实验在第 3.2 节所述的 3 个数据集上进行, 将

本文模型与其他具有代表性的主流识别模型进行横向对比, 包括 Bai<sup>[12]</sup>, Yin<sup>[13]</sup>和 Liu<sup>[14]</sup>的模型。作为对照组的 Bai<sup>[12]</sup>, Yin<sup>[13]</sup>和 Liu<sup>[14]</sup>的模型均采用与本模型一致的训练数据与训练方法, 评价指标为精确率 (Precision) 和归一化平均编辑距离 (Normalized Average Edit Distance, NAED)。测试时, 使用前述加噪后的公开数据集 RCTW17 和 CASIA-10k 中的测试集。

CASIA-10k 测试集在低强度噪声下平均 psnr 约为 19.1dB, 平均 ssim 约为 0.7, 在中强度噪声下平均 psnr 约为 15.7 dB, 平均 ssim 约为 0.5, 高强度噪声下平均 psnr 约为 13.0dB, 平均 ssim 约为 0.3。RCTW17 测试集在低强度噪声下平均 psnr 约为 21dB, 平均 ssim 约为 0.7, 在中强度噪声下平均 psnr 约为 17.5 dB, 平均 ssim 约为 0.5, 高强度噪声下平均 psnr 约为 14.2 dB, 平均 ssim 约为 0.3。表 2 和表 3 展示了不同模型在低, 中, 高三种强度的噪声的图像中的测试结果的精确率和归一化平均编辑距离。

从表 5 可以看出, 对于 CASIA-10k 测试集, 主流识别模型中, Bai 的模型效果最差, 原因主要在于其卷积结构较浅无法充分学习中文字符特征, 且模型无法应对噪声的干扰。此外, 结构中未采用 Dropout 层以及后续的 Bi-RNN 深度较深导致了过拟合。Liu 的模型卷积结构较深并且引入了残差结构, 因此对字符结构特征的学习能力相比 Bai 的模型有所提升。此外, 模型中卷积结构更深且引入了大量的 Dropout 层, 一定程度上抑制了过拟合, 因此性能超过了 Bai 和 Liu 的模型, 但仍无法有效应对噪声的干扰。

表 5 在 CASIA-10k 数据集上的结果对比  
Table 5 result comparison on CASIA-10k dataset

	精确率(%) / 归一化平均编辑距离					
	高斯噪声			瑞利噪声		
	低	中	高	低	中	高
	psnr=19.1 ssim=0.7	psnr=15.7 ssim=0.5	psnr=13.1 ssim=0.3	psnr=19.2 ssim=0.7	psnr=15.8 ssim=0.5	psnr=12.9 ssim=0.3
Bai et al. <sup>[12]</sup>	38.4/0.40	29.1/0.53	11.7/0.79	38.9/0.42	29.5/0.52	13.2/0.70
Liu et al. <sup>[14]</sup>	40.5/0.36	31.6/0.46	13.4/0.70	40.7/0.37	32.4/0.46	16.3/0.74
Yin et al. <sup>[13]</sup>	44.8/0.37	33.3/0.50	12.6/0.73	44.3/0.38	34.2/0.49	15.2/0.74
Ours	45.0/0.34	36.2/0.42	16.1/0.66	45.3/0.33	36.9/0.42	19.1/0.62

FENN 表现最佳, 得益于特征强化模块有效地学习到鲁棒的文本特征, 避免了噪声对特征学习的影响, 使得模型能够有效地学习区分性的中文字符特征。在低强度的高斯和瑞利噪声情况下, 尽管 FENN 性能有所提升但相比于第二名 Yin 的模型优势

不大, Precision/NAED 平均提升了 0.6%/0.04, 这是因为在低噪时, 相比于噪声信息, 图像中的有效文本信息占比较高, 特征强化模块抑制的噪声特征所占的比例较低, 导致特征强化模块的作用不明显, 因此模型性能提升有限。但随着噪声强度的增大, 本文

模型的优势逐渐明显。在中强度噪声下, Yin 的模型精确率高于 Liu 的模型但编辑距离劣于 Liu。因此, 将 FENN 与 Yin 的精确率和 Liu 的编辑距离进行对比, Precision/NAED 平均提升 2.8%/0.04。在高噪声强度下, Liu 的模型从精确率和编辑距离两个方面都超过了 Yin 的模型成为第二名, 此时 FENN 相比于第二名的 Precision/NAED 平均提升为 3.2%/0.08。

