中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-11

论文引用格式:Yu Yike, Ye Ziyi, Guan Heng, Wu Shenwei, Liu Chunxiao. Document Image Tampering Detection Network via Multimodal Prompt Learning[J/OL]. Journal of Image and Graphics, XXXX:1-11. DOI: 10.11834/jig.250160. (於依科, 叶子奕, 管恒, 吴沈炜, 刘春晓. 借助多模态提示词学习的文档图像篡改检测网络[J/OL]. 中国图象图形学报,XXXX:1-11. DOI: 10.11834/jig.250160.)[DOI:10.11834/jig.250160]

借助多模态提示词学习的文档图像篡改检测网络

於依科[1],叶子奕[1],管恒[1],吴沈炜[1],刘春晓[1,2]

1. 浙江工商大学计算机科学与技术学院,杭州　310018; 2. 浙江省大数据与未来电子商务技术重点实验室,杭州　310018

摘　要:**目的**　图像篡改检测是计算机视觉领域一个热点研究话题。为了减少篡改证书、证件等文档图像造成的负面影响,本文针对文档图像的篡改检测开展研究。相比于传统的自然图像篡改检测,文档图像篡改检测面临样本篡改类型的多样性、用于训练的有效篡改样本数量有限,以及篡改区域通常较小且难以识别等挑战。针对以上问题,本文提出了一种基于多模态提示词学习的文档图像篡改检测网络,同时构建了一个新的数据集。**方法**　首先,本文网络充分利用提示词学习引导机制,采用文本编码器对图像编码器进行引导的方法,以获取更具泛化性的特征表示,从而提升对困难样本的适应能力。其次,本文设计了跨模态特征对齐模块,目的是增强文本与图像之间的语义一致性,提升模型对不同类型特征的理解。此外,本文构建了自适应匹配解码器,将不同模态的特征结合并扩展至像素级,实现更精细的篡改区域定位。最后,为进一步降低过拟合风险并增强模型在不同场景下的鲁棒性,本文结合了不同损失函数以提升对比学习正则化能力。**结果**　本文在现有数据集和自制数据集上,对所提出的网络与8种代表性方法进行了实验与性能对比。实验结果表明,本文方法在多个评价指标和视觉效果上均优于现有模型。其中,在篡改区域的定位准确度方面,F1和IoU指标分别较已有方法在多个数据集上平均提升7.63%和8.42%。**结论**　本文提出的基于多模态提示词学习的文档图像篡改检测网络显著提升了检测精度,有效克服了困难样本中篡改区域识别困难以及模型泛化能力不足等挑战。

关键词:文档图像篡改检测;多模态;提示词学习;跨模态对齐;自适应解码

Document Image Tampering Detection Network via Multimodal Prompt Learning

Yu Yike[1], Ye Ziyi[1], Guan Heng[1], Wu Shenwei[1], Liu Chunxiao[1,2]

1. School of Computer Science and Technology, Zhejiang Gongshang University, Hangzhou 310018, China; 2. Zhejiang Key Laboratory of Big Data and Future E-Commerce Technology, Hangzhou 310018, China

Abstract: **Objective** Image tampering detection is a fundamental problem in the field of computer vision. To mitigate the negative impact caused by the tampering of important images such as certificates and official documents, conducting research on document image tampering detection is essential. Compared to traditional natural image tampering detection, document image tampering detection presents unique challenges, including diverse types of tampering, a limited number of

收稿日期:2025-04-04;**修回日期**:2026-05-11

基金项目:国家自然科学基金项目(U25A20440);浙江省自然科学基金项目(LY24F020004);浙江省重点研发计划资助(2023C01039);国家级大学生创新创业训练计划项目(202510353027);浙江省大学生创新创业训练计划项目(S202510353062)

Supported by: National Natural Science Foundation of China (U25A20440); Zhejiang Provincial Natural Science Foundation of China (LY24F020004); Key R&D Program of Zhejiang Province (2023C01039); National College Students Innovation and Entrepreneurship Training Program (202510353027); Zhejiang Provincial College Students Innovation and Entrepreneurship Training Program (S202510353062)

© 中国图象图形学报版权所有

paired tampered samples, and relatively small tampered regions. To address these challenges, we proposed a document image tampering detection network based on multimodal prompt learning, which effectively enhances the model's ability to capture subtle tampering traces and improves detection accuracy. **Method** First, the proposed network incorporates a prompt learning guidance mechanism, which uses a text encoder to explicitly guide the image encoder. This mechanism facilitates the extraction of more generalizable and discriminative feature representations, thereby enhancing the model's adaptability to challenging and diverse tampered samples. By integrating textual prompts into the feature extraction process, the network effectively captures high-level semantic information, improving its ability to generalize across different document tampering scenarios. Second, to further refine the interaction between textual and visual information, a cross-modal feature alignment module is introduced. This module is designed to reinforce the semantic consistency between text and image modalities, ensuring that the extracted features from both sources are well-aligned. By doing so, the model gains a more comprehensive understanding of different feature types, leading to improved tampering detection performance across various manipulation techniques. Additionally, an adaptive matching decoder is developed to integrate multimodal features and extend their representation to the pixel level. This component plays a crucial role in enhancing the granularity of tampering detection, enabling the model to localize manipulated regions with greater precision. By dynamically adjusting to the spatial and semantic characteristics of tampered areas, the decoder ensures more accurate detection results. Finally, to further reduce the risk of overfitting and enhance the model's robustness across diverse tampering scenarios, a specialized loss function is designed to strengthen the regularization capability of contrastive learning. This loss function promotes better feature discrimination and improves the model's ability to distinguish authentic and tampered regions, leading to more reliable detection outcomes in real-world applications. **Result** The effectiveness of the proposed network is thoroughly evaluated through comprehensive experiments, where it is compared against eight representative state-of-the-art methods. These evaluations are conducted on both publicly available benchmark datasets and a self-constructed dataset designed to better represent real-world document tampering scenarios. The experimental results demonstrate that the proposed approach significantly enhances the accuracy of authenticity classification, effectively distinguishing between genuine and tampered document images with greater reliability. Furthermore, in terms of tampered region localization, the proposed method achieves substantial performance gains over existing approaches. Specifically, it demonstrates an average improvement of 7.63% in the F1-score, indicating a more balanced precision and recall in detecting manipulated areas. Additionally, it achieves an 8.42% increase in Intersection over Union (IoU), highlighting its superior capability in accurately segmenting and localizing tampered regions within document images. These results confirm the robustness and effectiveness of the proposed network in improving both classification and localization performance in document image tampering detection tasks. **Conclusion** The proposed document image tampering detection network, based on multimodal prompt learning, significantly enhances detection accuracy. It effectively overcomes key challenges, such as the difficulty in recognizing tampered regions in complex or subtle manipulations and the limited generalization ability of existing models. By leveraging multimodal prompt learning, the network integrates textual and visual cues to extract more discriminative features, improving adaptability to diverse tampering patterns. Additionally, advanced cross-modal feature alignment and adaptive decoding mechanisms are incorporated to ensure precise localization of manipulated regions. Extensive experimental evaluations demonstrate superior performance over state-of-the-art methods, establishing the model as a robust and reliable solution for document image tampering detection.

