

基于空域与频域关系建模的篡改文本图像检测

王裕鑫, 张博强, 谢洪涛, 张勇东

(中国科学技术大学, 安徽 合肥 230026)

摘要: 近年来, 篡改文本图像在互联网的广泛传播为文本图像安全带来严重威胁。然而, 相应的篡改文本检测 (TTD, tampered text detection) 方法却未得到充分的探索。TTD 任务旨在定位图像中所有文本区域, 同时根据纹理的真实性判断文本区域是否被篡改。与一般的文本检测任务不同, TTD 任务需要进一步感知真实文本和篡改文本分类的细粒度信息。TTD 任务有两个主要挑战: 一方面, 由于真实文本和篡改文本的纹理具有较高的相似性, 仅在空域 (RGB) 进行纹理特征学习的篡改文本检测方法不能很好地区分两类文本; 另一方面, 由于检测真实文本和篡改文本的难度不同, 检测模型无法平衡两类文本的学习过程, 从而造成两类文本检测精度的不平衡问题。相较于空域特征, 文本纹理在频域中的不连续性能够帮助网络鉴别文本实例的真伪, 根据上述依据, 提出基于空域和频域 (RGB and frequency) 关系建模的篡改文本检测方法。采用空域和频域特征提取器分别提取空域和频域特征, 通过引入频域信息增强网络对篡改纹理的鉴别能力; 使用全局空频域关系模块建模不同文本实例的纹理真实性关系, 通过参考同幅图像中其他文本实例的空频域特征来辅助判断当前文本实例的真伪性, 从而平衡真实和篡改文本检测难度, 解决检测精度不平衡问题; 提出一个票据篡改文本图像数据集 (Tampered-SROIE) 来验证上述篡改文本检测方法的有效性, 该数据集包含 986 张图像 (626 张训练图像和 360 张测试图像)。该方法在 Tampered-SROIE 上的真实和篡改文本检测 F 值分别达到 95.97% 和 96.80%, 同时降低检测精度不平衡性 1.13%。该方法从网络结构与检测策略的角度为篡改文本检测任务提供了新的解决方案, 同时 Tampered-SROIE 为以后的篡改文本检测方法提供了评估基准。

关键词: 篡改文本检测; 空频域关系建模; 篡改文本检测数据集; 评估基准

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2022035

Tampered text detection via RGB and frequency relationship modeling

WANG Yuxin, ZHANG Boqiang, XIE Hongtao, ZHANG Yongdong

University of Science and Technology of China, Hefei 230026, China

Abstract: In recent years, the widespread dissemination of tampered text images on the Internet constitutes an

收稿日期: 2022-01-11; 修回日期: 2022-04-17

通信作者: 谢洪涛, htxie@ustc.edu.cn

基金项目: 国家自然科学基金 (62121002, 62022076, U1936210)

Foundation Item: The National Nature Science Foundation of China (62121002, 62022076, U1936210)

引用格式: 王裕鑫, 张博强, 谢洪涛, 等. 基于空域与频域关系建模的篡改文本图像检测[J]. 网络与信息安全学报, 2022, 8(3): 29-40.

Citation Format: WANG Y X, ZHANG B Q, XIE H T, et al. Tampered text detection via RGB and frequency relationship modeling[J]. Chinese Journal of Network and Information Security, 2022, 8(3): 29-40.

important threat to the security of text images. However, the corresponding tampered text detection (TTD) methods have not been sufficiently explored. The TTD task aims to locate all text regions in an image while judging whether the text regions have been tampered with according to the authenticity of the texture. Thus, different from the general text detection task, TTD task further needs to perceive the fine-grained information for real-world and tampered text classification. TTD task has two main challenges. One the one hand, due to the high similarity in texture between real-world texts and tampered texts, TTD methods that only learn from RGB domain features have limited capability to distinguish these two-category texts well. On the other hand, as the different detecting difficulty exists in real-world texts and tampered texts, the network cannot well balance the learning process of the two-category texts, resulting in the imbalance detection performance between real-world and tampered texts. Compared with RGB domain features, the discontinuity of text texture in frequency domain can help the network to identify the authenticity of text instances. Accordingly, a new TTD method based on RGB and frequency information relationship modeling was proposed. The features in the RGB and frequency domains were extracted by independent feature extractors respectively. Thus, the identification ability of tampered texture can be enhanced by introducing frequency information during the texture perception. Then, a global RGB-frequency relationship module (GRM) was introduced to model the texture authenticity relationship between different text instances. GRM referred to the RGB-frequency features of other text instances in the same image to assist in judging the authenticity of the current text instance, which solved the problem of imbalanced detection performance. Furthermore, a new TTD dataset (Tampered-SROIE) was proposed to evaluate the effectiveness of proposed method, which contains 986 images (626 training images and 360 test images). By evaluating on the Tampered-SROIE, the proposed method obtains 95.97% and 96.80% in F-measure for real-world and tampered texts respectively and reduces the imbalanced detection accuracy by 1.13%. The proposed method will give new insights to the TTD community from the perspective of network structure and detection strategy. Tampered-SROIE also provides an evaluation benchmark for future TTD methods.

Keywords: tampered text detection, RGB-frequency relationship modeling, tampered text detection dataset, evaluation benchmark

0 引言

文本作为一种重要的信息传播媒介, 蕴含大量的重要敏感信息^[1-3]。随着篡改技术的发展, 计算机可以自动将敏感内容转化为虚假信息, 用于欺诈、营销或其他非法目的。近年来, 经由深度学习篡改算法生成的文本图像在互联网上广泛传播^[4-6], 对多个领域产生了极大的负面影响^[7-8], 包括金融票据识别领域、证件识别领域、网页内容识别领域等。

篡改文本检测 (TTD, tampered text detection) 作为多媒体信息安全领域的一个新兴研究方向, 是指通过对文本图像中纹理特征的分析, 捕捉真实文本和篡改文本之间的纹理差异性, 以确定文本图像中文字区域的真伪性。如图 1 所示, 篡改文本检测技术包含两个步骤: 文本定位和文本真伪性鉴别。文本定位步骤需要对文本图像内容进