从表 6 可以看出, 对于 RCTW17 测试集, 各模型测试结果的相对关系与在 CASIA-10k 上大致相同, 在低强度噪声下, FENN 的精确率略低于 Yin 的模型但编辑距离更优, 说明尽管完全正确预测的样本略

少于 Yin 的模型, 但就编辑距离而言, 出现错字的比例更低。在中强度下, Yin 的模型精确率高于 Liu 的模型但编辑距离劣于 Liu。因此, 将本文模型与 Yin 的精确率和 Liu 的编辑距离进行对比, 本文模型的 Precision/NAED 平均提升为 0.65%/0.065。在高强度高斯噪声下, Liu 的模型在精确率和编辑距离两方面都位居第二, 本文模型与之相比 Precision/NAED 平均提升为 2.2%/0.03。在噪高强度瑞利噪声下, Yin 的模型的精确率高于 Liu 但编辑距离劣于 Liu。因此, 将本文模型与 Yin 的精确率和 Liu 的编辑距离进行对比, 本文模型的 Precision/NAED 平均提升为 1.2%/0.02。

表 6 在 RCTW17 数据集上的结果对比  
Table 6 result comparison on RCTW17 dataset

	精确率(%) / 归一化平均编辑距离					
	高斯噪声			瑞利噪声		
	低	中	高	低	中	高
	psnr=20.7 ssim=0.7	psnr=17.6 ssim=0.5	psnr=14.3 ssim=0.3	psnr=21.0 ssim=0.7	psnr=17.3 ssim=0.5	psnr=14.3 ssim=0.3
Bai et al. [12]	36.8/0.43	29.7/0.51	18.1/0.68	35.4/0.46	33.4/0.50	21.0/0.63
Liu et al. [14]	38.3/0.40	33.5/0.44	21.2/0.60	37.9/0.41	34.4/0.48	23.5/0.60
Yin et al. [13]	41.1/0.38	36.1/0.47	20.7/0.62	40.9/0.39	36.2/0.44	23.9/0.57
Ours	41.0/0.36	36.9/0.41	23.4/0.57	40.4/0.37	36.7/0.42	25.1/0.55

从表 7 可以看出, 对于实际电磁泄漏场景下的图像集, 在低噪污染下, 相比于 Bai、Liu 和 Yin 的模型, FENN 在 Precision/NAED 方面分别提升了 5.4%/0.06, 3.5%/0.05 和 0.6%/0.02。在中噪污染下, 本文模型相比于 Bai、Liu 和 Yin 在 Precision/NAED 方面分别提升 6.5%/0.11, 3.1%/0.05 和 0.7%/0.07。在高噪污染下, FENN 相比于 Bai、Liu 和 Yin 的模型在 Precision/NAED 方面分别提升 4.8%/0.1, 2.0%/0.03 和 2.5%/0.05。上述结果表明 FENN 能够有效应对电磁泄漏还原图像中的噪声干扰对其中的中文文本进行识别, 证明了模型的有效性。

表 7 在电磁泄漏还原图像集上的结果对比  
Table 7 Result comparison of different methods on reconstructed image samples

	Noise Level		
	low	medium	high
	psnr=20.0 ssim=0.71	psnr=16.4 ssim=0.45	psnr=12.1 ssim=0.28
Bai et al. [12]	34.6/0.45	26.7/0.56	16.3/0.73
Liu et al. [14]	36.5/0.44	30.1/0.50	19.1/0.66
Yin et al. [13]	39.4/0.41	32.5/0.52	18.6/0.68
Ours	40.0/0.39	33.2/0.45	21.1/0.63

此外, 各模型在实际泄漏数据集上的性能表现与在加噪公开数据集上大致相同, 说明合成数据集以及通过样本加噪模拟实际场景的方法是有效的, 训练数据符合实际数据分布。

表 8 展示了本文模型和其他 3 种被测模型对测试样本的识别情况。由图中结果可见, 对于清晰度较高且长度较短的中文文本, 4 种模型都可以正确识别。随着文本长度以及噪声污染的增加, 本文模型在识别率和编辑距离上都具有明显优势。然而, 对于文本中的标点, 本文模型存在漏检的情况, 这是因为标点尺寸较小, 受噪声污染影响大, 在 CTC 解码时被认为是文字间隔区而导致丢失。此外, 对于文本内容受到部分前景遮挡时, 4 种识别模型的结果都不理想, 说明对于高难度文本图像, 目前主流的识别模型还有待提升。

