中图法分类号:TP391.41　　**文献标识码**:A　　**文章编号**:1006-8961(XXXX)XX-0001-15

论文引用格式:Zeng Jianhong, Zhou Xinyue, Jiang Nanfeng, Wang Dahan. Dynamic receptive aggregation network for document image rectification under complex deformation[J/OL]. Journal of Image and Graphics, XXXX:1-15. DOI: 10.11834/jig.260432.(曾健烘, 周欣玥, 江楠峰, 王大寒. 面向复杂形变的文档图像动态感受聚合校正网络[J/OL]. 中国图象图形学报, XXXX:1-15. DOI: 10.11834/jig.260432.)[DOI:10.11834/jig.260432]

面向复杂形变的文档图像动态感受聚合校正网络

曾健烘,周欣玥,江楠峰*,王大寒

厦门理工学院计算机与信息工程学院,福建　厦门　361024　福建省模式识别与图像理解重点实验室,厦门　361024

摘　要:**目的**　针对复杂形变文档图像中局部几何恢复不足、背景交互冗余和通道响应重复等问题,提出动态感受聚合网络(dynamic receptive aggregation network,DRA-Net),以提高文档图像校正(document image rectification,DIR)的几何恢复质量。本文聚焦复杂形变条件下的页面几何结构恢复,并在公开测试集上进行定量评估。**方法**　DRA-Net采用层次化编码器与解码器结构,由动态感受聚合卷积(dynamic receptive aggregation convolution,DRAConv)、层次稀疏选择注意力(hierarchical sparse selective attention,HSSA)和显著性感知特征精炼模块(saliency-aware feature refinement module,SAFRM)构成。DRAConv通过通道重加权、感受野响应重排和空间调制聚合局部几何特征;HSSA将稀疏响应分支与稠密上下文分支按可学习比例融合,以兼顾选择性交互和结构信息传播;SAFRM结合局部卷积与门控调制精炼通道响应。网络通过反向映射监督、像素重建约束和结构相似性约束联合训练。**结果**　在DocUNet和DIR300完整公开测试集上与7种代表性方法比较,DRA-Net在局部失真(local distortion,LD)和对齐失真(aligned distortion,AD)两项几何指标上均取得本次对比的最佳结果。在DIR300上,LD和AD分别为5.5687和0.1911,相较次优结果分别降低约12.4%和19.4%。模块消融及损失权重实验进一步验证了所采用配置的有效性。在统一测试条件下,单图推理耗时为31.434 ms,峰值显存占用为885.06 MB。**结论**　DRA-Net通过局部特征聚合、空间交互选择和通道响应精炼的协同配置改善复杂文档的几何校正质量。其优势主要体现在本次公开基准的几何指标上,并不意味着在全部光学字符识别(optical character recognition,OCR)指标、理论计算量或商业系统整体性能上全面领先。未来的工作将进一步优化文本识别适配性与模型计算开销。

关键词:文档图像校正;动态感受聚合网络;动态感受聚合卷积;层次稀疏选择注意力;显著性感知特征精炼

Dynamic receptive aggregation network for document image rectification under complex deformation

Zeng Jianhong, Zhou Xinyue, Jiang Nanfeng*, Wang Dahan

School of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, Fujian, China

Abstract: **Objective** Document image rectification (DIR) aims to restore the geometric layout of camera-captured documents affected by page curvature, perspective distortion, local folding, and nonuniform deformation. It is an important preprocessing step for optical character recognition (OCR), document analysis, and digital archiving. Commercial scanning applications already provide useful capture, cropping, enhancement, and page-straightening functions. Their availability, however, does not remove the need to investigate geometric accuracy under controlled evaluation conditions. Complex

收稿日期:2026-07-16;**修回日期**:2026-10-25

*通信作者:江楠峰,男,博士,副教授,主要研究方向为计算机视觉,图像处理。E-mail: jnfrock@gmail.com

基金项目:国家自然科学基金资助项目(62501506,62576301)

Supported by: National Natural Science Foundation of China (Nos. 62501506, 62576301)

© 中国图象图形学报版权所有

pages may contain spatially varying distortions, irregular boundaries, and interference from surrounding objects. A visually acceptable scan does not by itself establish the accuracy of local geometry or the reliability of subsequent recognition. This work therefore focuses on improving the geometric restoration of complex documents using a reproducible learning-based model. The proposed dynamic receptive aggregation network (DRA-Net) addresses three representation problems: insufficient local deformation modeling, redundant spatial interactions, and repeated channel responses. Its objective is not to claim universal superiority over commercial systems, but to investigate a task-oriented combination of feature aggregation and selective interaction for backward-mapping prediction. **Method** DRA-Net uses a hierarchical encoder-decoder architecture with skip connections. An input projection first produces overlapping image features. Dynamic receptive aggregation convolution (DRAConv) then expands local receptive-field responses through grouped convolution, rearranges these responses into an enlarged spatial representation, and applies channel and spatial modulation before restoring the feature resolution. The rearrangement is a deterministic indexing operation rather than a learned displacement of sampling positions. Its role is to expose individual local responses to subsequent weighting. The input-dependent modulation coefficients provide the adaptive part of this operation. Hierarchical sparse selective attention (HSSA) combines a dense contextual branch with a sparse-response branch. The dense branch uses softmax normalization, whereas the sparse branch uses a squared rectified linear unit to set nonpositive attention scores to zero. Two learned scalar parameters determine the normalized fusion proportions. These parameters are optimized during training and are fixed during inference; the attention responses, rather than the scalar fusion proportions, change with the input. Sparsity in this design refers to response selection in one branch. It does not imply a sparse implementation of matrix multiplication or a reduction in the asymptotic complexity of attention. The saliency-aware feature refinement module (SAFRM) further processes the features through a convolutional projection, channel splitting, local depthwise convolution, and multiplicative gating. This configuration introduces local context into channel refinement and modulates redundant responses. Saliency here denotes learned feature-response selection, not a separately supervised saliency map. Receptive-field attention and sparse-dense attention fusion have antecedents in prior work; the present study examines their task-oriented configuration for document geometry rather than treating every constituent operation as an entirely new primitive. The network predicts a backward mapping and obtains the rectified image by bilinear sampling. Training combines direct supervision of the backward mapping, pixel reconstruction error, and a structural similarity index (SSIM) loss. The corresponding loss weights are 8, 0.1, and 0.06. **Result** Training uses the Doc3D dataset, and evaluation is performed on the complete public DocUNet and DIR300 test sets. Seven representative methods are included: DewarpNet, DocTr, DocGeoNet, FAM, UVDoc, DocRes, and DocScanner. The evaluation covers multi-scale structural similarity index (MS-SSIM), local distortion (LD), aligned distortion (AD), edit distance (ED), and character error rate (CER). DRA-Net obtains the lowest LD and AD among the compared methods on both test sets. On DIR300, its LD and AD are 5.5687 and 0.1911, respectively, approximately 12.4% and 19.4% lower than the corresponding second-best results. Relative to DocScanner, the reductions in LD are approximately 16.8% on DocUNet and 28.43% on DIR300. These results support an improvement in geometric restoration within the reported evaluation protocol. They do not imply that geometric improvement automatically produces the best recognition accuracy: DocScanner has lower ED and CER in the reported comparisons. Qualitative examples show the behavior of the methods in regions containing curved text lines and irregular page boundaries. Module-level and component-level ablations provide additional evidence. Replacing DRAConv with plain convolution increases LD from 5.5687 to 6.4224 and AD from 0.1911 to 0.2681. Removing receptive-field rearrangement increases LD to 7.3022. In HSSA, either attention branch alone and a fixed equal-weight fusion produce worse overall results than the learned fusion configuration; fixed fusion gives an LD of 5.8656 and an ED of 813.27. Removing both the local convolution and gate in SAFRM increases LD to 9.0241 and AD to 0.4232. Loss experiments likewise support the combined supervision, although the default weights are not optimal for CER among all tested settings. These ablations establish the utility of the tested configuration, not superiority over every related operator. Computational analysis reports 28.785 million parameters, 141.680 billion floating-point operations, 31.434 ms per forward inference, and 885.06 MB peak memory under the stated test conditions. DRA-Net has a latency close to, but slightly higher than, DocScanner and a lower measured memory peak. Its floating-point operation count is higher than those of all methods in the comparison, so low theoretical complexity is not claimed. **Conclu-**

© 中国图象图形学报版权所有

sion DRA-Net improves complex document rectification through a task-oriented combination of local receptive-field aggregation, selective spatial interaction, and channel refinement. The reported experiments support its geometric advantages on two public benchmarks while identifying limitations in recognition performance and theoretical computational cost. The study distinguishes a research model for geometric prediction from a complete scanning product, whose overall behavior also depends on acquisition, enhancement, recognition, and deployment choices. No direct ranking against commercial applications is established because matched product-level experiments were not conducted. Further work will investigate text-aware supervision, more efficient feature aggregation, and reproducible evaluation under additional capture conditions. Improving the balance between local geometric accuracy, recognition reliability, and deployment cost remains the central objective.