行分析, 定位出具有文本纹理特征的文本实例。真伪性鉴别步骤需要对文本实例的纹理真伪性进行分析, 鉴别当前文本实例是否为篡改文本。

篡改文本技术研究领域已经有较多公开发表的研究成果, 而篡改文本检测研究仍处于起步阶段, 未得到足够重视。篡改文本检测技术一方面能够与篡改文本技术相抗衡, 保护文本图像内容的真实性, 保障网络信息安全与人民财产安全; 另一方面, 篡改文本检测技术能有效地反映篡改文本技术的性能, 实现矛与盾的良性可持续发展。

篡改文本检测任务有两个主要挑战。① 局部纹理差异性捕捉困难。篡改文本与真实文本仅存在局部纹理差异。② 真实和篡改文本检测精度平衡困难。相较传统的文本检测任务, 篡改文本检测任务需要进一步区分篡改和真实文本。由于真实和篡改文本分类难度不一致, 训练过程中网络无法平衡两类的学习过程, 导致在测试过程中两

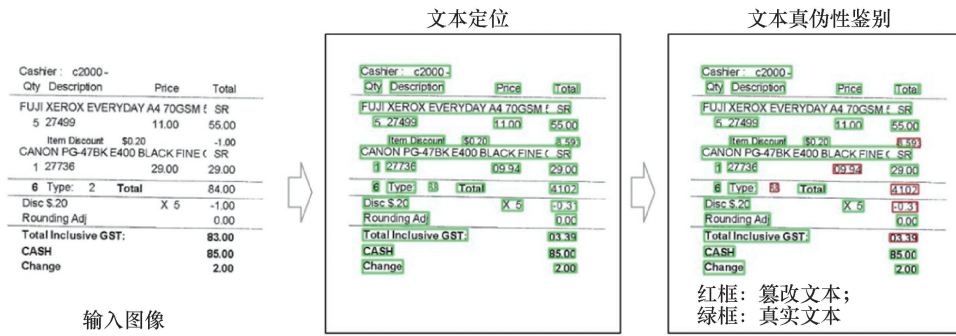


图1 篡改文本检测任务流程
Figure 1 The pipeline of tampered text detection task

类检测精度差异较大。上述挑战极大地限制了篡改文本检测方法的性能。因此, 如何准确地捕捉局部纹理差异性, 同时平衡篡改和真实类别学习难度, 是目前篡改文本检测研究的重要方向。

本文提出一种基于空域和频域 (RGB and frequency) 关系建模的篡改文本检测方法。为了准确地捕捉局部纹理差异性, 引入频域特征以增强对篡改纹理的鉴别能力。此外, 采用全局空频域关系模块建模文本实例之间的纹理真伪性关系, 通过参考同幅图像中其他所有文本区域的空频域特征辅助当前文本实例的真伪性鉴别, 平衡篡改和真实类别的学习难度。进一步地, 为了验证本文方法的性能, 同时为今后的篡改文本检测方法提供评估基准, 本文提出一个新的票据篡改文本图像数据集。

1 相关工作

早期的图像篡改检测方法主要针对不同的篡改手段设计对应的检测算法, 常见的有复制粘贴检测、拼接检测等。Fridrich 等^[9]在 2003 年提出一种检测复制粘贴篡改手段的方法, 利用离散余弦变换 (DCT, discrete cosine transform) 对图像块进行分类, 并且结合字典排序, 解决了计算量大的问题。基于矩不变量的复制粘贴检测算法^[10]和基于关键点 SIFT 特征的复制粘贴检测算法^[11]对旋转缩放等处理具有鲁棒性。Farid^[12]利用 JPEG (joint photographic experts group) 压缩特性的不一致性, 检测原始图片和经过拼接的篡改图片的 JPEG 压缩特性, 根据其差异性进行篡改检测。

1.1 文档图像篡改检测

常见的文档篡改操作包括文档区域复制移动、

文档拼接以及像素级的更改篡改字符区域等。文档篡改检测的一系列研究方法针对篡改操作所导致的缺陷展开。早期的文档篡改检测主要通过打印机分类和识别技术以确定在文档生成过程中所使用的硬件类型^[13-15], 这些方法利用了不同源类型之间所选特性的显著差异性, 但它们无法检测到文档图像内部复制粘贴和重打印伪造操作的痕迹。考虑到原始文档经机器重打印、扫描或经文字编辑软件处理之后, 会存在不均匀的垂直尺度问题, 文献[16]利用内在文档元素特性, 通过自动识别同一来源不同文档之间保持静态一致的模板区域来衡量文档内容的真实性, 从而检测同源文档中的篡改操作。

在大多数文档图像中, 同一单词或句子中所采用的字体大多一致。因此, 一些工作通过关注字体识别^[17-18]来进行篡改文本与真实文本的划分。文献[19]提出基于文档字体特征的自动伪造检测方法以检测同一单词内部不同字体拼接篡改, 其使用条件随机场, 通过分类字符字体及与邻域字符字体类型进行对比, 来区分真实字符区域和篡改字符区域。除使用字体属性外, 字符形状和文本对齐等属性也被用来检测篡改文档。文献[20]采用行级一致性检测的方式, 基于统计模型, 通过检测文本行对齐与旋转变化的, 来鉴别文档中每行文本的真实性。与基于文本字体、排列、形状布局等外观特征进行区域鉴别的方式不同, 文献[21]采用基于支持向量机的分类方式进行篡改文档检测, 利用局部二值模式特征描述算子 (LBP) 捕获篡改区域的可分辨纹理特征, 以寻找文档图像内在特征不一致性。文献[22]将自然场景