Key words: document image tampering detection; multimodal; prompt learning; cross-modal alignment; adaptive decoding

论文引用格式:DOI:10.11834/jig.250160

0 引 言

图像篡改检测是机器学习、计算机视觉与信息安全交叉领域的重要研究方向,尤其在文档图像的安全性保障方面具有重要意义。为了减少篡改证书证件等文档图像造成的负面影响,本文专门针对文档图像的篡改检测技术开展研究。与传统自然图像篡改检测相比,文档图像的篡改检测存在篡改类型

© 中国图象图形学报版权所有

多样、篡改检测样本少、篡改区域较小等问题，导致模型检测性能不佳。

早期的一些传统文档图像篡改取证工作（Bilal等，2011；Beusekom等，2013；Romain等，2015）利用字体特征不一致的信息来区分真实文本和篡改文本。Amr等（2014）通过识别同源文档的模板部分来寻找文档失真的伪造线索，但该方法的篡改识别准确率不高。Cruz等（2017）利用局部二值模式（Local Binary Pattern，LBP）来统计伪造区域上的纹理特征，并结合目标文字区域的邻近区域的多个模式描述符来建模上下文信息。Shivakumara等（2018）提出了国际移动设备识别码（International Mobile Equipment Identity，IMEI）检测伪造图像的算法。Dlamini等（2018）提出利用字符识别与傅里叶梅林变换来计算字符失真参数，并评估篡改概率。Sun等（2019）利用一阶图像差异捕捉拼接边界异常来定位伪造边界。James等（2020）利用随机森林检测金融文档中的字符小幅移动和缩放操作，以此判断文字是否篡改。Nandanwar等（2021）提出将离散余弦变换（Discrete Cosine Transform，DCT）的幅值与傅里叶变换的相位结合生成相位谱，用于伪造IMEI码和机票的检测。

随着深度学习的不断发展，文档图像篡改检测研究得到了有力推动。Ahmed等人（2020）利用文档图像中的文本匹配特性，提出了一种图像匹配网络，通过自动对比相同文档的静态内容来识别因篡改导致的形变区域。Xu等（2022）提出了一种双流网络来检测和定位文档图像的伪造区域，分别从空间信息和频域信息中捕获伪造痕迹，包括边界不自然、对比度不一致等。Yang等（2022）结合深度学习和错误级别分析（Error Level Analysis，ELA）提升检测效果。James等（2022）基于文本的结构化特性，将篡改检测重新建模为节点分类任务，在内容不同但字符特征一致的文本-图像对上训练变分自编码器（Variational Auto Encoder，VAE），利用字符识别技术提取文本框，并将其视作图节点输入包含VAE、图注意层及最终解码器的网络中，以识别拼接的文本区域，但这在很大程度上依赖于解码器结构的设计，会有误检的情况。

基于注意力机制的检测方法（Lin等，2019；Chen等，2021；Jiang等，2021；Ma等，2024；Li等，2024）也较为流行。Sun等（2022）设计了一种多层特征注意力的卷积编解码网络，编码器利用空洞空间金字塔池化（Atrous Spatial Pyramid Pooling，ASPP）提取多尺度上下文信息，并结合注意力机制抑制无关特征，最终通过上采样层恢复图像空间信息，以生成高质量的篡改预测图像。Liang等（2022）从高级特征图中捕获多尺度特征，并结合注意力机制，通过同时提取低级局部特征和高级全局特征来进行篡改定位。郑等（2023）提出了一种基于注意力机制定位篡改区域算法，来增强对篡改特征的捕捉能力。Qu等（2023）提出了一种新的框架，称为文档篡改检测器（document tampering detector，DTD），由频率感知头和多视图迭代解码器组成，以补偿视觉特征不明显造成的缺陷。为了解决现有方法对于模型训练未知的伪造样本的检测困难，Qu等（2024）通过改变图像中选定文本的纹理来制作高质量的数据集，但是其数据集来源比较单一且篡改类型缺乏多样性。

总之，传统的篡改检测方法主要依赖于频域变换与失真检测，如字符特征分析、边缘分析和压缩伪影分析等。基于深度学习的篡改检测方法通过大量数据训练，捕获篡改区域与真实区域的不一致特征，增加对篡改区域的感知，从而定位篡改伪造区域。然而，这些方法在处理文档图像篡改检测任务时仍存在局限：1）难以检测纹理平滑或篡改区域过小的样本；2）训练样本单一，缺乏多样性；3）检测模型跨数据集的泛化性能不佳。

为了解决上述问题，本文提出了一种基于多模态提示词学习的检测网络，并构建了一个新的数据集。本文在网络设计方面主要考虑以下策略：利用文本编码器引导图像编码器聚焦于篡改区域，从而增强对不同尺度篡改特征的捕捉能力；通过高效的跨模态对齐机制，提高模型在不同篡改类型上的表达能力，提升跨数据集的泛化性能；引入自适应匹配解码器，使模型能够融合文本与视觉信息，进一步提升检测精度。在五个文档图像篡改数据集上的实验结果表明，本文模型相较于现有方法在检测性能和泛化能力方面均取得了显著提升。

1 本文方法

1.1 整体网络框架

本文提出的整体网络结构如图1所示，主要由CLIP文本编码器、CLIP图像编码器、跨模态对齐模

© 中国图象图形学报版权所有

块、自适应匹配解码器组成。CLIP文本编码器负责处理输入的文本提示词，将不同类别的文本信息进行嵌入，采用可学习的类别嵌入引导模型关注潜在的篡改区域。CLIP图像编码器提取文档图像的局部和全局特征，得到图像块嵌入。CLIP多模态模型架构的核心思想是利用对比学习来对齐文本和图像的语义空间，通过大规模的图文对数据集进行训练，使得相应的文本和图像在嵌入空间中更加接近，而无关的文本和图像嵌入则相互远离。跨模态对齐模块的核心作用是结合文本信息与图像特征，采用类别嵌入作为查询信息，输出跨模态融合特征，使文本信息影响视觉特征提取。自适应匹配解码器模块负责整合跨模态对齐后的特征与原始图像特征，生成最终的篡改检测结果。