Key words: document image rectification; dynamic receptive aggregation network; dynamic receptive aggregation convolution; hierarchical sparse selective attention; saliency-aware feature refinement

论文引用格式：曾健烘，周欣玥，江楠峰，王大寒. 2026. 面向复杂形变的文档图像动态感受聚合校正网络. 中国图象图形学报. DOI: 10.11834/jig.260432.

0 引　言

文档图像校正(document image rectification, DIR)旨在将包含卷曲、褶皱、起伏和透视畸变的文档图像恢复为平整、可读的页面，是光学字符识别(optical character recognition, OCR)、数字档案构建、移动端文档采集和自动化文档分析的重要前置步骤。真实场景下的文档通常同时伴随局部拉伸、复杂背景、阴影遮挡以及非均匀光照等干扰，因此，如何从单幅图像中可靠恢复文档的二维平面结构，仍然是一个具有挑战性的研究问题。

文档图像校正算法已经植入进实际应用产品，并不意味着复杂形变下的几何恢复问题已经得到充分解决。Adobe Scan 的官方说明已列出自动边界检测和书页曲率校正等功能(Adobe, n. d.)，因此不能将成熟系统简单概括为仅能进行全局透视变换。本文关注的研究问题是：在局部弯曲、褶皱与复杂背景同时出现时，如何在统一公开测试条件下进一步降低页面结构误差，并明确这一改进所需的参数量、计算量和推理开销。功能可用、视觉观感良好与局部几何准确属于不同层面的评价目标，已有产品的普及不能替代对上述问题的量化研究。

国内相关研究同样表明，文档图像预处理与后续识别、版面分析和信息抽取相互关联。Liu 等(2023)将变形文档分析与识别列为值得持续研究的问题；Lin 等(2023)讨论了文本位置偏差和识别错误对信息抽取流程的影响。因此，本研究以页面几何结构恢复为直接目标，同时保留 OCR 指标与计算开销分析，以考察几何收益能否兼顾识别适配性和运行成本，而不是仅以单一视觉效果作为改进依据。

早期 DIR 方法多依赖几何先验或三维重建，通过恢复文档表面形状并完成反投影实现校正。这类方法在受控环境中具有一定效果，但通常依赖额外硬件、多视角采集或较强的表面假设，难以

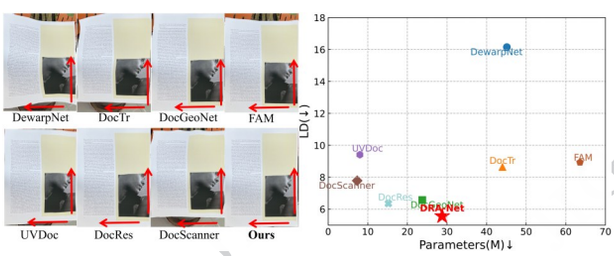


图 1　性能与参数量对比

Fig. 1　Comparison of performance and parameter count

适应移动拍摄和复杂自然场景。随着深度学习的发展，基于卷积神经网络(convolutional neural network, CNN)和 Transformer 的方法逐渐成为主流。CNN 能够直接学习从畸变图像到平整图像的映射关系，但受限于固定感受野和参数共享机制，在建模空间变化明显的非均匀形变时仍存在不足；Transformer 虽然具备更强的全局依赖建模能力，但其密集注意力会对所有空间位置进行两两交互，容易引入背景与无关区域噪声，并带来较高的计算开销。

以 DocTr(Feng 等，2021)、DocGeoNet(Feng 等，2022)、FAM(Li 等，2023)和 DocRes(Zhang 等，2024)为代表的现有方法，已在文档几何校正上取得持续进展。本文进一步考察局部响应聚合与选择性交互能否改善复杂形变下的几何预测。图 1 展示了

© 中国图象图形学报版权所有

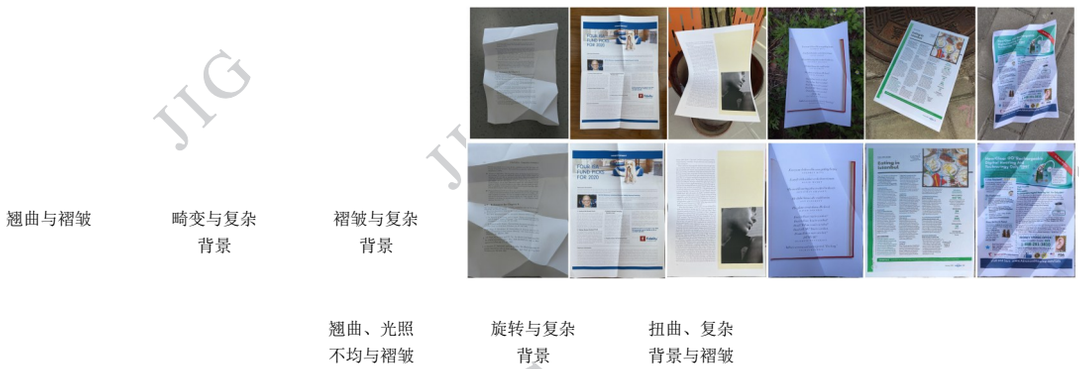


图 2　不同失真程度的文档图像示例

Fig. 2　Examples of document images with different distortion degrees

参与比较的方法在几何性能与参数规模上的差异；参数量不能直接代表实际推理效率，因此第3.3节进一步报告浮点运算量、推理耗时和峰值显存占用。

为此，本文提出一种面向文档图像校正的动态感受聚合网络（dynamic receptive aggregation network，DRA-Net）。该网络围绕“动态建模局部空间变化”和“抑制无效特征交互”两个目标展开设计，由动态感受聚合卷积（dynamic receptive aggregation convolution，DRAConv）、层次稀疏选择注意力（hierarchical sparse selective attention，HSSA）和显著性感知特征精炼模块（saliency-aware feature refinement module，SAFRM）组成。其中，DRAConv通过位置相关的动态聚合增强局部畸变建模能力；HSSA利用稀疏分支与稠密分支的协同机制，在保证全局上下文传播的同时抑制噪声交互；SAFRM进一步对通道响应进行显著性感知重整，以减少冗余特征并提升恢复质量。

综上，本文的主要贡献如下：

（1）提出一种用于文档图像校正的动态感受聚合网络DRA-Net，重点提升复杂几何形变场景下的局部结构恢复能力。

（2）设计了由DRAConv、HSSA和SAFRM构成的协同建模框架，分别从局部动态建模、选择性交互和通道精炼三个层面增强特征表达。

（3）在Doc3D上训练，并在DocUNet和DIR300完整公开测试集上评估。所提方法在本次比较的LD和AD几何指标上取得优势；其中在DIR300上，LD相较次优结果降低约12.4%。

1　相关工作

1.1　传统方法

传统文档图像校正方法主要建立在几何建模和三维重建基础上。早期方法通常利用文档边界、形状先验或光照约束恢复页面几何结构（Brown 等，2006；Zhang 等，2008）。随后，部分研究进一步引入多视角几何、度量校正和三维曲面重建，以提高弯曲文档的校正精度（Tsoi 等，2007；Koo 等，2009；Meng 等，2012；You 等，2018）。近年来，也有研究从边缘检测、秩分解和轻量化几何校正角度改进传统方法（Xu 等，2023；Li 等，2025）。这类方法通常先估计文档表面的三维形状，再映射回二维平面完成校正。为提高重建精度，部分工作引入参数化曲面、边界约束或多视角几何信息（Brown 等，2006；You 等，2018；He 等，2013；Liang 等，2008）。然而，这类方法往往依赖额外硬件设备（Meng 等，2012；Zhang 等，2008）、多视角采集（Brown 等，2006；You 等，2018）或较强的形状先验，在复杂起伏、局部折叠和非均匀形变场景下鲁棒性有限，实际应用成本也较高。