图像篡改检测任务中基于图像块的重复项检测方法迁移到扫描文档篡改区域检测任务中,探索特定阈值和参数影响下基于图像块的检测方法对复制移动到文档的伪造文本的检测性能。

相较上述使用人工设计的特征及评分方式,以数据为驱动的基于深度学习的篡改文档检测方法获得了更好的检测性能。文献[23]将拼接检测问题定义为图节点分类问题,通过光学字符识别(OCR)技术获取文档图像中文本块的位置及内容,并基于此结果构建图神经网络。具体地,该方法以各文本块为图节点,节点间的连接与否由检测框距离决定,基于预训练过的变分自动编码器对文档图像进行特征提取,进而引入图注意力机制捕获更强的文本块上下文特征,以提升分类准确度。上述方法仅在 RGB 域捕获浅层语义特征差异,然而篡改操作通常会留下高频痕迹,此类高频信息很难在 RGB 域中被捕获。因此,本文方法通过引入频率特征并将其与 RGB 域特征融合来增强网络对篡改纹理的鉴别能力,同时使用关系模块对篡改类别与真实文本类别进行关系建模,平衡真实文本和篡改文本的学习难度,实现精确度更高的篡改文本检测。

1.2 频域信息在篡改检测中的应用

随着图像伪造技术手段的进一步提升,篡改检测方法已经难以从视觉上检测出图像是否被篡改。因此,越来越多的研究转向引入更多的信息和先验知识^[24-25]作为篡改检测的辅助信息。

在难以从视觉图像中获取有效伪造线索的情况下,有研究^[26-28]发现,原图和相应篡改后的图像在频域上相同位置特征不一致。因此,图像的频域信息对于准确的篡改检测至关重要。在面部伪造检测中,由于背景、性别、年龄、伪造方法的多样性,基于固定的频域信息提取方法不足以从频域中捕捉细微的伪造模式^[26],因此,文献[27]引入两种提取频域特征的方法 FAD (frequency-aware decomposition) 和 LFS (local frequency statistics),并设计了一个融合模块来融合双流网络的特征,从而在频域内实现对面部伪造图像视频的检测。这种方法将传统学习和深度学习结合,在低质量的伪造图片检测上取得了不错的结果。文献[28]提出新颖的频率感知判别特征学习框架

来进行篡改鉴别,利用自适应的频率特征生成模块以数据驱动的方式挖掘频率线索,从而避免使用太多不全面的先验知识,同时结合度量学习,提出单中心损失来学习更多的判别特征,进一步提高模型的篡改检测能力。

上述频域信息的捕获主要针对人脸或其他非文本目标区域。由于文本独特的性质,直接使用上述频域信息提取方式会限制篡改文本检测器的性能,因此需要针对文本特性,设计符合文本特性(如局部存在、长宽比例变化大)的频域特征提取器。

2 方法设计

2.1 概述

本文提出的基于空频域关系建模的篡改文本检测方法有以下几点贡献。

(1) 创新性地在篡改文本检测任务中引入频域信息增强篡改纹理特征,通过同时捕捉空域和频域的信息提升网络对局部纹理差异性的鉴别能力。

(2) 全局空频域关系模块提供了一种简洁、有效的平衡篡改和真实类别学习难度的方法,通过建模全局文本实例之间的空频域特征关系,借助其他文本实例空频域信息来辅助鉴别当前文本实例的真伪性,帮助网络更好地平衡真实和篡改文本的学习难度,提升检测精度。

基于空域和频域关系建模的篡改文本检测方法流程如图 2 所示。本文采用两阶段目标检测的框架,包含文本区域建议框的生成、文本区域建议框微调、文本几何预测和文本真伪性鉴别。文本区域建议框微调仅在测试过程使用,用于微调文本区域建议框的位置,提升检测器对多尺度文本的检测能力。首先,输入图像经过主干网络提取空域特征,同时通过离散余弦变换提取频域特征。其次,通过空域特征结合区域建议网络(RPN, region proposal network)生成文本候选框。然后,使用 N 个区域候选框对文本的空域特征和频域特征进行裁剪,通过感兴趣区域对齐(RoI align)归一化特征尺寸,再输入全局空频域关系模块进行关系建模。最后,将全局空频域关系模块的输出特征用于真伪性鉴别和文本区域建议框微调。为了表示任意形状文本,文本几何预测分支采用基于轮廓点

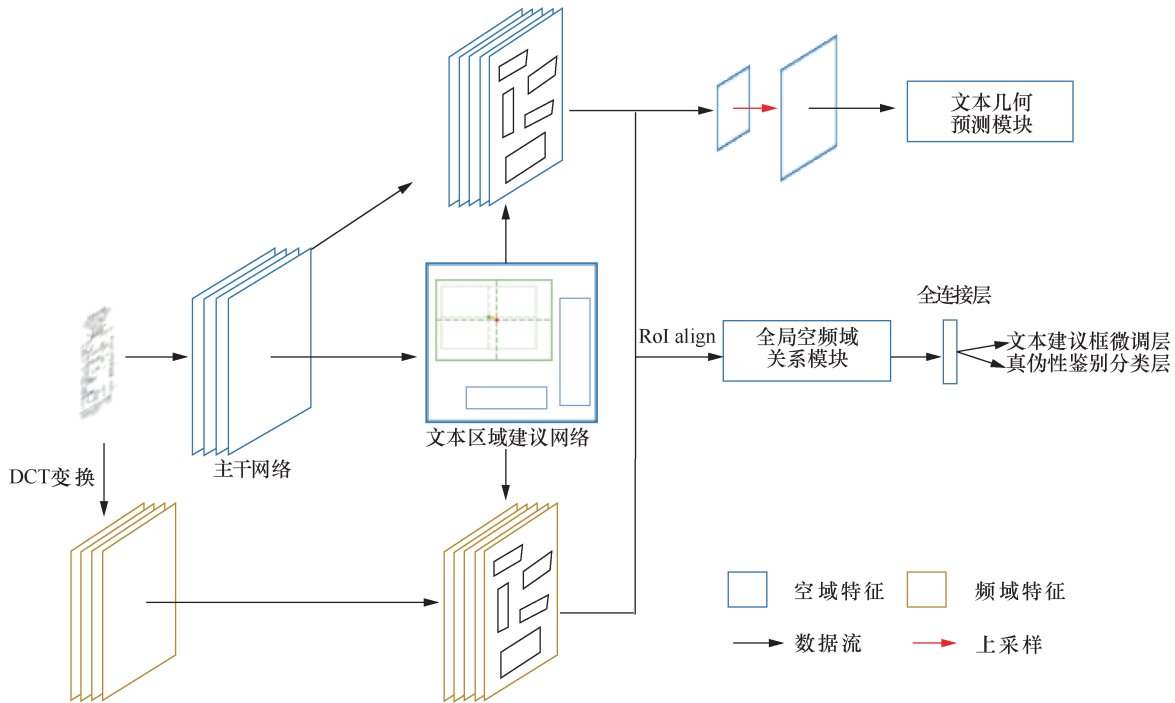


图2 基于空域和频域关系建模的篡改文本检测方法流程
Figure 2 The pipeline of tampered text detection via RGB and frequency relationship modeling