综合上述对比结果, 本文模型在从低噪到高噪条件下都比主流识别模型有着更好的表现, 体现了模型的泛用性。随着噪声强度的增加, 模型借助特征强化部分对于鲁棒特征的提取, 最大程度地降低了噪声对图像特征破坏的影响, 充分说明了该模型在解决电磁泄漏还原图像中的中文文本识别问题时的有效性。



表 8 不同模型对电磁泄漏还原图像的识别结果对比

Table 8 Comparison of recognition results between different methods

电磁泄漏还原图像	Bai	Liu	Yin	Ours
	健康农业	健康农业	健康农业	健康农业
	型车辆驶入请绕	型车辆驶入请绕	型车辆驶入请绕	型车辆驶入请绕
	门助银行服务	自助银行服务	中助银行服务	自助银行服务
	消随全请文明经市	创国会国文明热术	创镀金锁文明储寅	尔到金国文明城市
	电池回收	电池回收	电池回收	电池回收
	清将烟头灭后人桶	请将烟头灭后人桶	清将烟头灭后入桶	清将烟头灭后入桶
	推歌铜登	雅新钢景速	雅新钢昌城	雅新钢器璃
	禁止鸣笛	禁止鸣笛	禁止鸣笛	禁止鸣笛

5 结论

本文针对电磁泄漏还原图像的特点提出了一种本识别而无需去噪等常规预处理。模型能够在不进行常规去噪等预处理的情况下直接对带噪中文文本图像实现无分割识别。在电磁泄漏实景数据以及公开数据集RCTW17和CASIA-10K上的测试结果表明, 相比与其他主流识别模型, FENN 在电磁泄漏还原图像中的中文识别率最高提升 5.4%, 体现出明显优势。对比实景数据集与公开数据集可以发现, 尽管公开数据集对比结果表明方法虽然更好, 但提高不太明显, 这是因为实景数据中包含大量难以建模的次要噪声模型, 这些噪声对模型的识别率造成了影响。由模型结构和实际数据测试结果可知, 模型的识别效果部分程度上局限于去噪自编码器对图像噪声模型的拟合能力, 如果噪声模型过于复杂超出去噪自编码器的拟合能力, 以至于无法达到良好的重建效果, 则其编码器部分提取的特征就不够充分, 最终的识别率也会受到影响。

参考文献

[1] van Eck W. Electromagnetic Radiation from Video Display Units: An Eavesdropping Risk[J]. *Computers & Security*, 1985, 4(4): 269-286.

[2] Elibol F, Sarac U, Erer I. Realistic Eavesdropping Attacks on Computer Displays with Low-Cost and Mobile Receiver System[C]. *The 20th European Signal Processing Conference (EUSIPCO)*, 2012: 1767-1771.

[3] Tosaka T, Yamanaka Y, Fukunaga K. Method for Determining Whether or not Information is Contained in Electromagnetic Disturbance Radiated from a PC Display[J]. *IEEE Transactions on Electromagnetic Compatibility*, 2011, 53(2): 318-324.

[4] Kuhn M G. Cipher Instruction Search Attack on the Bus-Encryption Security Microcontroller DS5002FP[J]. *IEEE Transactions on Computers*, 1998, 47(10): 1153-1157.

[5] Backes M, Dürmuth M, Unruh D. Compromising Reflections-or-how to Read LCD Monitors around the Corner[C]. *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 2008: 158-169.

[6] Kuhn M G. Electromagnetic Eavesdropping Risks of Flat-Panel Displays[C]. *The International Workshop on Privacy Enhancing Technologies*, 2004: 88-107.

[7] Tsai S S, Chen H Z, Chen D, et al. Mobile Visual Search on Printed Documents Using Text and Low Bit-Rate Features[C]. *2011 18th IEEE International Conference on Image Processing*, 2011: 2601-2604.

[8] Barber D B, Redding J D, McLain T W, et al. Vision-Based Target Geo-Location Using a Fixed-Wing Miniature Air Vehicle[J]. *Journal of Intelligent and Robotic Systems*, 2006, 47(4): 361-382.