1.2　深度学习方法

随着深度学习的发展，基于数据驱动的DIR方法逐渐占据主流。DocUNet（Ma 等，2018）利用U-Net预测像素级形变场；DewarpNet（Das 等，2019）借助三维几何建模提升了复杂场景下的恢复能力；后续工作还通过平滑约束、文本方向建模和两阶段校正等策略进一步改善预测稳定性与可读性（Li 等，2019；Xie 等，2020；Markovitz 等，2020；Qin 等，2023；Feng

© 中国图象图形学报版权所有

等，2024）。

近年来，Transformer因其全局依赖建模能力被广泛应用于DIR任务。DocTr（Feng等，2021）实现了几何畸变与光照退化的联合处理，DocGeoNet（Feng等，2022）加强了几何表示学习，FAM（Li等，2023）通过前景与文本行约束改善局部恢复，UVDoc（Verhoeven等，2023）和DocRes（Zhang等，2024）进一步提升了复杂场景下的校正精度。尽管如此，现有方法大多仍依赖密集特征交互，对背景与无关区域缺乏足够抑制。基于这一观察，本文从动态局部建模和选择性交互两个角度对DIR网络进行改进。

与本文结构选择相关的通用视觉研究还包括感受野注意力卷积（receptive-field attention convolution，RFAConv）（Zhang X等，2024）和自适应稀疏Transformer（adaptive sparse transformer，AST）（Zhou等，2024）。前者讨论了在感受野响应上进行注意力调制，后者采用稀疏与稠密注意力分支及特征精炼。这些工作说明，响应重排、双分支融合和门控精炼均有已有研究基础。本文的研究重点是将相关机制配置于文档反向映射预测流程，并通过模块替换、内部组件和损失消融分析其在文档几何校正中的作用。

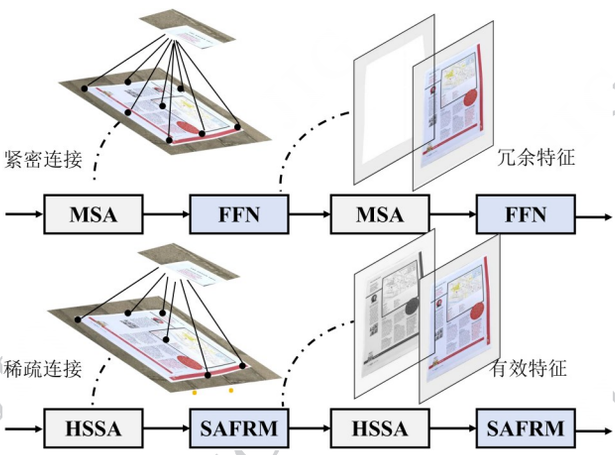


图3　DRA-Net与标准Transformer的特征交互方式对比

Fig. 3　Comparison of feature interaction mechanisms between DRA-Net and the standard Transformer

2　方　法

2.1　动机

如图2所示，真实文档图像中的几何畸变具有明显的空间非均匀性：不同区域在尺度、弯曲程度、拉伸方向和局部起伏方面往往存在显著差异。即便属于同一类失真，其具体表现形式和严重程度也可能相差较大。因此，DIR模型不仅需要具备较强的全局结构建模能力，还应能够自适应刻画局部区域中不断变化的畸变模式。

现有CNN方法受限于固定感受野和参数共享机制，在应对空间变化明显的复杂形变时往往缺乏足够的适应性；而标准Transformer虽然具备较强的长程依赖建模能力，但通常将所有标记统一送入多头自注意力模块进行计算，不仅会引入背景与无关区域的噪声交互，还会在前馈网络阶段继续放大冗余特征，进而影响恢复质量和计算效率。

本文围绕局部形变表征、空间交互选择和通道响应精炼组织网络。DRAConv将局部感受野响应显式重排后进行通道与空间调制，以增强输入投影阶段的局部几何表征；HSSA结合稀疏响应与稠密上下文传播，并在训练中学习二者的融合比例；SAFRM在空间交互后引入局部卷积和乘性门控，以精炼用于几何预测的特征。三者共同用于反向映射估计，最终通过双线性采样生成校正图像。该设计建立在已有响应加权与特征精炼思想之上，其研究价值需要由明确的任务配置和实验结果支撑，而非仅由模块命名或组合形式确定。图3展示了相应的特征交互方式。

2.2　网络结构

DRA-Net的整体结构如图4所示，采用四阶段层次化编码器-解码器框架。给定一幅失真文档图像，网络首先通过3×3卷积获得带重叠的patch嵌入；随后利用DRAConv生成初始特征表示$\boldsymbol{Y}$，以增强网络对空间变化畸变的感知能力。编码器和解码器各阶段均由HSSA与SAFRM构成的基本块堆叠而成，并在中间设置瓶颈阶段以进一步扩大感受范围、捕获更长距离的结构依赖。此外，在4个尺度的跳跃连接中引入经过适配的Chen等（2024）的隐式神经表示（implicit neural representation，INR），各分支对编码特征的3×3邻域进行展开，并通过多层感知机完成特征映射，保持输入特征的空间尺寸和通道数不变，再与对应的上采样解码特征沿通道维拼接。最终，网络通过3×3卷积预测反向映射，再结合双线性采样获得校正结果。下面分别介绍各关键模块。

1）动态感受聚合卷积：给定输入特征张量

© 中国图象图形学报版权所有

$\boldsymbol{X} \in R^{B\times C\times H\times W}$，式中 B、C、H、W 分别表示批量大小、输入通道数、高度和宽度，k 表示局部卷积核的边长。DRAConv 先由全局平均池化（global average pooling，GAP）和两层线性映射生成通道权重；再通过分组卷积（grouped convolution，GConv）生成局部响应并进行重排；最后将通道权重与空间权重作用于重排特征，并通过步长为 k 的卷积恢复空间分辨率。其计算过程为：

$$\begin{aligned}\boldsymbol{m}_c &= \sigma\left({}_2\left(\delta\left({}_1\left(GAP(\boldsymbol{X})\right)\right)\right)\right),\\ \boldsymbol{U} &= GConv_{k\times k}(\boldsymbol{X}),\\ \boldsymbol{U} &\quad R^{B\times Ck^2\times H\times W},\\ ,\boldsymbol{R} &= S\left(\tilde{\boldsymbol{U}}()\right)\boldsymbol{m}_s = \sigma\left(f_{3\times 3}\left(Concat\left[Max_c(\boldsymbol{R}),Mean_c(\boldsymbol{R})\right]\right)\right),\boldsymbol{Y} = Conv_{k\times k}^{s=k}\left(\boldsymbol{R}\quad \boldsymbol{m}_c\quad \boldsymbol{m}_s\right).\end{aligned} \tag{1}$$

式中，$_1$ 和 $_2$ 为线性映射，δ 为修正线性单元（rectified linear unit，ReLU）激活函数，σ 为 Sigmoid 函数，$f_{3\times 3}$ 为生成空间权重的卷积映射。View 表示维度重塑，S 表示响应重排，*Concat* 表示通道拼接，Max_c 和 $Mean_c$ 分别沿通道维取最大值和平均值；表示带广播的逐元素相乘。$\boldsymbol{m}_c$ $R^{B\times C\times 1\times 1}$ 在空间维度广播，$\boldsymbol{m}_s$ $R^{B\times 1\times kH\times kW}$ 在通道维度广播，输出为 $\boldsymbol{Y}$ $R^{B\times C_o\times H\times W}$，式中 C_o 为输出通道数，s 为卷积步长。图 4 中的 LRLS 与 CS 分别是上述通道权重生成和空间权重生成过程的简写。