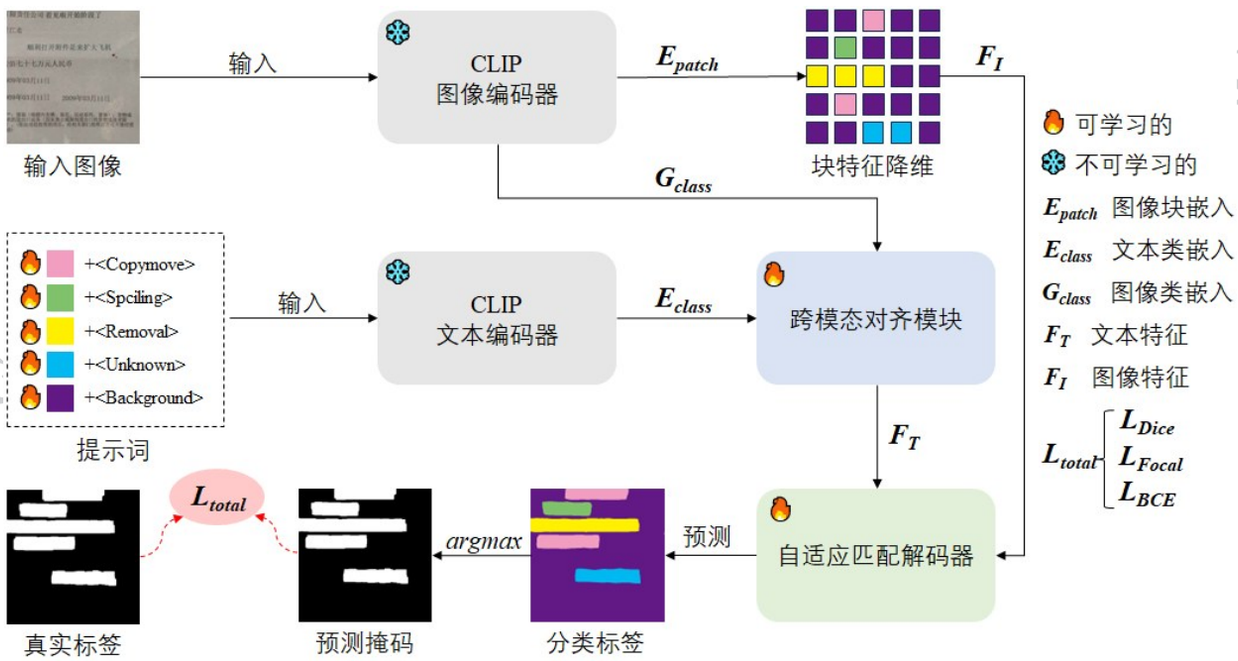


图1　整体网络框架

Fig. 1　The overall framework of our network

1.2　跨模态对齐模块

原始的图像与文本编码器是在大规模的图文对数据集上预训练的，直接应用于文档图像篡改检测任务上可能存在适配性问题。为了更好地整合模型在预训练中学到的匹配能力，并提高跨模态对齐的适应性，有效泛化到文档篡改检测任务，需要在解码器之前设计一个专门的跨模态对齐模块。具体来说，首先，本文将原始的文本类嵌入$\boldsymbol{E}_{class}$表示为所有类别的文本嵌入，其形式为：

$$\boldsymbol{E}_{class}=\left[\boldsymbol{t}^1,\boldsymbol{t}^2,\cdots,\boldsymbol{t}^i,\cdots,\boldsymbol{t}^C\right]\in\boldsymbol{R}^{C\times D}\#(1)$$

式中，$\boldsymbol{C}$表示类别总数，$\boldsymbol{D}$为嵌入维度。t^i表示$\boldsymbol{E}_{class}$中第i类文本嵌入，具体形式表示为：

$$\boldsymbol{t}^i=\left[\boldsymbol{t}_1^i,\boldsymbol{t}_2^i,\cdots,\boldsymbol{t}_j^i,\cdots,\boldsymbol{t}_D^i\right]\in\boldsymbol{R}^D\#(2)$$

式中，$\boldsymbol{t}_j^i$为文本类嵌入$\boldsymbol{E}_{class}$中第i类文本嵌入的第j个维度值。

其次，为了清晰地表示第i类文本嵌入与图像嵌入之间的匹配关系$\boldsymbol{r}^i$，将其定义成向量为：

$$\boldsymbol{r}^i=\left[\boldsymbol{r}_1^i,\boldsymbol{r}_2^i,\cdots,\boldsymbol{r}_j^i,\cdots,\boldsymbol{r}_D^i\right]\in\boldsymbol{R}^D\#(3)$$

式中，$\boldsymbol{r}_j^i$表示第i类的文本嵌入与图像嵌入在第j个维度上的匹配关系。

然后，通过计算文本类嵌入和图像类嵌入之间的匹配分数，以衡量它们的相似性，计算公式为：

$$\boldsymbol{S}_i=\sum_j^D\boldsymbol{r}_j^i=\sum_j^D\boldsymbol{t}_j^i\boldsymbol{g}_j\#(4)$$

式中，$\boldsymbol{S}_i$表示第i类文本类嵌入与图像类嵌入的匹配分数，$\boldsymbol{g}_j$为图像类嵌入$\boldsymbol{G}_{class}$中第j个维度值。

接着，为了使模型能够更精准地捕捉类别相关信息，需要保留原始的文本嵌入$\boldsymbol{E}_{class}$并结合文本与图像的匹配关系进行增强。具体而言，将第i类文本类嵌入$\boldsymbol{t}^i$与匹配关系$\boldsymbol{r}^i$进行拼接，有利于增强查询的图像适应性。于是，对于每个类的解码器输入文本查询可以表示为：

© 中国图象图形学报版权所有

$$\boldsymbol{F}_{T,i}=\boldsymbol{concat}\left[\boldsymbol{r}^i,\boldsymbol{t}^i\right]=\boldsymbol{concat}\left[\boldsymbol{t}^i\odot\boldsymbol{g},\boldsymbol{t}^i\right]\#(5)$$