的分割算法，通过预测并连接文本区域轮廓点描述任意形状文本。

全局空频域关系模块的输入为文本候选框的空域特征和频域特征，文本候选框的空域特征直接通过对空域的特征进行裁剪得到。文本候选框频域特征的获取方式如下。① 将输入图像进行离散余弦变换得到频域特征。如图3所示，整个过程包含RGB图像向YCbCr图像转变、DCT变换、几何变换和级联过程。② 通过三层卷积增强频域特征的表达能力，同时与空域特征维度对齐。③ 采用空域特征得到的文本建议框，对频域特征中的对应区域进行裁剪，生成对应文本候选框的频域特征。相较人脸篡改检测的频域信息获取，本文方法能够捕捉到局部纹理的频域特征，通过结合文本区域建议框的多尺度特性，提取符合文本特征的局部、多尺度频域特征。

2.2 全局空频域关系模块

若简单地根据当前文本候选框内的特征进行纹理真伪性鉴别，由于缺乏全局信息的感知能力，将导致网络无法平衡真实和篡改文本类别的学习难度，从而在测试过程中两类的检测精度差异较

大，造成检测精度不平衡的问题。为了捕获全局信息辅助当前文本候选框的真伪性鉴别，本文提出全局空频域关系模块，通过感知当前文本候选框与其他文本候选框的空频域特征相似性，平衡真实和篡改类别的学习难度。

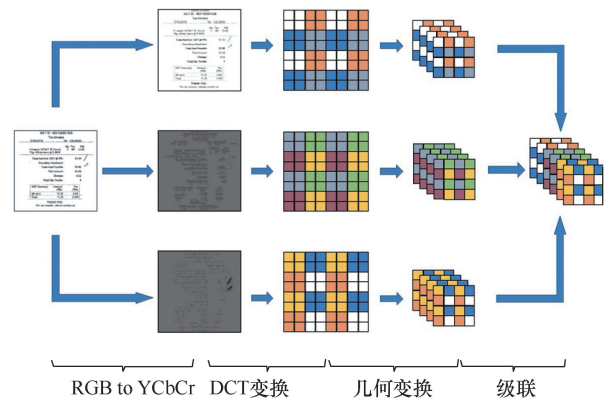


图3 DCT过程
Figure 3 The process of DCT

全局空频域关系模块结构如图4所示，主要包含两个部分：融合模块和关系模块。首先，输入文本候选框的空域特征和频域特征并通过融合模块进行特征融合。然后，将融合后的特征输入

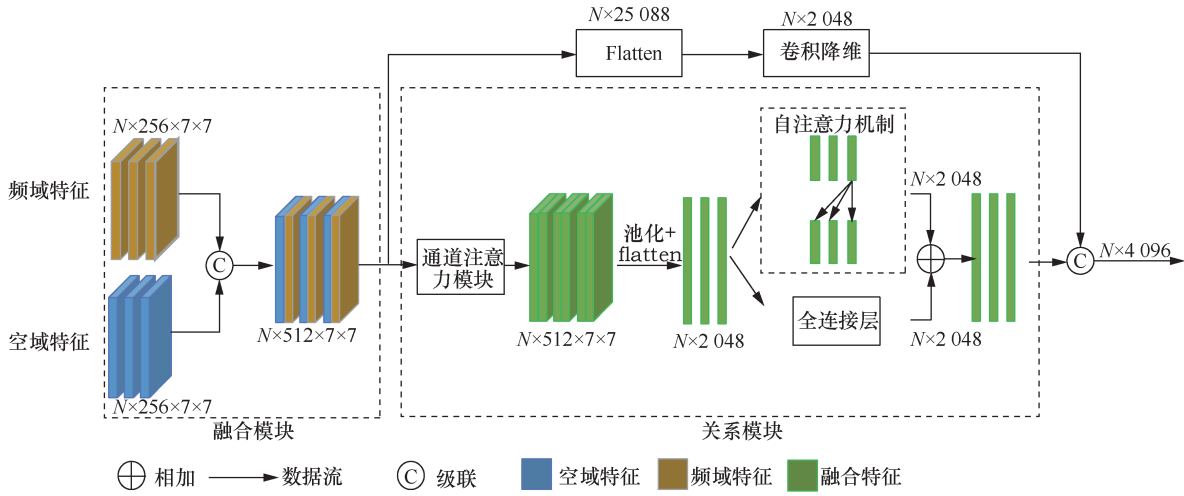


Figure 4 The structure of global RGB-frequency relationship module

关系模块，建模不同文本候选框之间的空频域特征相似性。

在关系模块中，首先采用通道注意力机制控制空域和频域的信息表达。通道注意力机制如图 5 所示，输入特征通过池化和全连接层获得通道注意力图，然后通过通道注意力图控制不同通道的特征表达。基于通道注意力机制的信息增强方法有效地解决了空域和频域特征之间的特征异质性问题，针对不同的文本候选框，自适应地考虑不同域的信息，提升特征的表达能力。然后，使用池化操作将通道注意力机制输出特征的尺寸变为 $N \times 512 \times 2 \times 2$ ，并通过展开操作将特征尺寸变换为 $N \times 2048$ 。最后，采用自注意力机制 (self-attention)，融合其他文本候选框的空频域特征对当前文本候选框的空频域特征增强。自注意力机制的计算方法如式(1)所示。

$$\text{Attention}(q, k, v) = \text{softmax}\left(\frac{qk^T}{\sqrt{d}}\right)v \quad (1)$$

$$q = W_q f_{\text{input}} \quad (2)$$

$$k = W_k f_{\text{input}} \quad (3)$$

$$v = W_v f_{\text{input}} \quad (4)$$

其中， f_{input} 是输入特征， d 是输入特征 k 的通道数， W_q 、 W_k 、 W_v 是可学习的参数。在自注意力机制中，首先计算当前文本框和其他文本框空频域特征的相似性；然后通过相似性矩阵，在每个