- [9] Kisacanin B, Pavlovic V, Huang T S. Preface to Workshop on Real-Time Vision for Human-Computer Interaction[J]. *2004 Conference on Computer Vision and Pattern Recognition Workshop*, 2004: 150.
- [10] Desouza G N, Kak A C. Vision for Mobile Robot Navigation: A Survey[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2002, 24(2): 237-267.
- [11] Merler M, Galleguillos C, Belongie S. Recognizing Groceries in Situ Using in Vitro Training Data[C]. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007: 1-8.
- [12] Ham Y K, Kang M S, Chung H K, et al. Recognition of Raised Characters for Automatic Classification of Rubber Tires[J]. *Optical Engineering*, 1995,34(1): 102-110.
- [13] Boyat A K, Joshi B K. A Review Paper: Noise Models in Digital Image Processing[J]. *Signal & Image Processing: an International Journal*, 2015, 6(2): 63-75.
- [14] Nazaré, Tiago S, Contato W A, et al. Deep convolutional neural networks and noisy images[C]. *The Iberoamerican Congress on Pattern Recognition*, 2017:416-424.
- [15] Wang Runtian, Sang Nong, Ding Ding, et al. Text Detection in Natural Scene Image: A Survey[J]. *Acta Automatica Sinica*, 2018,44(12): 2113-2141.
- [16] Hinton G, Deng L, Yu D, et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups[C]. *IEEE Signal processing magazine*, 2012: 29.
- [17] Krizhevsky A, Sutskever I, Hinton G E. ImageNet Classification with Deep Convolutional Neural Networks[J]. *Communications of the ACM*, 2017, 60(6): 84-90.
- [18] Campos T E, Babu B R, Varma M. Character recognition in natural images[C]. *VISAPP*, 2009: 05-08.
- [19] Epshtein B, Ofek E, Wexler Y. Detecting Text in Natural Scenes with Stroke Width Transform[C]. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010: 2963-2970.
- [20] Weinman J J, Learned-Miller E, Hanson A R. Scene Text Recognition Using Similarity and a Lexicon with Sparse Belief Propagation[J]. *IEEE Trans Pattern Anal Mach Intell*, 2009, 31(10): 1733-1746.
- [21] Pal C, Sutton C, McCallum A. Sparse Forward-Backward Using Minimum Divergence Beams for Fast Training of Conditional Random Fields[C]. *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, 2006: V.
- [22] Mishra A, Alahari K, Jawahar C V. Top-down and Bottom-up Cues for Scene Text Recognition[C]. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2012: 2687-2694.
- [23] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]. *The 18th International Conference on Machine Learning*, 2001:282-289.
- [24] Wang T, Wu D J, Coates A, et al. End-to-End Text Recognition with Convolutional Neural Networks[C]. *The 21st International Conference on Pattern Recognition*, 2012: 3304-3308.
- [25] Bissacco A, Cummins M, Netzer Y, et al. PhotoOCR: Reading Text in Uncontrolled Conditions[C]. *2013 IEEE International Conference on Computer Vision*, 2013: 785-792.
- [26] Karatzas D, Shafait F, Uchida S, et al. ICDAR 2013 Robust Reading Competition[C]. *12th International Conference on Document Analysis and Recognition*, 2013: 1484-1493.
- [27] Russell S J, Norvig P. Artificial Intelligence: A Modern Approach. Prentice Hall[EB/OL]. 2009.
- [28] 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*, 2017, 39(11): 2298-2304.
- [29] Yin F, Wu Y C, Zhang X Y, et al. Scene Text Recognition with Sliding Convolutional Character Models[EB/OL]. 2017
- [30] Liu W, Chen C F, Wong K Y, et al. STAR-Net: A Spatial Attention Residue Network for Scene Text Recognition[C]. *The British Machine Vision Conference 2016*, 2016: 19-22.
- [31] Mohammed, Faisal, al-amer, et al. Al-Hasani and Uhood. Noisy character recognition technique based on moments of inertia[J]. *International Journal of Advancements in Research & Technology*, 2013,2(5),19-23.
- [32] Nautiyal, Thapliyal C, Singh S, et al. Noisy Character Recognition. *Global Journal of Pure and Applied Mathematics*[J], 2017,13(6): 1875-1892.
- [33] Singha S, Imran S, M A, et al. A Robust System for Noisy Image Classification Combining Denoising Autoencoder and Convolutional Neural Network[J]. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS*, 2018, 9(1): 224-235.
- [34] Ren X W, Wang T, Li J Y, et al. Research on Handwritten Chinese Character Recognition Based on Deep Learning with Different Noise[J]. *Application Research of Computers*, 2019, 36(12): 3878-3881.  
(任晓文, 王涛, 李健宇, 等. 基于深度学习的异噪声下手写汉字识别的研究[J]. *计算机应用研究*, 2019, 36(12): 3878-3881.)
- [35] LeCun Y, Boser B, Denker J S, et al. Backpropagation Applied to Handwritten Zip Code Recognition[J]. *Neural Computation*, 1989, 1(4): 541-551.
- [36] Cireşan D C, Meier U, Masci J, et al. High-Performance Neural Networks for Visual Object Classification[OB/EL]. 2011:ArXiv Preprint ArXiv:1102.0183.
- [37] Rico A, Fornés A. Camera-Based Optical Music Recognition Us-