式中，$\boldsymbol{F}_{T,i}$表示最终由文本编码器端输入自适应匹配解码器的第i类文本查询向量，$\odot$表示哈达玛积(Hadamard Product)，即对应元素相乘。最终，总体图像引导的文本查询向量的形状为：

$$\boldsymbol{F}_T=\left[\boldsymbol{F}_{T,1},\boldsymbol{F}_{T,2},\cdots,\boldsymbol{F}_{T,i},\cdots,\boldsymbol{F}_{T,C}\right]\in\boldsymbol{R}^{C\times 2D}\#(6)$$

式中，$\boldsymbol{F}_{T,i}$表示第i类的文本特征。每个类别的文本嵌入不仅包含了原始的文本特征$\boldsymbol{t}^i$，还结合了匹配关系$\boldsymbol{r}^i$，以确保解码器能够根据图像特征自适应调整查询信息。这种设计不仅增强了文本查询对图像特征的适应能力，还能够在解码阶段充分利用预训练的跨模态匹配知识。在完成跨模态对齐后，模型需进一步提高像素级分类能力，因此引入轻量级自适应匹配解码器，以实现高效特征匹配。

1.3 自适应匹配解码器

大规模预训练模型CLIP在图像文本匹配任务中展现出了卓越的特征感知能力，但是像素级分类任务则需要更精细的局部特征对齐，因此如何有效地将上游模型的泛化能力迁移到图像篡改检测任务仍然是一个挑战。为此，本文设计了一种以轻量级Transformer为主体的自适应匹配解码器，如图2所示，其核心目标是在单阶段框架下以较低的计算开销高效利用CLIP编码器特征进行像素级分类。此外，还将篡改检测任务建模为一个匹配问题，使得类别查询能够与图像patch特征进行自适应匹配，以实现准确分类。

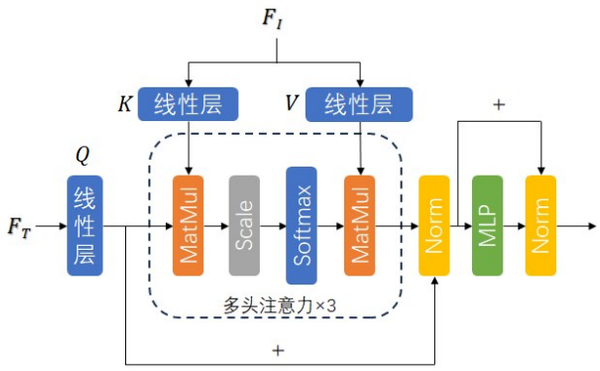


图2 自适应匹配解码器

Fig. 2 Adaptive matching decoder

本文方法基于CLIP的文本类嵌入与图像patch特征进行匹配，以实现像素级分类。图像块嵌入$\boldsymbol{E}_{patch}$在通过1×1卷积逐通道降维操作后获得图像特征F_I定义如下：

$$\boldsymbol{F}_I=\left[\boldsymbol{F}_{I,1},\boldsymbol{F}_{I,2},\cdots\boldsymbol{F}_{Ij},\cdots,\boldsymbol{F}_{I,C}\right]\in\boldsymbol{R}^{C\times D}\#(7)$$

式中，$\boldsymbol{F}_{I,j}$表示第j个图像块的特征向量，C表示块总数，D为嵌入维度。为了建立文本类嵌入与图像块之间的交互，采用线性投影映射层$\varnothing$，生成查询$\boldsymbol{Q}$、键$\boldsymbol{K}$和值$\boldsymbol{V}$的矩阵，具体定义如下：

$$\boldsymbol{Q}=\varnothing_q\left(\boldsymbol{F}_T\right)\in\boldsymbol{R}^{C\times D}\#(8)$$

$$\boldsymbol{K}=\varnothing_k\left(\boldsymbol{F}_I\right)\in\boldsymbol{R}^{N\times D}\#(9)$$

$$\boldsymbol{V}=\varnothing_v\left(\boldsymbol{F}_I\right)\in\boldsymbol{R}^{N\times D}\#(10)$$

式中，$\varnothing_q,\varnothing_k,\varnothing_v$分别为$\boldsymbol{Q}$、$\boldsymbol{K}$和$\boldsymbol{V}$的线性变换函数。在获得$\boldsymbol{Q}$和$\boldsymbol{K}$矩阵后，采用多头注意力机制来计算匹配掩码$\boldsymbol{M}$，具体计算公式如下：

$$\boldsymbol{M}=\frac{\boldsymbol{QK}^T}{\sqrt{\boldsymbol{d}_k}}\in\boldsymbol{R}^{C\times N}\#(11)$$

式中，$\sqrt{d_k}$为键$\boldsymbol{K}$的维度，作为比例因子来稳定梯度计算，防止数值过大。匹配掩码$\boldsymbol{M}$的值越高表示匹配程度越高。随后，在掩码矩阵的类别维度上进行$Argmax$操作：

$$\hat{y}=Argmax\left(M,dim=0\right)\#(12)$$

选择最大响应类作为背景类，其余类别定义为篡改区域，最终得到类别预测结果这一操作确保每个图像块都被分配到最可能的类别，最终得到篡改检测结果。

1.4 损失函数

在文档图像篡改检测任务中，模型通常被设计为像素级二分类问题。本文采用基于Sigmoid激活函数的二值交叉熵损失，减少类别间的互斥影响，计算公式如下：

$$L_{BCE}=-\frac{1}{hw}\sum_{i=1}^{hw}y_i\log p_i+\left(1-y_i\right)\log\left(1-p_i\right)\#(13)$$

式中，p_i为模型预测的像素级概率，y_i为真实标签，hw为图像总像素数。本文引入Focal Loss，对难分类的困难样本赋予更高的损失权重，从而增强模型对篡改区域的关注度，其定义如下：

$$L_{Focal}=-\frac{1}{hw}\sum_{i=1}^{hw}\frac{\left[\left(1-p_i\right)^m y_i\log p_i\right.}{\left.+p_i^m\left(1-y_i\right)\log\left(1-p_i\right)\right]}\#(14)$$

式中，m为调节因子，实验中设定$m=2$。本文进一步使用Dice Loss以优化区域级别的分类性能，其定义如下：

$$L_{Dice}=1-\frac{2\sum_{i=1}^{hw}p_i y_i}{\sum_{i=1}^{hw}p_i^2+\sum_{i=1}^{hw}y_i^2}\#(15)$$