为明确重排规则，使用从 0 开始的索引。令 b_0、c、h、w 分别为批量、通道和原空间位置索引，a,b $\{0,\cdots,k\quad 1\}$ 为局部响应位置，$t = ak + b$ 为长度为 k^2 的响应轴索引，则

$$\boldsymbol{R}_{b_0,c,kh+a,kw+b} = . \tag{2}$$

这里的重排不引入可学习采样偏移，也不直接改变文档像素坐标；输入相关的调制来自 $\boldsymbol{m}_c$ 和 $\boldsymbol{m}_s$。

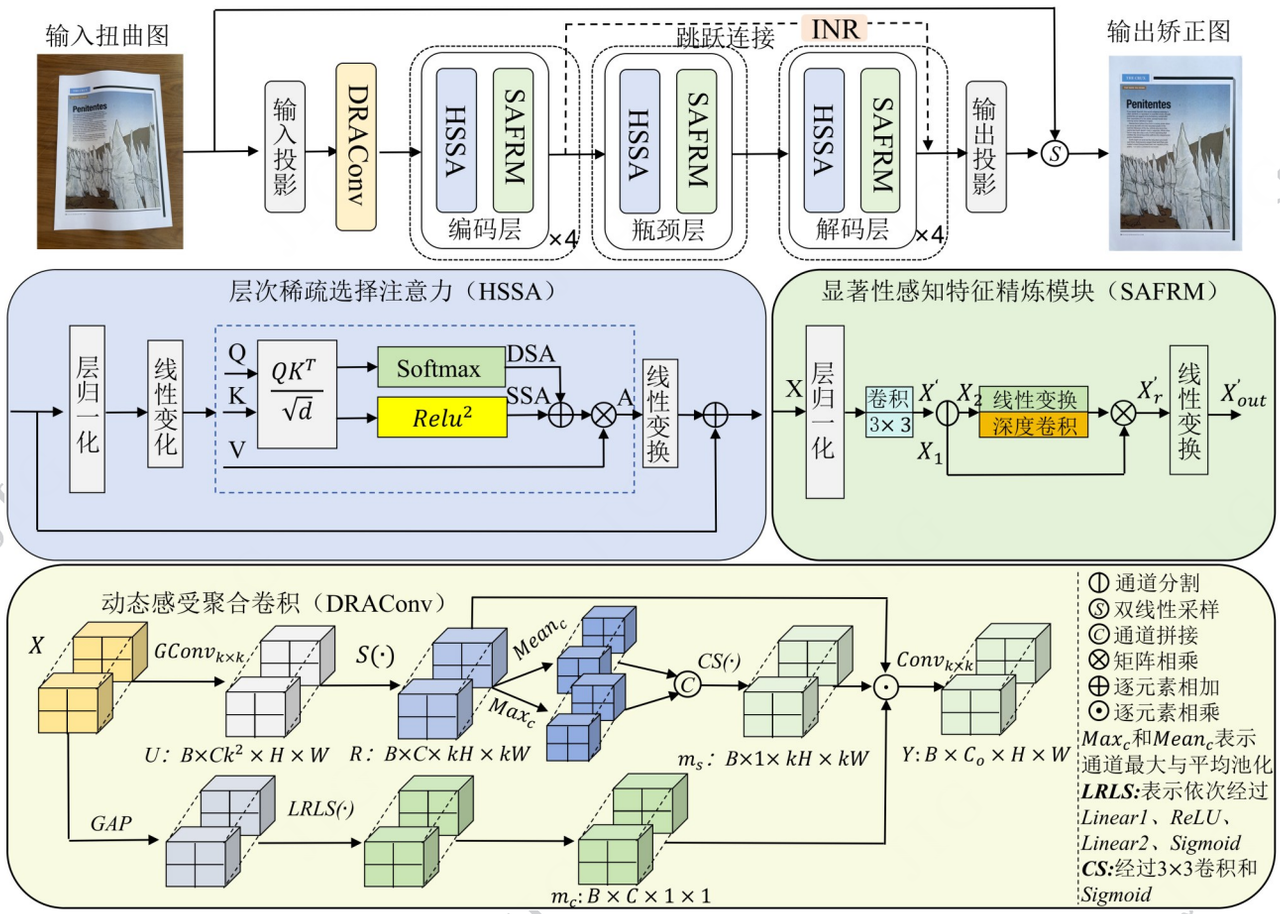


图 4　所提出的 DRA-Net 整体架构

Fig. 4　Overall architecture of the proposed DRA-Net

© 中国图象图形学报版权所有

因此，DRAConv 中的“动态”指响应权重随输入变化，而不是卷积采样网格本身发生形变。

2）层次稀疏选择注意力：HSSA 由密集自注意力（dense self-attention，DSA）与稀疏自注意力（sparse self-attention，SSA）两个分支组成。为便于说明，以下省略批量和注意力头索引。设一个注意力计算域中包含 N 个标记，每个标记的查询、键和值特征维度为 d，则 $\mathrm{Q},\mathrm{K},\mathrm{V}\in\mathrm{R}^{\mathrm{N}\times\mathrm{d}}$ 分别为查询、键和值矩阵。相对位置偏置矩阵为 $\mathrm{B}_{\mathrm{pos}}\in R^{\mathrm{N}\times\mathrm{N}}$。密集分支的注意力矩阵为

$$A_{\mathrm{DSA}}=Softmax\left(\frac{\mathrm{QK}^{T}}{\sqrt{d}}+\mathrm{B}_{pos}\right),\tag{3}$$

式中，*Softmax* 沿键对应的维度归一化，T 表示转置。稀疏分支为

$$A_{SSA}=ReLU^{2}\left(\frac{\mathrm{QK}^{T}}{\sqrt{d}}+\mathrm{B}_{pos}\right),\tag{4}$$

式中，$ReLU^{2}(z)=\left[max(0,z)\right]^{2}$，该操作逐元素将非正分数置零，并不执行 Top-k 选取。两分支的输出融合为：

$$A=\left(w_{1}A_{SSA}+w_{2}A_{DSA}\right)V,\tag{5}$$

式中，$A\in R^{\mathrm{N}\times d}$ 为融合后的特征，w_1 和 w_2 为分支权重，由两个可学习标量 a_1、a_2 归一化得到：

$$w_{1}=\frac{\mathbf{e}^{a_1}}{\mathbf{e}^{a_1}+\mathbf{e}^{a_2}},\ w_{2}=\frac{\mathbf{e}^{a_2}}{\mathbf{e}^{a_1}+\mathbf{e}^{a_2}}\tag{6}$$

a_1 和 a_2 均初始化为 1。按照式（6）的定义，融合比例在训练中学习，推理时固定；随输入改变的是注意力响应，而不是由额外输入门控网络即时生成的 w_1,w_2。HSSA 在层次化编码器与解码器的不同特征阶段使用，两分支分别提供选择性响应和稠密上下文。

需要区分响应稀疏与计算稀疏。只要密集分支保留正权重，融合后的注意力一般仍为稠密矩阵。使用普通稠密矩阵运算时，一个注意力计算域内的主要计算复杂度仍为 $O\left(N^{2}d\right)$，注意力矩阵存储量仍为 $O\left(N^{2}\right)$。因此，该模块的作用是调整交互响应，不构成渐近计算复杂度降低的依据。

3）显著性感知特征精炼模块：SAFRM 在空间交互之后处理通道特征。相较于只通过逐位置线性映射处理特征的前馈网络（feed-forward network，FFN），该模块通过逐点卷积引入局部信息，并以乘性门控对通道响应进行调制。其计算过程按本文网络定义写为：

$$\begin{aligned}\boldsymbol{X}\quad&=PConv\left(LN\left(\boldsymbol{X}\right)\right),\\\left(\boldsymbol{X}_1,\boldsymbol{X}_2\right)&=Split\left(\boldsymbol{X}\quad\right),\\\boldsymbol{X}\quad_r&=\boldsymbol{X}_1\quad DWConv\left(Linear\left(\boldsymbol{X}_2\right)\right),\\\boldsymbol{X}\quad_{out}&=Linear\left(\boldsymbol{X}\quad_r\right).\end{aligned}\tag{7}$$

式中，$\boldsymbol{X}$ 为输入特征张量，层归一化（layer normalization，LN）用于归一化特征；*PConv* 表示分支划分之前的逐点卷积，$\boldsymbol{X}$ 为其中间输出，*Split* 沿通道维将其划分为 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$。第二分支经过线性映射 *Linear* 和逐通道卷积（depthwise convolution，DWConv）生成局部调制特征，其输出维度与 $\boldsymbol{X}_1$ 相同，以进行逐元素相乘。$\boldsymbol{X}\quad_r$ 为门控后的中间特征，$\boldsymbol{X}\quad_{out}$ 为输出特征。这里的“显著性”指训练得到的响应选择作用，未额外引入显著图标注，也不预设每个通道与某类文档结构具有一一对应关系。