文本框空频域特征中，融合其余文本框的空频域特征。因此，全局空频域关系模块有效地通过感知全局的空频域信息，辅助当前文本框的真伪性鉴别，平衡真实和篡改文本检测的学习难度。

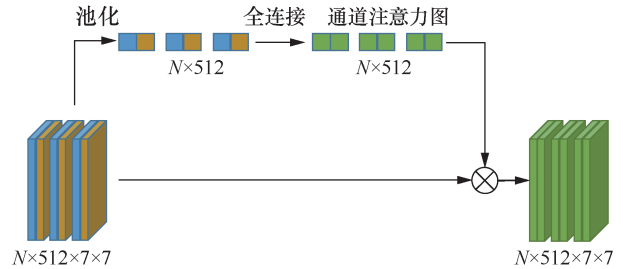


Figure 5 The channel attention mechanism

2.3 文本几何预测模块

在文本几何预测模块中，采用预测轮廓点分割图表示任意形状文本。由于分割过程易受假阳性背景噪声点的影响（如栅格、波浪线等），本文参考了 ContourNet^[1]中正交纹理感知模块结构和正交融合操作。

首先，将 RoI align 后的特征图上采样，感知细节纹理特征。然后，使用两个正交卷积核捕捉水平和垂直方向上的纹理，预测两个方向上具有文字纹理响应的轮廓点。在训练过程中，将两个正交轮廓点分割图的损失函数相加，作为文本几何预测模块的学习目标。在测试过程中，首先，使用正交的卷积核分别感知水平和垂直方向上的纹理；然后，使用 0.5 阈值过滤正交轮廓点分割图，得到轮廓点候选集合；最后，采用正交融合操

作^[1]抑制背景的假阳性噪声点,生成最终文本轮廓点,通过连接最终文本轮廓点,实现任意形状文本的检测。

2.4 网络优化指标

基于空频域关系建模的篡改文本检测网络优化指标由4个部分组成:文本区域建议网络损失函数(L_{tpn})、真伪性鉴别损失函数(L_{cls})、文本区域建议框微调损失函数(L_{ref})与文本几何预测损失函数(L_{geo})。整体的损失函数如式(5)所示。

$$L = L_{\text{tpn}} + L_{\text{cls}} + L_{\text{ref}} + L_{\text{geo}} \quad (5)$$

在文本区域建议网络损失函数部分,本文采用交并比(IoU)优化指标,通过计算预测文本候选框和标签文本候选框的IoU,并优化IoU指标来指导文本区域建议网络对文本粗定位的学习。如式(6)所示,其中 P 和 G 分别是预测框和标签框, $L_{\text{cross_entropy}}$ 为交叉熵损失函数。

$$L_{\text{tpn}} = -\log \frac{|P \cap G|}{|P \cup G|} + L_{\text{cross_entropy}} \quad (6)$$

在真伪性鉴别损失函数部分,本文采用交叉熵损失函数优化分类网络。在文本建议框微调损失函数部分,本文借鉴Mask r-CNN^[29],采用smooth L_1 损失函数优化微调层。smooth L_1 损失函数如式(7)所示。

$$\text{smooth}_{L_1} = \begin{cases} 0.5x^2, & |x| < 1 \\ |x| - 0.5, & \text{其他} \end{cases} \quad (7)$$

在文本几何损失函数部分,本文参考了ContourNet^[1],通过平衡的交叉熵损失函数指导文本几何预测的学习。

$$L_{\text{geo}} = -\frac{N_{\text{neg}}}{N} y_i \log p_i - \frac{N_{\text{pos}}}{N} (1 - y_i) \log(1 - p_i) \quad (8)$$

其中, N_{neg} 和 N_{pos} 表示负样本和正样本个数, y_i 和 p_i 表示标签和预测样本, N 表示正负样本的总和。

3 篡改文本数据集制作

本文提出的票据篡改文本图像数据集(Tampered-SROIE)是通过对当前主流的SROIE票据数据集中文本图像篡改得到的^[30]。SROIE数据集公布于ICDAR2019扫描收据光学字符识别和信息提取(scanned receipts OCR and information

extraction)挑战赛。该数据集有986幅完整的扫描收据图像,其中训练图像626幅,测试集360幅。数据集文本内容主要由数字和英文字符组成,文本位置标注为文本矩形包围框的4个角点坐标。特别地,该数据集中一些票据纸张墨水和印刷质量较差,且存在扫描失真、折叠等干扰因素,这使其上开展的篡改文本检测任务更具挑战。

本文所提出的篡改数据集Tampered-SROIE的制作包括两个步骤。1) 成对篡改文本的选取。本文对SROIE中每幅图像随机挑选包含数字且不包含字母的文本实例进行篡改,并选择1/2至1/3数量文本实例作为篡改对象(源字符串),同时保证每幅图像至少篡改一个文本实例。为了增加篡改文本的多样性,源字符串对应的篡改字符串通过随机生成,并使1/3数量篡改字符串的长度增加一位。2) 篡改操作。本文使用SRNet^[31]进行文本篡改。为了训练一个强大的篡改网络,本文基于准备好的源篡改词对生成5万对合成训练样本,并在1张2080Ti GPU上进行10万次迭代训练。本文使用篡改对象的标签包围框来裁剪SROIE图像中的文本实例图像,并使用训练后的SRNet生成篡改文本实例图像,随后将篡改后的文本实例图放回原始图像中相应位置。