© 中国图象图形学报版权所有

式中，p_i为模型预测的像素级概率，y_i为真实标签。最终，本文将上述三种损失进行加权组合，得到如下的最终损失函数L_{total}：

$$L_{total} = \alpha L_{Focal} + \beta L_{Dice} + \gamma L_{BCE} \#(16)$$

式中，α,β,γ为权重系数，用于平衡不同损失项的影响。

2 实验与分析

2.1 实验环境与参数设置

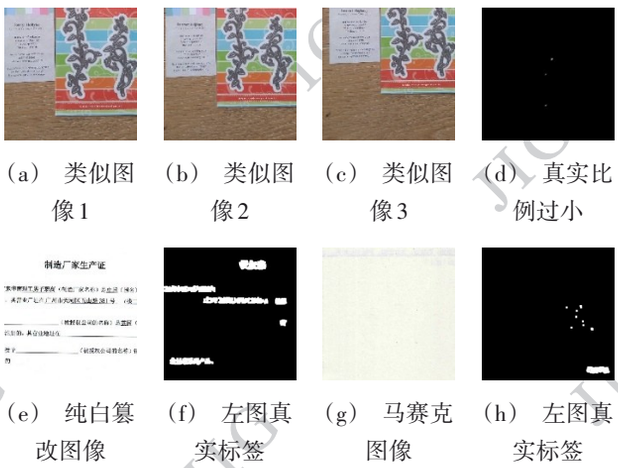
本文在Ubuntu系统上借助PyTorch(1.13.1)实现了这个模型，使用两块NVIDIA RTX3090 GPU进行训练，在每轮训练后，保留损失函数最小的模型为最佳模型。本文设计的实验基于预训练CLIP的VIT-B-16模型，在训练过程中使用了Adam优化器，其两个动量参数分别为β_1=0.9和β_2=0.999。

2.2 数据集与评价指标

2.2.1 数据集

现有数据集包括SACP、RIFLC、TTI，都为阿里云平台的用于竞赛的数据集。SACP数据集有2005张篡改图像，RIFLC数据集有4000张篡改图像，TTI有19000张篡改图像，主要模拟真实攻击手段，包括替换、裁剪、遮盖等。然而，在以上数据集中存在一些问题。如图3(a-c)，出现过多重复的样本，疑似是单个图像随机裁剪增广导致的，SACP数据集中约17%的样本为重复图像。如图3(d)，存在真实标签比例过小的样本，难以训练模型。如图3(e-f)所示，纯白背景的样本且篡改内容仅涉及对白色像素的修改。图3(g-h)展示了在马赛克图像上再次添加马赛克。

为了解决有效样本不足的情况并增强模型的泛化能力，本文自制新的数据集HTD(hybrid tampering dataset)来对模型进行训练与测试。首先，选取带有文本和图像的数据集作为自制篡改数据集的素材来源。其次，利用光学字符识别模型进行文本检索和定位，并对文本位置进行标记。接着，设定文本检测的置信度阈值，对于置信度超过阈值的文本，通过内容相似度进行匹配，选择相似度高的样本进行篡改操作，对于置信度低于阈值的文本图像，采用随机抖动等方式进行数据增广。最后，为了提升模型的鲁棒性，在篡改样本中添加噪声并调整图像风格，进一步提升模型对困难样本的适应能力。



(a) 类似图像1　(b) 类似图像2　(c) 类似图像3　(d) 真实比例过小

(e) 纯白篡改图像　(f) 左图真实标签　(g) 马赛克图像　(h) 左图真实标签

图3 现有数据集中存在的问题

Fig. 3 Problems existed in existing datasets

图4 本文制作的混合篡改数据集样本

Fig. 4 Some samples from our hybrid tampering dataset



篡改图像　真实标签　篡改图像　真实标签

按照以上篡改样本制作流程，本文共制作了15000张样本，如图4所示，覆盖了复制粘贴、拼接、移除等常见篡改方式。本文实验部分使用SACP、RIFLC、TTI、HTD和DocTamper五个数据集进行网络模型的训练和测试，其中SACP、RIFLC、TTI和DocTamper数据集经过清洗，去除了脏样本，而HTD数据集为自制数据集，其基本信息如表1所示。

2.2.2 评价指标

本文采用精确率(Precision)、召回率(Recall)、F1分数、IoU分数作为篡改检测的评价指标(Li等，

© 中国图象图形学报版权所有

表1　本文HTD数据集基本信息

Table 1　Basic information of our HTD dataset

HTD数据集		图像数量
数据集来源	ICDAR竞赛未篡改图像	4285
	CHIP2022发票数据集	1000
	SCID中文票据数据集	19999
	MODD多媒体文档图像	3000
篡改类型	复制粘贴	5000
	拼接移除	5000
	生成式篡改	3000
数据集划分	训练集	13000
	测试集	2000

2021)，以客观全面地评估模型在图像篡改检测任务上的性能表现，进而为模型的优化提供指导。

2.3　对比实验

本文选用的8种用于性能比较的网络模型分别是CAT-Net(Kwon等，2018)、RRU-Net(Xiu等，2019)、MantraNet(Yue等，2019)、MVSS-Net(Liu等，2022)、Lseg(Li等，2022)、MFAN(Ding等，2023)、PidNet(Xu等，2023)、MSU-Net(Hao等，2024)。各网络模型在五个数据集上的定量比较如表2所示。不难看出，本文提出的方法在五个数据集上的两项评价指标均为最高，明显优于两分支的网络框架(如MVSS-Net、CAT-Net)和端到端的模型(如RRU-Net、MantraNet、MFAN等)，具有较强的稳定性。尽管基于多模态的Lseg模型表现出较高性能，但与本文方法比较仍存在不足，这表明本文提出的方法在融合文本与图像信息、提取图像特征以及处理复杂场景上具有显著优势。即使在多个模型检测准确率都较低的TTI数据集上，本文方法仍展现出优异的性能。

表2　不同模型在五个数据集上的定量比较

Table 2　Quantitative comparison of different models on five datasets

方法	SACP		RIFLC		TTI		HTD		DocTamper	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
CAT-Net	0.166	0.139	0.127	0.102	0.162	0.221	0.201	0.135	0.553	0.564
RRU-Net	0.625	0.581	0.651	0.642	0.539	0.347	0.334	0.406	0.467	0.459
MantraNet	0.243	0.308	0.408	0.415	0.239	0.303	0.351	0.350	0.209	0.226
MFAN	0.677	0.654	0.636	0.642	0.474	0.352	0.376	0.372	0.572	0.594
Lseg	0.673	0.719	0.711	0.709	0.537	0.465	0.470	0.452	**0.608**	0.601
MVSS-Net	0.573	0.673	0.692	0.701	0.447	0.332	0.347	0.324	0.424	0.457
PidNet	0.654	0.603	0.633	0.630	0.328	0.420	0.420	0.410	0.462	0.432
MSU-Net	0.703	0.638	0.712	0.727	0.572	0.621	0.422	0.420	0.449	0.456
本文结果	**0.771**	**0.792**	**0.739**	**0.740**	**0.627**	**0.639**	**0.590**	**0.558**	0.605	**0.610**