2.3 损失函数

为同时约束反向映射和校正图像，本文联合使用映射误差、像素重建误差与结构相似性约束。

令 $\boldsymbol{I}_{bm}$、$\boldsymbol{I}_{bm}^{gt}$ 分别表示预测和参考反向映射，$\boldsymbol{I}_{out}$、$\boldsymbol{I}_{d}^{gt}$ 分别表示校正输出和参考平整图像，上标 *gt* 表示真实标注。L_1 和 L_2 分别表示按训练实现计算的 L_1 和 L_2 误差；*SSIM* 表示结构相似性指数（structural similarity index，SSIM）（Wang 等，2004）。为避免映射项重复乘权，L_{bm} 定义为未加权映射损失，各项为：

$$L_{bm}=L_{1}\left(\boldsymbol{I}_{bm},\boldsymbol{I}_{bm}^{gt}\right),\tag{8}$$

$$L_{ssim}=1\quad SSIM\left(\boldsymbol{I}_{out},\boldsymbol{I}_{d}^{gt}\right),\tag{9}$$

$$L_{d}=\lambda_{2}L_{2}\left(\boldsymbol{I}_{out},\boldsymbol{I}_{d}^{gt}\right)+\lambda_{3}L_{ssim},\tag{10}$$

$$L_{total}=\lambda_{1}L_{bm}+L_{d}.\tag{11}$$

L_d 为图像重建损失，L_{total} 为总损失，$\lambda_1,\lambda_2,\lambda_3$ 分别为映射、像素重建和结构相似性项的权重，默认设为 8、0.1 和 0.06。这里的 SSIM 用于训练约束，与实验评价使用的多尺度结构相似性指标相区分。

2.4 结构联系与任务适配分析

1）局部几何响应聚合。感受野响应重排及后续注意力调制与 RFAConv 中的相关结构具有联系（Zhang X 等，2024）。本文将这一类机制置于文档输入特征形成阶段，并通过反向映射误差学习与校正任务相关的响应。由式（1）可见，最终聚合系数由卷积参数、通道权重和局部空间权重共同决定；即使卷积参数共享，不同输入位置也可以得到不同的有

© 中国图象图形学报版权所有

效响应系数。该解释说明了其相较固定响应聚合的功能，但不将响应重排或通道注意力本身作为独立原创算子。表4中普通卷积替换及表7中的组件移除实验用于检验这一配置对几何恢复的作用。

2）空间交互与通道精炼。AST已研究稀疏与稠密注意力结合及特征精炼（Zhou等，2024），因此仅采用这两类分支或门控乘法不能单独证明结构新颖性。本文将相应机制用于层次化文档几何预测：HSSA处理标记间交互，SAFRM则利用局部信息调制空间交互后的通道响应。式（6）所定义的训练可学习比例、式（7）中的局部调制及反向映射监督共同确定本文具体配置。其必要性分别由注意力替换、固定融合、局部卷积移除和门控移除实验检验，不能仅凭实验有效性推断所有组成操作均为首次提出。

3）实验支撑与结论边界。现有消融表明，在相同训练与测试设置下，完整配置的几何指标优于所测试的替换或删减配置。例如，普通卷积替代DRAConv后LD为6.4224，HSSA固定融合后LD为5.8656，同时去除SAFRM局部卷积与门控后LD为9.0241，均高于完整模型的5.5687。这些结果支持各功能环节在本文任务中的实用贡献；当前实验并未提供与所有同类注意力或精炼算子的逐一等预算比较，因此不据此宣称通用算子层面的全面优势。

3 实　验

3.1 数据集和实现细节

本文采用Doc3D数据集（Das等，2019）进行训练。该数据集包含10万幅合成失真文档图像，并提供三维坐标、法线、深度、二维纹理坐标(u,v)映射和反向映射等标注，可为DIR模型提供充分监督。测试阶段选用DocUNet（Ma等，2018）和DIR300（Feng等，2022）两个公开基准测试集。为保证实验结果的可比性，分别使用完整公开测试集进行评估，其中DocUNet包含130幅真实拍摄文档图像，DIR300包含300幅复杂场景文档图像。

模型在单张NVIDIA GeForce RTX 4090图形处理器（graphics processing unit，GPU）上训练93个epoch，全局批量大小为4。优化器采用自适应矩估计（adaptive moment estimation，Adam），权重衰减为5×10^{-4}，并采用余弦退火策略，最大学习率为2×10^{-4}。训练图像块大小为256×256。评价采用多尺度结构相似性指数（multi-scale structural similarity index，MS-SSIM）（Wang等，2003）、局部失真（local distortion，LD）（You等，2018）、对齐失真（aligned distortion，AD）（Ma等，2022）、字符错误率（character error rate，CER）（Morris等，2004）和编辑距离（edit distance，ED）（Levenshtein，1966）。前3项用于评估图像结构及几何校正质量，后2项衡量OCR识别结果。所有对比方法均在相同硬件与训练设置下重新训练并测试，以保证比较公平。

跳跃连接中的INR分支采用3个隐藏层，每层包含256个单元，隐藏层使用ReLU激活，各尺度分支不共享参数。DRA-Net完整模型及其各消融配置均保留相同的分支结构设置，相关参数随各实验模型共同训练。复杂度与推理效率测试采用包含该分支的实际模型。

3.2 与最先进方法的比较

本文将所提方法与7种代表性DIR算法进行比较，包括DewarpNet（Das等，2019）、DocTr（Feng等，2021）、DocGeoNet（Feng等，2022）、FAM（Li等，2023）、UVDoc（Verhoeven等，2023）、DocRes（Zhang等，2024）和DocScanner（Feng等，2025）。为保证公平性，所有对比方法均基于公开实现，在相同训练与测试平台上完成复现和评估。

表1　DocUNet数据集上的定量比较

Table 1　Quantitative experiments on the DocUNet dataset

方法	MS-SSIM↑	LD↓	AD↓	ED↓	CER↓
DewarpNet	0.4549	11.8551	0.4758	1029.96	0.4614
DocTr	0.5098	12.1657	0.4543	974.88	0.4237
DocGeoNet	0.5083	11.5791	0.4495	952.68	0.4235
FAM	0.5009	12.6745	0.4534	960.90	0.4259
UVDoc	**0.5281**	**11.0128**	**0.3917**	945.06	0.4265
DocRes	0.4921	11.2881	0.4704	983.46	0.4451
DocScanner	0.5092	11.6316	0.4264	**857.60**	**0.3844**
DRA-Net	**0.5102**	**9.6794**	**0.3913**	**927.52**	**0.4157**

注：下划线加粗表示最优结果，仅加粗表示次优结果。

表1和表2给出了在DocUNet与DIR300数据集上的定量比较结果。整体来看，DRA-Net在LD和AD两项几何校正指标上均优于对比方法，说明

© 中国图象图形学报版权所有

表 2　DIR300数据集上的定量比较

Table 2　Quantitative experiments on the DIR300 dataset

方法	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
DewarpNet	0.4588	16.1594	0.3633	957.49	0.7111
DocTr	0.5868	8.6439	0.2950	625.00	**0.3569**
DocGeoNet	**0.6300**	6.5746	0.2476	698.70	0.6791
FAM	0.5786	8.9340	0.2488	654.083	0.3570
UVDoc	0.5953	9.4162	0.2424	688.43	0.5606
DocRes	<u>0.6315</u>	6.3537	**0.2370**	675.50	0.5458
DocScanner	0.6050	7.7811	0.2371	<u>526.47</u>	<u>0.2691</u>
DRA-Net	0.6235	<u>5.5687</u>	<u>0.1911</u>	**575.64**	0.5119

注：下划线加粗表示最优结果，仅加粗表示次优结果。

其在页面布局恢复、局部形变校正和结构对齐方面具有较强优势。CER 与 ED 衡量本文 OCR 评估流程中的识别结果，不能与几何指标简单互换。表 1 和表 2 中，DocScanner 的 CER 和 ED 更低，而 DRA-Net 的 LD 和 AD 更低。相较于 DocScanner，DRA-Net 在 DocUNet 上的 LD、AD 分别降低约 16.8% 和 8.23%，在 DIR300 上分别降低约 28.43% 和 19.40%。这些结果表明几何误差降低并不必然带来本次 OCR 指标的同步最优。

图 5 和图 6 给出了 DocUNet 与 DIR300 上的定性比较结果。与对比方法相比，DRA-Net 在复杂弯曲、边界缺失和局部拉伸场景下能够恢复更自然的页面结构，文本行更平直，边缘区域也更完整；

3.3　模型复杂度与推理效率分析

本文比较参数量（parameters，Params）、浮点运算量（floating-point operations，FLOPs）、单图推理耗时（latency）和峰值显存占（memory）。

所有方法在同一 NVIDIA GeForce RTX 4090 GPU 上测试，输入分辨率为 256 × 256，批量大小为 1。浮点运算量采用同一工具统计；推理前进行预热以减少初始化对计时的影响，随后通过统一计算设备架构（compute unified device architecture，CUDA）