在Tampered-SROIE训练集的10 251个数字文本实例中有3 947个被标记为篡改类,在测试集的5 829个数字文本实例中有2 251个被标记为篡改类。Tampered-SROIE数据集的可视化结果如图6所示。Tampered-SROIE篡改文本图像质量较高,能够有效反映篡改文本检测方法的检测性能。

4 实验分析

4.1 实验设置

本文在一张2080Ti显卡上部署网络模型,并使用随机梯度下降法(SGD)优化网络参数。在训练过程中,使用多尺度数据增广提升网络对多尺度文本的拟合能力。具体地,将长边固定为2 000,短边选择400、600、720、1 000、1 200对图像进行等比例缩放。测试过程中,将图像缩放到1 200×2 000进行预测。非极大值抑制算法在测试过程中用于滤除冗余的检测结果。

4.2 评估指标

本文采用召回率、准确率、 F 值评估真实文

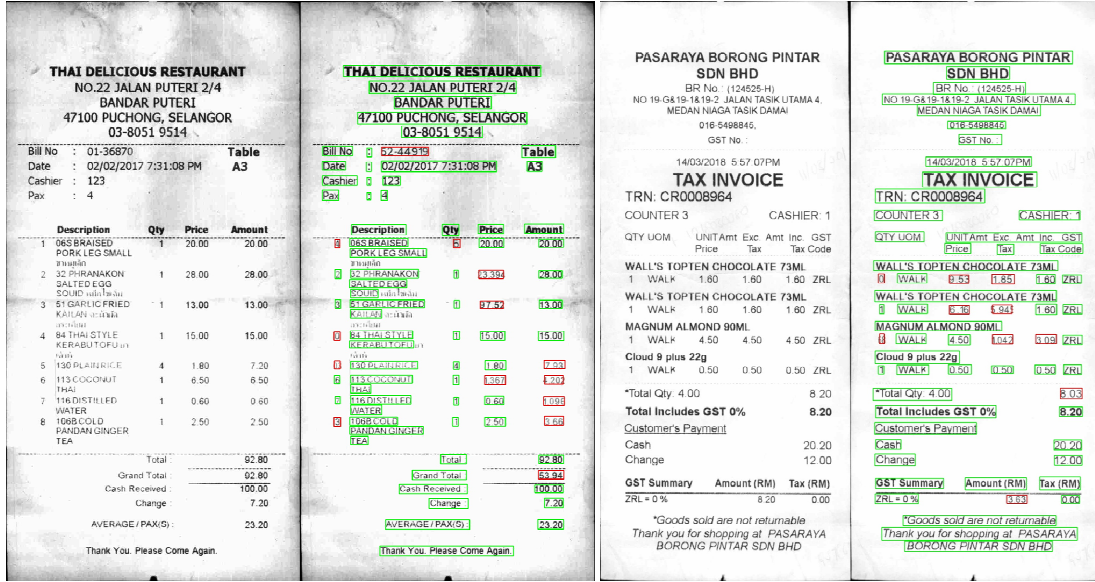


图 6 Tampered-SROIE 数据集可视化
(左: 原始图像; 右: 篡改图像; 红框: 篡改文本; 绿框: 真实文本)

Figure 6 The visualization of Tampered-SROIE dataset
(left: origin image; right: tampered image; texts in red box: tampered texts; texts in green box: real-world texts)

本和篡改文本的检测结果。 F 值的计算方式如式(9)所示。

$$F = \frac{2 \times \text{召回率} \times \text{准确率}}{\text{召回率} + \text{准确率}} \quad (9)$$

为了评估真实类别和篡改类别检测精度不平衡性, 本文进一步引入 $\text{Gap-}F$ 。 $\text{Gap-}F$ 值为真实类别和篡改类别 F 值的差, $\text{Gap-}F$ 的计算方式如式(10)所示。

$$\text{Gap-}F = F_{\text{真实}} - F_{\text{篡改}} \quad (10)$$

召回率、准确率、 F 值和 $\text{Gap-}F$ 评估指标能有效地反映篡改文本检测方法的检测性能, 从多方面反映真实和篡改文本检测效果。

4.3 全局空频域关系模块消融实验

通过部署全局空频域关系模块, 篡改文本检测模型能够感知全局空频域关系以辅助当前文本框真伪性判断, 有效地平衡篡改和真实文本的学习难度。全局空频域关系模块的消融实验结果如表 1 所示, 相较于不使用全局空频域关系模块的方法, 本文方法有效提升了网络在真实类别和篡改类别的检测精度。具体地, 针对真实文本, 召回率、准确率、 F 值提升分别为 2.58%、0.08%、1.35%; 针对篡改文本, 召回率、准确率、 F 值提

升分别为 0.08%、0.12%、0.11%。进一步地, 通过平衡真实类别和篡改类别的学习难度, 本文方法降低了真实类别和篡改类别之间的检测性能差距 (1.96% 和 0.72%)。

全局空频域关系模块有效解决了真实类别和篡改类别检测精度不平衡的问题。在篡改文本检测任务中, 本文证明了该不平衡问题是由于全局信息感知能力缺失所导致的: 因为无法参考其他文本实例的纹理真伪性, 导致真实和篡改文本的学习难度无法平衡。得益于本文全局空频域关系模块的简洁性和有效性, 该模块能够方便地移植到其他篡改文本检测算法中, 同时能够通过增强每个子模块的性能, 进一步提升网络对篡改文本的鉴别能力。

虽然手动调节真实和篡改文本损失函数能够一定程度上缓解检测精度不平衡的问题, 但是针对不同的检测方法, 检测精度不平衡是无法预先确定的。如表 2 所示, $\text{EAST}^{[32]}$ 和 $\text{ATRR}^{[33]}$ 分别倾向于对篡改文本和真实文本具有更好的检测性能。这表明基于人工的损失函数调节是烦琐的, 同时损失函数的系数改变会引入更多不确定性, 因此很难调节到最优的检测性能。但本文提出的全局空频域关系模块是可学习的, 经过实验证明, 该模块能够自适应地缓解检测精度不平衡问题,