注：加粗字体为最优值。

从图5的视觉效果比较可以看出，本文提出的模型能够准确识别被篡改的区域。在背景纹理较为平滑、篡改区域较为细小的情况下，模型依然能够有效检测，不受复杂背景和微小篡改的干扰。这表明该模型在篡改检测任务中具有较强的泛化性和适应性，能够应对多种复杂场景。

2.4　消融实验

为了验证所提出设计的有效性，本文从三个方面进行消融实验的比较：

(1)不同文本图像对齐模式的影响：在多模态提示词学习的框架下，文本与图像对齐的模式会影响模型对篡改区域的关注程度，如表3所示。为了方便书写，在表3中将文本类嵌入$\boldsymbol{E}_{class}$与图像类嵌入$\boldsymbol{G}_{class}$分别简化表示为t与g。不难看出，使用$t\odot g$，t格式的模型整体上展示出最优的表现，在跨模态交互时更聚焦于语义信息。

(2)不同损失函数权重的影响：损失函数的设计对模型训练的收敛速度和最终性能具有关键影响，

© 中国图象图形学报版权所有

图5　不同网络模型的效果比较

Fig. 5　Visual comparison of different models

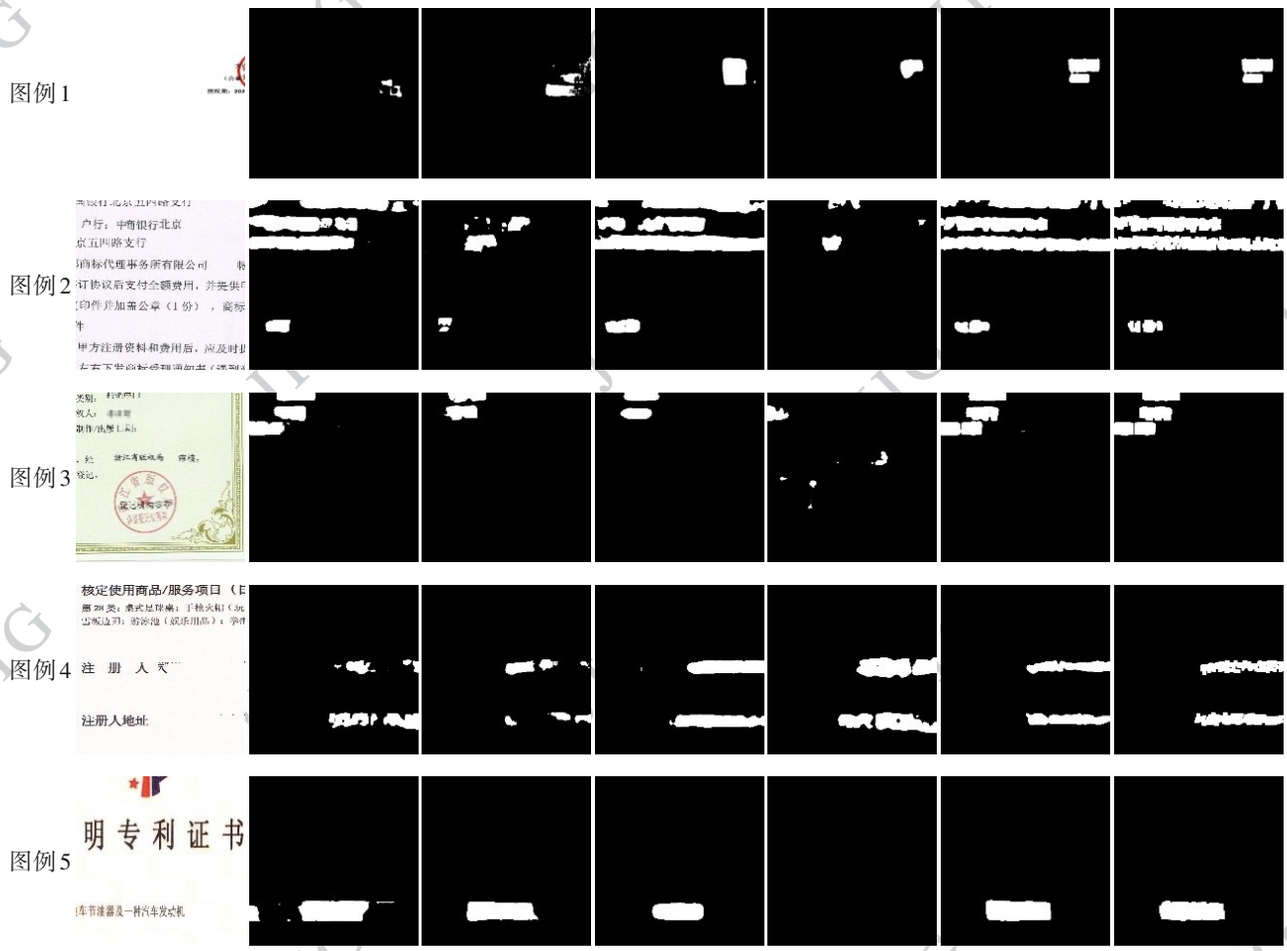


表3　不同对齐模式在SACP和RIFLC数据集上的定量比较

Table 3　Quantitative comparison of different alignment modes on SACP and RIFLC datasets

对齐模式	维度	SACP		RIFLC	
		F1	IoU	F1	IoU
t		0.506	0.612	0.685	0.506
$t \odot g$	512	0.702	0.672	0.725	0.707
$t + g$		0.653	0.575	0.505	0.531
t,g		0.631	0.663	0.571	0.628
$t \odot g,t$	512*2	**0.771**	**0.792**	0.709	**0.758**
$t + g,t$		0.540	0.595	**0.729**	0.684

注：加粗字体为最优值

如表4所示。表中结果表明，不同的损失函数权重α,β,γ对模型性能有显著影响，当$\alpha = 0.2,\beta = 0.6,\gamma = 0.2$时，SACP和RIFLC数据集上的F1和IoU指标均达到最高，说明提高Dice Loss的权重有助于优化目标区域完整性，提高检测性能。