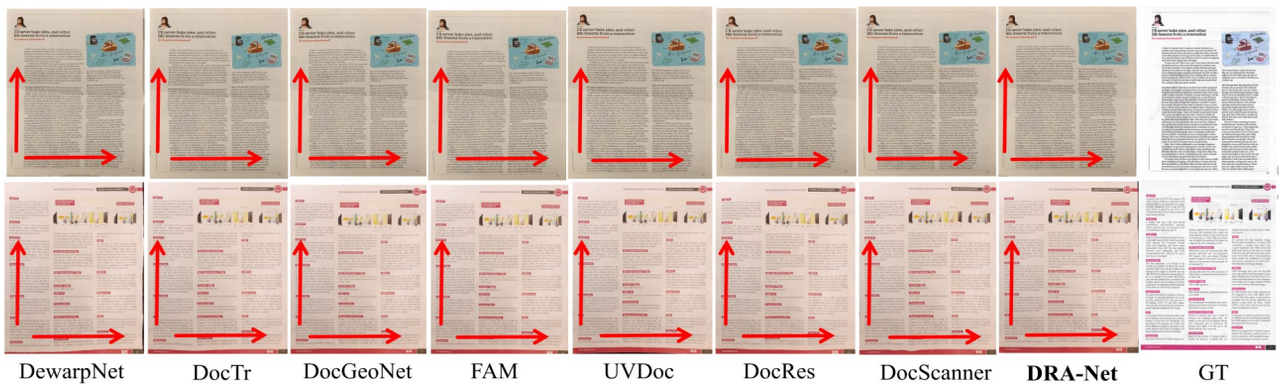


图 5　DocUNet数据集上的定性比较

Fig. 5　Qualitative experiments on the DocUNet dataset

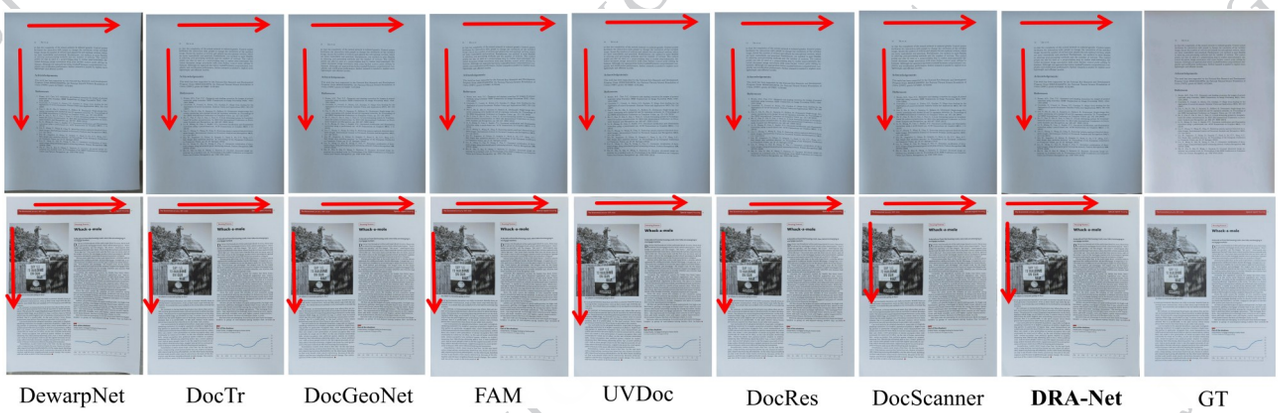


图 6　DIR300数据集上的定性比较

Fig. 6　Qualitative experiments on the DIR300 dataset

© 中国图象图形学报版权所有

事件计时器记录多次前向推理的平均耗时。显存指标为相应推理过程记录的峰值。本节数值只代表这一硬件与输入设置下的测试结果，不直接等同于移动端耗时或完整产品处理延迟。

由表3可见，DRA-Net的参数量为28.785M，低于DewarpNet、DocTr和FAM，但高于DocGeoNet、DocRes、UVDoc和DocScanner。其FLOPs为141.680G，高于本次对比的其他方法。结合前文结果，DRA-Net主要面向复杂形变下的几何恢复精度，而不以最低理论计算量为目标。

值得注意的是，尽管DRA-Net的FLOPs较高，但其单图推理耗时为31.434 ms/张，接近DocScanner（Feng等，2025）的30.031 ms/张，并明显低于FAM（Li等，2023）和DocRes（Zhang等，2024）；同时，DRA-Net的峰值显存占用为885.06 MB，仅高于UVDoc（Verhoeven等，2023）和DewarpNet（Das等，2019），显著低于DocTr（Feng等，2021）、DocGeoNet（Feng等，2022）、FAM（Li等，2023）、DocRes（Zhang等，2024）和DocScanner（Feng等，2025）。这说明DRA-Net虽然理论计算量较大，但实际显存占用较低，推理速度仍处于可接受范围。结合前文几何校正结果，DRA-Net主要体现为几何校正性能与实际推理开销之间的折中。

与DocScanner相比，DRA-Net的单图推理耗时增加1.403 ms，约增加4.67%，峰值显存占用降低约73.27%；同时参数量和FLOPs更高。因此，其实际权衡是更低的几何误差和所测显存峰值，而不是全面降低计算代价。对成熟扫描系统，本研究尚未进行限定软件版本、采集流程和设备配置的同条件测试，故不作产品间的精度或速度排序。本文结果可为扫描流程中的几何校正环节提供研究依据，不能替代对完整系统的评测。

表3　计算开销对比结果

Table 3　Comparison results of computational overhead

方法	Params ↓	FLOPs ↓	Latency ↓	Memory ↓
DewarpNet	45.120 M	4.495 G	13.442 ms	724.53 MB
DocTr	44.010 M	37.692 G	21.821 ms	8701.29 MB
DocGeoNet	23.763 M	32.012 G	23.739 ms	4366.31 MB
FAM	63.584 M	131.238 G	114.821 ms	4639.08 MB
UVDoc	8.003 M	3.414 G	7.469 ms	251.68 MB
DocRes	15.204 M	91.807 G	58.338 ms	14272.74 MB
DocScanner	7.329 M	61.586 G	30.031 ms	3310.72 MB
DRA-Net	28.785 M	141.680 G	31.434 ms	885.06 MB

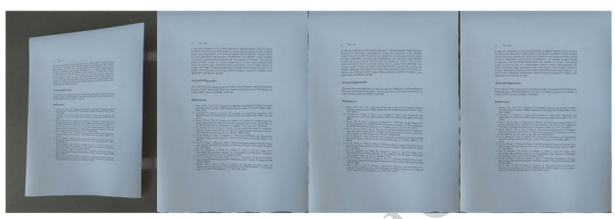
3.4　消融实验

3.4.1　模块间消融

为验证各关键模块的有效性，本文在DIR300数据集上进行了消融实验。

DRAConv的作用：本文在输入投影阶段引入DRAConv，以在局部窗口内动态生成位置相关权重，从而增强对空间变化畸变的建模能力。为验证其有效性，本文设置三组对比：1）移除DRAConv后的模型，记为Rm-DRAConv，模型的LD和AD分别退化至7.6347和0.3121；2）使用普通卷积替代DRAConv后的模型记为Pc-DRAConv，模型的LD和AD分别退化至6.4224和0.2681，说明固定卷积核难以充分适应文档图像中空间非均匀的局部形变；3）所提出的完整模型。表4表明，引入DRAConv后，模型在所有评价指标上均有明显提升；图7进一步说明，完整模型在复杂失真区域能够恢复更自然的局部结构和更平滑的页面形态。

HSSA的作用：为分析所提注意力机制的贡献，本文将HSSA分别替换为标准多头自注意力（multi-



(c)Pc-DRAConv　**(d)DRA-Net**

(b)Rm-DRAConv　(a)输入扭曲图

图7　在DRAConv模块上的定性消融实验

Fig. 7　Qualitative ablation experiments on the DRAConv module

© 中国图象图形学报版权所有

表 4　DRAConv模块替换与移除实验的定量结果

Table 4　Quantitative ablation results for different DRAConv module configurations

方法	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
Rm-DRAConv	0.5654	7.6347	0.3121	1028.43	0.7438
Pc-DRAConv	0.6136	6.4224	0.2681	795.37	0.6450
DRA-Net	**0.6235**	**5.5687**	**0.1911**	**575.64**	**0.5119**