实现更加鲁棒的平衡过程。

4.4 频域信息消融实验

通过对频域信息的感知,本文方法能够更好地捕捉局部纹理的差异性。如表1所示,通过融合空域和频域的特征,本文方法将真实文本和篡改文本的检测精度提升到一个新的高度。具体地,在真实文本和篡改文本类别上,检测的 F 值分别达到95.97%和96.80%。进一步分析表1中的结果,虽然频域信息的引入导致网络Gap- F 指标略微升高(0.11%),但是针对篡改和真实文本的精度提升仍然是可观的。因此,本文总结Gap- F 性能的略微下降原因为:频域信息的引入极大地增强了篡改文本类别的检测性能,使篡改文本检测精度相较真实文本检测精度增长速度更快(F 值增长:真实0.18%和篡改0.29%),从而导致Gap- F 性能略微下降。

值得注意的是,本文通过多尺度的文本区域建议框对频域特征进行裁剪,相较一般的频域信息提取,本文所引入的频域特征提取过程是符合文本的局部存在和多尺度特性的。因此,该方法能够广泛适用于篡改文本检测模型,通过提取符合文本特性的频域信息,提升检测模型对篡改纹理特征的鉴别能力。同时,本文的频域信息提取方式为今后篡改文本检测模型引入额外信息辅助检测提供参考,即可以通过文本区域建议框裁剪的形式,引入符合文本特性的辅助信息帮助提升模型检测性能。

4.5 实验结果

为了展现本文方法的有效性,本文从单阶段和两阶段分别选取了最具代表性的方法进行性能对比。具体地,本文复现了EAST^[32]和ATTR^[33]方法,模型配置和训练细节参考EAST和ATTR论文原文进行部署。为了将传统文本检测算法向篡改文本检测任务迁移,本文引入额外的文本分类操作对上述传统文本检测算法的文本定位过程进行了改动。例如,在EAST中,将原始的中心区域预测图分为真实文本中心预测图和篡改文本中心预测图,并分别对两个中心文本预测图进行优化;对于文本几何预测过程,本文对真实和篡改文本共享文本几何预测图。

实验结果如表2所示,本文方法在真实文本和篡改文本上都展现了领先的检测水平。相较同为两阶段的检测算法^[33],本文提出的基于空频域关系建模的篡改文本检测方法在真实文本和篡改文本检测精度上都取得了更好的效果,同时可以有效解决检测精度不平衡问题,防止篡改检测网络实现对单一类文本的准确检测。

通过对实验结果的进一步分析发现,检测精度不平衡问题普遍存在于篡改文本检测算法(|Gap- F |指标在EAST和ATTR算法分别为2.99%和2.37%),且该不平衡性呈现随机出现的特点,即检测精度可能出现向真实文本或篡改文本倾斜的情况。本文通过对全局信息的感知,捕获全局真伪性信息,有效地解决了该不平衡问

表1 消融实验结果 (GRM表示全局空频域关系模块)
Table 1 The result of ablation study (GRM is the global RGB-frequency relationship module)

方法	频域信息	真实文本			篡改文本			Gap- F
		召回率	准确率	F 值	召回率	准确率	F 值	
—	—	92.00%	96.96%	94.44%	95.51%	97.32%	96.40%	1.96%
√	—	94.58%	97.04%	95.79%	95.59%	97.44%	96.51%	0.72%
√	√	94.71%	97.26%	95.97%	96.07%	97.55%	96.80%	0.83%

表2 Tampered-SROIE 实验效果
Table 2 The experiment result on Tampered-SROIE

方法	真实文本			篡改文本			Gap- F
	召回率	准确率	F 值	召回率	准确率	F 值	
EAST ^[32]	91.14%	84.62%	87.76%	91.91%	89.61%	90.75%	2.99%
ATTR ^[33]	96.70%	95.24%	95.96%	94.71%	92.49%	93.59%	2.37%
本文方法	94.71%	97.26%	95.97%	96.07%	97.55%	96.80%	0.83%

题,且本文可学习的平衡方式能够更好地帮助网络同时提升真实文本和篡改文本的检测精度,使两类文本的检测性能都达到最好的效果。

另外,本文将检测结果进行了可视化。如图 7 所示,相较 EAST^[32]和 ATRR^[33],本文的篡改文本检测方法能够更准确地区分篡改和检测文本,同时生成更加紧致的文本包围框,因此在实际应用场景中,本文所提出的篡改文本检测算法更具有实用价值。

5 结束语

本文提出了一种基于空频域关系建模的篡改文本检测方法,在高质量篡改文本图像中,通过引入频域信息同时建模不同文本实例之间的空频域特征关系,实现准确的篡改文本检测。虽然对

传统文本检测方法的简单修改能够适用于篡改文本检测任务,但本文实验证明,篡改文本检测任务中局部纹理差异性感知困难和检测精度平衡困难问题极大地限制了篡改文本检测算法性能。

此外,本文所提出的思想可以扩展到其他篡改文本检测算法中,通过直接使用或者简单修改文中的模块,可以实现篡改文本检测精度的显著提升,这为篡改文本检测技术的相关研究提供了新的方向和思路。

参考文献:

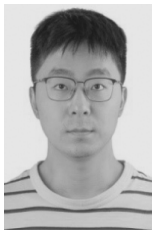
- [1] WANG Y, XIE H, ZHA Z J, et al. Contournet: taking a further step toward accurate arbitrary-shaped scene text detection[C]//CVPR. 2020: 11753-11762.
- [2] WANG Y, XIE H, FANG S, et al. From two to one: a new scene