表4　不同α,β,γ取值在SACP和RIFLC数据集上的定量比较

Table 4　Quantitative comparison of different α,β,γ values on SACP and RIFLC datasets

参数设置	SACP		RIFLC	
	F1	IoU	F1	IoU
$\alpha = 0.4,\beta = 0.2,\gamma = 0.4$	0.753	0.763	0.722	0.727
$\alpha = 0.2,\beta = 0.4,\gamma = 0.4$	0.721	0.698	0.691	0.687
$\alpha = 0.4,\beta = 0.4,\gamma = 0.2$	0.741	0.712	0.723	0.731
$\alpha = 0.2,\beta = 0.6,\gamma = 0.2$	**0.771**	**0.792**	**0.739**	**0.740**

注：加粗字体为最优值

© 中国图象图形学报版权所有

（3）不同模块的影响：各模块的设计差异是影响模型训练收敛速度与最终性能的关键因素，如表5所示。从表格看出，自适应匹配解码器在两个数据集上的F1和IoU表现上均优于一般分割网络。

表5　自适应匹配解码器在SACP和RIFLC数据集上的优势

Table 5　Advantages of Adaptive Matching Decoder on SACP and RIFLC datasets

网络差异	SACP		RIFLC	
	F1	IoU	F1	IoU
分割网络	0.737	0.728	0.692	0.687
自适应匹配解码器	**0.771**	**0.792**	**0.739**	**0.740**

注：加粗字体为最优值

2.5　鲁棒性分析

为了模拟文档图像在实际传输过程中可能遭受的攻击，本文在RIFLC数据集上对两种模型进行了JPEG压缩处理、高斯噪声、缩放等操作，以模拟图像质量下降的不同方式。如表6所示，面对模拟攻击时，在轻度条件下（如JPEG参数90、高斯噪声10、随机裁剪比例0.9）所提出的模型仍能保持较高检测效果。整体来看，本文模型在多种干扰条件下均表现出较强的鲁棒性，能够较好地适应不同的攻击方式，保证检测性能的稳定性。

表6　本文方法和Lseg模型在RIFLC数据集上的鲁棒性分析

Table 6　The robustness analysis of the proposed method and the Lseg model on the RIFLC dataset

攻击方式	参数	本文方法		Lseg	
		F1	IoU	F1	IoU
无后处理	–	0.739	0.740	0.711	0.709
JPEG压缩	90	0.685	0.721	0.691	0.680
	80	0.650	0.697	0.670	0.649
	70	0.592	0.612	0.582	0.601
高斯噪声	10	0.713	0.690	0.682	0.677
	20	0.653	0.648	0.617	0.628
	30	0.627	0.579	0.584	0.573
缩放	0.9	0.712	0.725	0.679	0.670
	0.7	0.688	0.691	0.642	0.637
	0.5	0.647	0.652	0.583	0.594

3　结　论

本文针对当前文档图像篡改检测领域篡改手段多样化、后处理操作复杂化、公开标注数据集稀缺的核心挑战，提出了一种借助多模态提示词学习的篡改检测网络，构建了包含7类常见篡改方式的多模态语义标注数据集，通过图像编码器引导文本编码器生成语义匹配的提示词，结合跨模态特征对齐模块与自适应匹配解码器实现了精细的篡改区域定位；实验结果表明，该方法在5个主流数据集上的F1分数和IoU指标分别实现了7.63%和8.42%的平均提升，在复杂后处理困难样本和跨数据集泛化场景下的表现均显著优于传统手工特征方法、单模态深度学习方法及其他多模态检测方法，其核心优势在于通过自适应提示词生成机制实现了视觉信息与文本语义的双向互补。未来，可进一步探索轻量化多模态模型设计、大规模生成式伪造数据集构建、增量式提示词学习方法，并将该技术扩展至多语言文档与视频文档篡改检测领域。

参考文献（References）

Ahmed B, Gulliver T and Alzahir S. 2020. Image splicing detection using Mask-RCNN. Signal Image and Video Processing, 14(5): 1035-1042 [DOI: 10.1007/s11760-020-01636-0]

Amr G and Faisal S. 2014. Forgery detection based on intrinsic document contents//Proceedings of the 11th IAPR International Workshop on Document Analysis Systems. Tours, France: IEEE: 252-256 [DOI: 10.1109/DAS.2014.26]

Beusekom J, Faisal S and Thomas M. 2013. Text-line examination for document forgery detection. International Journal on Document Analysis and Recognition, 16(2): 189-207 [DOI: 10.1007/s10032-011-0181-5]

Bilal B, Siti N and Khairudin O. 2011. A statistical global feature extraction method for optical font recognition//Proceedings of the Asian Conference on Intelligent Information and Database Systems. Daegu, Korea, Springer: 257-267 [DOI: 10.1007/978-3-642-20039-7_26]

Bi X L, Wei Y, Xiao B and Li W S. 2019. RRU-Net: The ringed residual u-net for image splicing forgery detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 30-39 [DOI: 10.1109/CVPRW.2019.00010]

Chen X R, Dong C B, Ji J Q, Cao J and Li X R. 2021. Image manipula-

© 中国图象图形学报版权所有

tion detection by multi-view multi-scale supervision// Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: #1392 [DOI: 10.1109/ ICCV48922.2021.01392]

Dong C B, Chen X, Hu R X, Cao J and Li X L. 2022. MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 3539-3553 [DOI: 10.1109/CVPR.2022.3180556]

Chen B J, Tan W J, Coatrieux G, Zheng Y H and Shi Y Q. 2020. A serial image copy-move forgery localization scheme with source/target distinguishment. IEEE Transactions on Multimedia, 23: 3506-3617 [DOI: 10.1109/TMM.2020.3026868]

Cruz F, Sidere N, Coustaty M, Andecy V and Ogier J M. 2017. Local binary patterns for document forgery detection//Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE: 1223-1228 [DOI: 10.1109/ICDAR.2017.202]

Dlamini N, Mthethwa S and Barbour G. 2018. Mitigating the challenge of hardcopy document forgery//Proceedings of the International Conference on Advances in Big Data, Computing and Data Communication Systems. Durban, South Africa: IEEE: 1-6 [DOI: 10.1109/ICABCD.2018.8465401]

Hao Y, Su L C, Dai C W and Wang J L. 2024. MSU-Net: the multi-scale supervised U-Net for image splicing forgery localization. Pattern Analysis and Applications, 27(3): #86 [DOI: 10.1007/S10044-024-01305-9]