表 5　HSSA模块替换实验的定量结果

Table 5　Quantitative ablation results for different HSSA module configurations

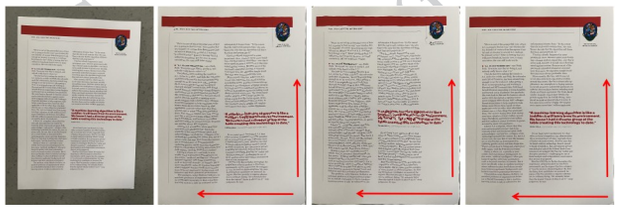
方法	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
DRA-Net/MSA	0.5823	6.4575	0.2582	781.30	0.6049
DRA-Net/W-SA	0.5634	6.7947	0.2778	1021.9	0.6382
DRA-Net	**0.6235**	**5.5687**	**0.1911**	**575.64**	**0.5119**

head self-attention，MSA)(Vaswani 等，2017)和基于窗口的自注意力(window-based self-attention，WSA)，对应模型记为 DRA-Net/MSA 和 DRA-Net/W-SA。其中 W-SA 为该窗口注意力替换配置的标识。由表5和图8可见，两种替代方案的性能均低于完整模型。这表明，HSSA 通过联合稀疏建模与密集建模，能够更有效地筛选有用交互、抑制背景噪声，并提升复杂失真场景下的结构恢复能力。

SAFRM 的作用：为验证通道精炼策略的有效性，本文分别以多层感知机(multilayer perceptron，MLP)和局部增强前馈网络(locally-enhanced feed-forward network，LEFF)(Yuan 等，2021)替换 SAFRM，对应模型记为 DRA-Net/MLP 和 DRA-Net/LEFF。表6和图9表明，采用SAFRM的完整模型在各项指标上均优于DRA-Net/LEFF替代方案，仅在CER指标上逊色于DRA-Net/MLP，说明该模块能够有效强化显著通道响应并抑制冗余特征，从而进一步提升网络的整体恢复能力。

3.4.2　模块内消融

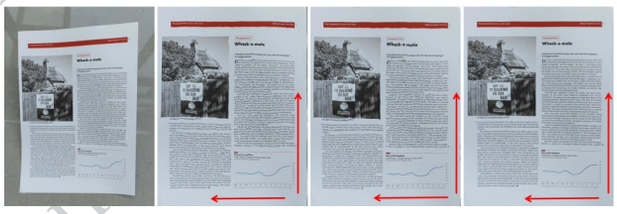
为进一步分析各模块内部关键设计的贡献，本文在 DIR300 数据集上补充细粒度消融实验。不同于前文移除或替换整个模块的实验，本节分别考察 DRAConv 中通道重加权、空间重加权和感受野重排，HSSA 中稀疏分支、稠密分支和动态融合权重，以及 SAFRM 中局部卷积和门控机制的作用。所有实验均采用相同训练设置，仅改变对应模块内部



(a)输入扭曲图　(b)DRA-Net/MSA　(c)DRA-Net/W-SA　**(d)DRA-Net**

图 8　在HSSA模块上的定性消融实验

Fig. 8　Qualitative ablation experiments on the HSSA module



(a)输入扭曲图　(c)DRA-Net/LEFF　**(d)DRA-Net**　(b)DRA-Net/MLP

图 9　在SAFRM模块上的定性消融实验

Fig. 9　Qualitative ablation experiments on the SAFRM module

© 中国图象图形学报版权所有

表 6　SAFRM模块替换实验的定量结果

Table 6　Quantitative ablation results for different SAFRM module configurations

方法	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
DRA-Net/MLP	0.6005	6.6242	0.2935	1085.73	**0.4663**
DRA-Net/LEFF	0.6010	6.3320	0.2665	915.60	0.6656
DRA-Net	**0.6235**	**5.5687**	**0.1911**	**575.64**	0.5119

结构。

表 7 给出了 DRAConv 内部结构的消融结果。去除通道重加权(w/o channel-reweight)后,MS-SSIM 降至 0.5734,LD 和 AD 分别退化至 6.8598 和 0.2821,表明通道重加权有助于突出与文档结构相关的有效响应。去除感受野重排(w/o refine-shape)后,LD 进一步退化至 7.3022,ED 增大至 1068.93,说明将 k^2 个局部响应重新组织到扩展空间维度对于局部几何聚合具有重要作用。去除空间重加权(w/o spatial-reweight)后,AD 和 CER 分别退化至 0.2900 和 0.7791,说明空间权重调制能够增强模型对不同弯曲区域和褶皱区域的自适应建模能力。完整 DRAConv 在所有指标上均取得最佳结果,验证了通道重加权、感受野重排和空间权重调制的互补性。

表 7　DRAConv内部组件消融的定量结果

Table 7　Quantitative ablation results for internal components within the DRAConv module

方法	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
w/o channel-reweight	0.5734	6.8598	0.2821	989.741	0.6544
w/o refine-shape	0.5681	7.3022	0.2834	1068.93	0.6390
w/o spatial-reweight	0.6016	6.7385	0.2900	984.87	0.7791
DRA-Net	**0.6235**	**5.5687**	**0.1911**	**575.64**	**0.5119**

表 8 给出了单分支、固定等权融合和可学习融合的比较。仅使用 SSA 时,MS-SSIM 为 0.5833,LD 和 AD 为 6.9412 和 0.2739,均弱于完整模型;仅使用 DSA 时,LD 为 7.1815。固定等权融合的 LD 和 AD 优于两种单分支配置,但其 CER 为 0.6405,高于仅用 SSA 时的 0.5868,因此不能认定固定融合在所有指标上均更优。完整模型在该表的 5 项指标上均取得最佳结果,支持训练可学习融合比例的作用。这一结果不表示推理时存在依赖输入的标量权重生成器,也不直接证明稀疏矩阵计算得到加速。

表 8　HSSA内部组件消融的定量结果

Table 8　Quantitative ablation results for internal components within the HSSA module

方法	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
DSA only	0.5894	7.1815	0.2572	849.35	0.6810
SSA only	0.5833	6.9412	0.2739	852.0	0.5868
fixed fusion	0.6087	5.8656	0.2395	813.27	0.6405
DRA-Net	**0.6235**	**5.5687**	**0.1911**	**575.64**	**0.5119**

表 9 分析了 SAFRM 中局部卷积和门控机制的作用。可以看到,去除局部卷积(w/o Local-Conv)后,模型的 MS-SSIM 从 0.6235 降至 0.5904,LD 和 AD 分别退化至 6.4803 和 0.2513,说明局部卷积能够为通道精炼分支提供必要的邻域结构信息。去除门控机制(w/o Gate)后,模型同样出现明显性能下降,表明简单的线性变换难以有效抑制冗余通道响应。进一步地,当同时去除局部卷积和门控机制(w/o Local-Conv&Gate)时,模型性能下降最为明显,LD 和 AD 分别退化至 9.0241 和 0.4232,ED 也增大至 1616.58。这说明 SAFRM 中的局部结构建模和门控调制具有协同作用,能够帮助网络突出与文本行、页面边界和局部弯曲区域相关的几何特征。

表 9　SAFRM内部组件消融的定量结果

Table 9　Quantitative ablation results for internal components within the SAFRM module

方法	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
w/o Local-Conv	0.5904	6.4803	0.2513	865.7	0.6170
w/o Gate	0.5717	6.5125	0.2749	835.73	0.7488
w/o Local-Conv&Gate	0.5330	9.0241	0.4232	1616.58	0.7703
DRA-Net	**0.6235**	**5.5687**	**0.1911**	**575.64**	**0.5119**

3.4.3　各损失项及其权重系数的消融实验

为验证各损失项对文档图像校正性能的影响,本文对损失函数进行消融实验。具体而言,L_{bm} 用于

© 中国图象图形学报版权所有

直接约束预测反向映射与真实反向映射之间的差异，L_2 用于约束校正图像与真实平整图像之间的像素一致性，$L_{ssim}=1-SSIM(\cdot)$ 用于增强结构一致性。所有实验均采用相同网络结构和训练设置，仅改变损失项组成。

由表10可见，仅使用映射损失 L_{bm} 时，LD 和 AD 分别为8.2716和0.2787，说明单独采用反向映射监督尚不足以获得完整联合损失的校正效果。去除像素重建损失 L_2 后，LD 和 AD 分别为6.4889和0.2694，ED 与 CER 也高于完整模型，表明图像像素一致性约束对当前配置有补充作用。去除 L_{ssim} 后，MS-SSIM 降至0.5684，LD 和 ED 分别为7.1803和971.07，支持结构一致性约束的作用。完整联合损失在表中5项指标上均优于所测试的删减配置，说明这3类监督在本文训练设置下具有互补作用。