图 7 检测结果可视化
Figure 7 The visualization of detection results

- text recognizer with visual language modeling network[C]//ICCV. 2021:14194-14203.
- [3] 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/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7098-7107.
 - [4] WU L, ZHANG C, LIU J, et al. Editing text in the wild[C]//ACM MM. 2019: 1500-1508.
 - [5] YANG Q, HUANG J, LIN W. Swaptxt: image based texts transfer in scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14700-14709.
 - [6] ROY P, BHATTACHARYA S, GHOSH S, et al. Stefann: scene text editor using font adaptive neural network[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13228-13237.
 - [7] 王志红, 黄惠芬, 常玉红. 基于SVD分解和Hadamard变换的图像半脆弱水印技术[J]. 网络与信息安全学报, 2017, 3(5): 26-31.
WANG Z H, HUANG H F, CHANG Y H. Semi-fragile watermarking technology based on SVD decomposition and Hadamard transform[J]. Chinese Journal of Network and Information Security, 2017, 3(5): 26-31.
 - [8] 常玉红, 黄惠芬, 王志红. 检测图像篡改的脆弱水印技术[J]. 网络与信息安全学报, 2017, 3(7): 47-52.
CHANG Y H, HUANG H F, WANG Z H. Fragile watermarking technique for detecting image[J]. Chinese Journal of Network and Information Security, 2017, 3(7): 47-52.
 - [9] FRIDRICH J, SOUKAL D, LUKAS J. Detection of copy-move forgery in digital images[C]//Proceedings of Digital Forensic Research Workshop. 2003: 55-61.
 - [10] RYU S J, KIRCHNER, et al. Rotation invariant localization of duplicated image regions based on Zernike moments[J]. IEEE Transactions on Information Forensics and Security, 2013, 8(8): 1355-1370.
 - [11] HUANG H, GUO W, ZHANG Y. Detection of copy-move forgery in digital images using SIFT algorithm[C]//IEEE Pacific-Asia Workshop on Computational Intelligence and Industrial Application. 2009: 272-276.
 - [12] FARID H. Exposing digital forgeries from JPEG ghosts[J]. IEEE Transactions on Information Forensics and Security, 2009, 4(1): 154-160.
 - [13] MIKKILINENI A K, CHIANG P J, ALI G N, et al. Printer identification based on texture features[C]//NIP & Digital Fabrication Conference. Society for Imaging Science and Technology. 2004: 306-311.
 - [14] LAMPERT C H, MEI L, BREUEL T M. Printing technique classification for document counterfeit detection[C]//2006 International Conference on Computational Intelligence and Security. 2006: 639-644.
 - [15] SCHULZE C, SCHREYER M, STAHL A, et al. Using DCT features for printing technique and copy detection[C]//IFIP International Conference on Digital Forensics. 2009: 95-106.
 - [16] AHMED A G H, SHAFIT F. Forgery detection based on intrinsic document contents[C]//2014 11th IAPR International Workshop on Document Analysis Systems. 2014: 252-256.
 - [17] BATAINEH B, ABDULLAH S N H S, OMAR K. A statistical global feature extraction method for optical font recognition[C]//Asian Conference on Intelligent Information and Database Systems. 2011: 257-267.
 - [18] ZRAMDINI A, INGOLD R. Optical font recognition using typographical features[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(8): 877-882.
 - [19] BERTRAND R, TERRADES O R, GOMEZ-KRÄMER P, et al. A conditional random field model for font forgery detection[C]//2015 13th International Conference on Document Analysis and Recognition (ICDAR). 2015: 576-580.
 - [20] VAN-BEUSEKOM J, SHAFIT F, BREUEL T M. Text-line examination for document forgery detection[J]. International Journal on Document Analysis and Recognition (IJAR), 2013, 16(2): 189-207.
 - [21] CRUZ F, SIDERE N, COUSTATY M, et al. Local binary patterns for document forgery detection[C]//2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). 2017: 1223-1228.
 - [22] ABRAMOVA S. Detecting copy-move forgeries in scanned text documents[J]. Electronic Imaging, 2016, 2016(8): 1-9.
 - [23] JAMES H, GUPTA O, RAVIV D. Learning document graphs with attention for image manipulation detection[R].
 - [24] DANG H, LIU F, STEHOUWER J, et al. On the detection of digital face manipulation[C]// CVPR. 2020: 5780-5789.
 - [25] QI H, GUO Q, XU J F, et al. DeepRhythm: exposing deepfakes with attentional visual heartbeat rhythms[C]//ACM Multimedia. 2020: 4318-4327.
 - [26] IACOPO MASI, ADITYA KILLEKAR, ROYSTON MARIAN MASCARENHAS, et al. Two-branch recurrent network for isolating deepfakes in videos[C]//ECCV. 2020: 667-684.
 - [27] QIAN Y Y, YIN G J, SHENG L, et al. Thinking in frequency: face forgery detection by mining frequency-aware clues[C]//ECCV. 2020: 86-103.
 - [28] LI J M, XIE H T, LI J H, et al. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection[C]// CVPR. 2021: 6458-6467.
 - [29] HE K, GKIOXARI G, DOLL'AR P, et al. Mask r-CNN[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 2961-2969.
 - [30] HUANG Z, CHEN K, HE J, ET al. Icdar2019 competition on scanned receipt ocr and information extraction[C]//2019 International Conference on Document Analysis and Recognition (ICDAR). 2019: 1516-1520.
 - [31] WU L, ZHANG C, LIU J, et al. Editing text in the wild[C]//ACM MM. 2019:1500-1508.
 - [32] 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: 5551-5560.
 - [33] WANG X, JIANG Y, LUO Z, et al. Arbitrary shape scene text detection with adaptive text region representation[C]//CVPR. 2019: 6449-6458.

[作者简介]



王裕鑫（1996—），男，四川成都人，中国科学技术大学博士生，主要研究方向为场景文字检测与识别、人工智能。



谢洪涛（1983—），男，河南信阳人，中国科学技术大学教授、博士生导师，主要研究方向为多媒体内容安全、医学影像智能分析。



张博强（2000—），男，山西忻州人，中国科学技术大学硕士生，主要研究方向为场景文字检测与识别、人工智能。



张勇东（1973—），男，山西运城人，中国科学技术大学教授、博士生导师，主要研究方向为多媒体内容分析、网络空间安全。