James H, Gupta O and Raviv D. 2022. Learning document graphs with attention for image manipulation detection//Proceedings of the International Conference on Pattern Recognition and Artificial Intelligence. Paris, France: Springer: 263-274 [DOI: 10.1007/978-3-031-09037-0_22]

Jiang X Y and Liu C X. 2021. Edge and region inconsistency-guided image splicing tamper detection network. Journal of Image and Graphics, 26(10): 2411-2420 (蒋小玉, 刘春晓. 2021. 边缘与区域不一致性引导下的图像拼接篡改检测网络. 中国图象图形学报, 26(10): 2411-2420) [DOI: 10.11834/jig.200298]

Kwon M J, Yu I J, Nam S H and Lee H K. 2021. CAT-Net: Compression artifact tracing network for detection and localization of image splicing//Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 375 384 [DOI: 10.1109/WACV48630.2021.00042]

Luo D L, Liu Y L, Yang R, Liu X J, Zeng J S, Zhou Y and Bai X. 2025. Toward real text manipulation detection: New dataset and new solution. Pattern Recognition, 157: 110828. [DOI: 10.1016/J.PATCOG.2024.110828]

Liang W, Dong L and Wang R. 2022. Robust document image forgery localization against image blending//Proceedings of the IEEE International Conference on Trust, Security and Privacy in Computing and Communications. Shenyang, China: IEEE: 810-817 [DOI: 10.1109/TrustCom56396.2022.00113]

Li, B, Weinberger K Q, Belongie S J, Koltun V and Ranftl R. 2025. Language-driven semantic segmentation[EB/OL]. [2025-3-14] https://arxiv.org/pdf/2201.03546.

Li S, Ma W, Guo J, Xu S, Li B and Zhang X. 2024. UnionFormer: Unified-learning transformer with multi-view representation for image manipulation detection and localization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12523-12533 [DOI: 10.1109/cvpr52733.2024.01190]

Lin C, Lu W, Huang X C, Liu K, Sun W, and Lin H H. 2019. Region duplication detection based on hybrid feature and evaluative clustering. Multimedia Tools and Applications, 78: 20739-20763 [DOI: 10.1007/ s11042-019-7342-9]

Ma X, Du B, Liu X, Hammadi A and Zhou J. 2023. IML-ViT: Image manipulation localization by vision transformer[EB/OL]. [2024-10-10] https://arxiv.org/pdf/2307.14863

Nandanwar L, Shivakumara P, Kanchan S, Basavaraja V, Guru D S, Pal U and Lu T. 2021. DCT-phase statistics for forged IMEI numbers and air ticket detection. Expert Systems with Applications, 164: 114014 [DOI: 10.1106/j.eswa.2020.114014]

Qu C, Liu C, Liu Y, Chen X, Peng D, Guo F and Jin L. 2023. Towards robust tampered text detection in document image: New dataset and new solution//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 5937-5946 [DOI: 10.1109/CVPR52729.2023.00575]

Qu C, Zhong Y, Guo F and Jin L. 2024. Generalized tampered scene text detection in the era of generative AI[EB/OL]. [2024-10-10] https://arxiv.org/pdf/2407.21422

Romain B, Oriol R, Petra G, Patrick F and Jean-Marc O. 2015. A conditional random field model for font forgery detection//Proceedings of 13th International Conference on Document Analysis and Recognition. Nancy, France: IEEE: 576 - 580 [DOI: 10.1109/ICDAR.2015.7333827]

Sun Y, Ni R and Zhao Y. 2022. MFAN: Multi-level features attention network for fake certificate image detection. Entropy, 24(1): #118 [DOI: 10.3390/e24010118]

Sun K, Cao G, Zhao Q and Zhang J L. 2019. Differential abnormality-based tampering detection in digital document images//Proceedings of the 18th International Conference on Computer and Information Science. Beijing, China: IEEE: 145-149 [DOI: 10.1109/ICIS46139.2019.8940190]

Shivakumara P, Basavaraja V, Gowda H S, Guru D, Pal U and Lu T. 2018. A new RGB based fusion for forged IMEI number detection in mobile images//Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition. NewYork, USA: IEEE: 386-391 [DOI: 10.1109/ICFHR-2018.2018.00074]

© 中国图象图形学报版权所有

Tahira N，Marriam N，Momina M and Ali J. 2022. Copy move forgery detection and segmentation using improved mask region-based convolution network. Applied Soft Computing，131：#109778［DOI：10.1016/j.asoc.2022.109778］

Wu Y，Abd-Almageed W and Natarajan P. 2019. ManTra-Net：Manipulation tracing network for detection and localization of image forgeries with anomalous features//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach，USA：IEEE：9535-9544［DOI：10.1109/CVPR.2019.00977］

Xu W，Luo J and Zhu C. 2022. Document images forgery localization using a two-stream network. International Journal of Intelligent Systems，37(8)：5272-5289［DOI：10.1002/int.22792］

Xu J C，Xiong Z X，Shankar P. Bhattacharyya. 2023. PIDNet：A real-time semantic segmentation network inspired by PID controllers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver，Canada：IEEE：19529-19539［DOI：10.1109/CVPR.52729.2023.01871］

Yang P，Fang W，Zhang F，Bai L F and Gao Y Y. 2022. Document image forgery detection based on deep learning models//Proceedings of the International Symposium on Electrical，Electronics and Information Engineering. Chiang Mai，Thailand：IEEE：36-41［DOI：10.1109/ISEEIE55684.2022.00014］

Zhen J T，Li B and Zeng J H. 2023. Document image tampering and desensitization localization based on attention mechanism. Journal of Xidian University，50(6)：207-218（郑锫涛，李斌，曾锦华. 应用注意力机制的文档图像篡改与脱敏定位. 西安电子科技大学学报，50(6)：207-218）［DOI：10.19665/j.issn1001-2400.20230105］

作者简介

於依科，1999年生，男，硕士研究生，主要研究方向为图像篡改检测。E-mail：759974498@qq.com

叶子奕，2005年生，女，本科生，主要研究方向为机器学习与计算机视觉。E-mail：15715762821@163.com

管恒，2004年生，男，本科生，主要研究方向为机器学习与计算机视觉。E-mail：hengheng0813@qq.com

吴沈炜，2005年生，男，本科生，主要研究方向为图像篡改检测与深度伪造检测。E-mail：2034563169@qq.com

刘春晓，通讯作者，男，副教授，硕士生导师，主要研究方向为机器学习与计算机视觉、视觉安全与隐私保护。E-mail：cxliu@mail.zjgsu.edu.cn

© 中国图象图形学报版权所有