- ing a Convolutional Neural Network[C]. *The 14th IAPR International Conference on Document Analysis and Recognition*, 2017: 27-28.
- [38] Simonyan, Karen, Zisserman A. Very deep convolutional networks for large-scale image recognition[EB/OL]. 2014:ArXiv Preprint ArXiv:1409.1556.
- [39] Huang G, Liu Z, van der Maaten L, et al. Densely Connected Convolutional Networks[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 2261-2269.
- [40] Bengio Y, Simard P, Frasconi P. Learning Long-Term Dependencies with Gradient Descent is Difficult[J]. *IEEE Transactions on Neural Networks*, 1994, 5(2): 157-166.
- [41] Gers F A, Schmidhuber J, Cummins F. Learning to Forget: Continual Prediction with LSTM[J]. *1999 Ninth International Conference on Artificial Neural Networks ICANN 99 (Conf Publ No 470)*, 2: 850-855vol.2.
- [42] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.
- [43] Gers F A, Schraudolph N N, Schmidhuber J. Learning precise timing with LSTM recurrent networks[J]. *Journal of machine learning research*, 2002, 3(8):115-143.
- [44] Graves A, Mohamed A R, Hinton G. Speech Recognition with Deep Recurrent Neural Networks[C]. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013: 6645-6649.
- [45] Graves A, Fernández S, Gomez F, et al. Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks[C]. *The 23rd international conference on Machine learning - ICML '06*, 2006: 369-376.
- [46] Shi B G, Yao C, Liao M H, et al. ICDAR2017 Competition on Reading Chinese Text in the Wild (RCTW-17)[C]. *The 14th IAPR International Conference on Document Analysis and Recognition*, 2017: 1429-1434.
- [47] He W H, Zhang X Y, Yin F, et al. Multi-Oriented and Multi-Lingual Scene Text Detection with Direct Regression[J]. *IEEE Transactions on Image Processing*, 2018, 27(11): 5406-5419.
- [48] Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition[EB/OL]. 2014
- [49] Jin G J, Xiao H, Fu L, et al. Construction and deep processing of modern Chinese corpus. *Applied Linguistics*[J], 2005(2):111-120. (靳光瑾, 肖航, 富丽, 等. 现代汉语语料库建设及深加工[J]. *语言文字应用*, 2005(2): 111-120.)
- [50] Kingma D P, Ba J. Adam: A method for stochastic optimization[EB/OL]. 2014: ArXiv Preprint ArXiv:1412.6980.
- [51] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks[C]. *The thirteenth international conference on artificial intelligence and statistics*, 2010: 249-256.
- [52] Vincent P, Larochelle H, Lajoie I, et al. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion[J]. *Journal of machine learning research*, 2010,11(12):3371-3408.
- [53] Vincent P, Larochelle H, Bengio Y, et al. Extracting and Composing Robust Features with Denoising Autoencoders[C]. *The 25th international conference on Machine learning - ICML '08*, 2008: 1096-1103.



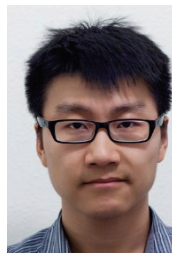
吕志强 于 2007 年在哈尔滨工业大学微电子学与固体电子学专业获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为信号处理及系统实现。研究兴趣包括：高噪声图像处理与文本识别。Email: lvzhiqiang@iie.ac.cn



张磊 于 2017 年在北京工业大学大学电子信息工程专业获得学士学位。现在中国科学院大学网络空间安全专业攻读硕士学位。研究领域为网络空间安全。研究兴趣包括：图像处理、文本识别。Email: zhanglei1995@iie.ac.cn



夏宇琦 于 2016 年在武汉科技大学电子信息工程专业获得学士学位。现在中国科学院大学通信与信息系统专业攻读硕士学位。研究领域为声音安全、跨网攻防。研究兴趣包括：计算机网络、隐蔽通信、声音安全。Email: xiayuqi@iie.ac.cn



张宁 于 2013 年在哥伦比亚大学电气工程专业获得硕士学位。现任中国科学院信息工程研究所工程师。研究领域为网络空间安全。研究兴趣包括：硬件安全等。Email: zhangning@iie.ac.cn