表10　损失项消融实验

Table 10　Ablation experiments on loss terms

方法	MS-SSIM ↑	LD ↓	AD ↓	ED ↓	CER ↓
only L_{bm}	0.5566	8.2716	0.2787	978.77	0.7833
w/o L_2	0.6020	6.4889	0.2694	920.075	0.6861
w/o L_{ssim}	0.5684	7.1803	0.2669	971.07	0.7129
DRA-Net	**0.6235**	**5.5687**	**0.1911**	**575.64**	**0.5119**

表11　权重系数敏感性分析

Table 11　Sensitivity analysis of weight coefficients

$(\lambda_1,\lambda_2,\lambda_3)$	*MS-SSIM*	*LD*	*AD*	*ED*	*CER*
(7, 0.1, 0.06)	0.5958	7.0100	0.2783	976.41	0.4124
(10, 0.1, 0.06)	0.6132	6.3184	0.2499	798.60	0.5742
(8, 0.06, 0.06)	0.6009	6.8077	0.2795	860.05	0.6134
(8, 0.3, 0.06)	0.6186	6.2365	0.2939	955.43	**0.4117**
(8, 0.1, 0.01)	0.6107	6.3191	0.2824	877.18	0.6687
(8, 0.1, 0.10)	0.6187	6.5053	0.2733	821.32	0.6236
(8, 0.1, 0.06)	**0.6235**	**5.5687**	**0.1911**	**575.64**	0.5119

表11中的 $(\lambda_1,\lambda_2,\lambda_3)$ 分别表示未加权映射损失 L_{bm}、像素重建损失 L_2 和结构损失 L_{ssim} 的系数，例如 $(7,0.1,0.06)$ 为 $7L_{bm}+0.1L_2+0.06L_{ssim}$。每组仅调整一个系数，其余训练配置保持不变。

由表11可见，默认权重 $(8,0.1,0.06)$ 在所测试设置中取得最高 MS-SSIM 和最低 LD、AD、ED。分别调整任一系数后，上述4项指标均不同程度退化。不过，$(7,0.1,0.06)$ 和 $(8,0.3,0.06)$ 的 CER 分别为0.4124和0.4117，均低于默认设置的0.5119。因此，默认权重反映本文对几何恢复与图像结构的侧重，并不是全部指标的共同最优解；本实验仅说明测试范围内的敏感性，不据此宣称全局最优或统计显著性。

3.5　结果讨论

结合两组公开测试和消融结果，本文的主要收益体现在几何预测，而不是独立的纹理生成或图像增强。DRAConv 的局部响应调制、HSSA 的分支融合和 SAFRM 的局部门控配置，在本次训练协议下共同改善了 LD 与 AD。模块间替换和内部删减实验给出了功能性证据，但不能单独区分结构变化、参数分配与优化过程各自带来的全部影响，也不能替代与最接近已有结构的等预算比较。

几何评价与识别评价存在不一致性。DocScanner 在本次测试的 ED 和 CER 上更优；在 SAFRM 替换实验中，MLP 配置的 CER 也低于完整模型。因此，不能仅由几何收益推断对下游识别的普遍提升。未来需要结合识别反馈，并在明确识别引擎、图像分辨率和文本归一化规则的条件下分析误差来源。

4　结　论

本文面向复杂形变文档的反向映射预测，构建了结合动态感受聚合、稀疏与稠密响应融合以及局部门控精炼的校正网络。其方法特点是围绕局部几何表征、空间信息传播和通道响应选择进行任务配置，并以映射、像素和结构约束联合训练。在 DocUNet 与 DIR300 完整公开测试集上，模型取得了本次对比中最优的 LD 和 AD；模块及损失消融进一步支持各环节对当前几何预测任务的有效性。

本文的结论同时受到评价范围约束：已有操作的任务配置和消融有效性不等同于每个算子均具有独立原创性；较低的几何误差也未带来全部 OCR 指标最优。DRA-Net 的理论 FLOPs 较高，所测推理延迟与显存优势只适用于本文硬件和输入条件，尚不能推广为对商业系统或移动端部署的全面领先。后续研究将进一步研究文本识别适配及更高效的几何特征聚合，以改进校正精度与实际部署成本之间的

© 中国图象图形学报版权所有

平衡。

参考文献(References)

Adobe. n.
d. Take scans: Adobe Scan for iOS[EB/OL]. [2026-10-08]. https://www.adobe.com/devnet-docs/adobescan/ios/en/scan.html

Brown M S and Tsoi Y C. 2006. Geometric and shading correction for images of printed materials using boundary. IEEE Transactions on Image Processing, 15(6): 1544 - 1554 [DOI: 10.1109/TIP.2006.871082]

Chen X, Pan J and Dong J. 2024. Bidirectional multi-scale implicit neural representations for image deraining//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos, CA, USA: IEEE: 25627 - 25636 [DOI: 10.1109/CVPR52733.2024.02421]

Das S, Ma K, Shu Z, Samaras D and Shilkrot R. 2019. DewarpNet: single-image document unwarping with stacked 3D and 2D regression networks//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, South Korea: IEEE: 131 - 140 [DOI: 10.1109/ICCV.2019.00022]

Feng H, Wang Y, Zhou W, Deng J and Li H. 2021. DocTr: document image transformer for geometric unwarping and illumination correction//Proceedings of the 29th ACM International Conference on Multimedia. Chengdu, China: Association for Computing Machinery: 273 - 281 [DOI: 10.1145/3474085.3475388]

Feng H, Zhou W, Deng J, Wang Y and Li H. 2022. Geometric representation learning for document image rectification//Computer Vision - ECCV 2022. Cham: Springer: 475 - 492 [DOI: 10.1007/978-3-031-19836-6_27]

Feng H, Zhou W, Deng J, Tian Q and Li H. 2025. DocScanner: robust document image rectification with progressive learning. International Journal of Computer Vision, 133(8): 5343 - 5362 [DOI: 10.1007/s11263-025-02431-5]

Feng J, Chi Y, Zhou Y T and He J F. 2024. Two image rectification networks for distorted and warped documents. Journal of Data Acquisition and Processing, 39(1): 167 - 180 (冯瑾, 池越, 周亚同, 何静飞. 2024. 面向畸变扭曲文档的两种图像矫正网络. 数据采集与处理, 39(1): 167 - 180) [DOI: 10.16337/j.1004-9037.2024.01.015]

He Y, Pan P, Xie S, Sun J and Naoi S. 2013. A book dewarping system by boundary-based 3D surface reconstruction//Proceedings of the 12th International Conference on Document Analysis and Recognition. Washington, DC, USA: IEEE: 403 - 407 [DOI: 10.1109/ICDAR.2013.88]

Koo H I, Kim J and Cho N I. 2009. Composition of a dewarped and enhanced document image from two view images. IEEE Transactions on Image Processing, 18(7): 1551 - 1562 [DOI: 10.1109/TIP.2009.2019301]

Levenshtein V I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady, 10(8): 707 - 710

Li H, Wu X, Chen Q and Xiang Q. 2023. Foreground and text-lines aware document image rectification//Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 19574 - 19583 [DOI: 10.1109/ICCV51070.2023.01793]

Li Q H, Feng M and Yang C H. 2025. Document image skew correction based on rank decomposition. Journal of Nanchang Hangkong University (Natural Sciences), 39(3): 57 - 64 (李千红, 冯镅, 杨词慧. 2025. 基于秩分解的文档图像倾斜校正. 南昌航空大学学报(自然科学版), 39(3): 57 - 64)

Li X, Zhang B, Liao J and Sander P V. 2019. Document rectification and illumination correction using a patch-based CNN. ACM Transactions on Graphics, 38(6): 168:1 - 168:11 [DOI: 10.1145/3355089.3356563]

Liang J, DeMenthon D and Doermann D. 2008. Geometric rectification of camera-captured document images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(4): 591 - 605 [DOI: 10.1109/TPAMI.2007.70724]

Lin Z N, Wang J P and Jin L W. 2023. Visual information extraction deep learning method: a critical review. Journal of Image and Graphics, 28(8): 2276 - 2297 (林泽柠, 汪嘉鹏, 金连文. 2023. 视觉信息抽取的深度学习方法综述. 中国图象图形学报, 28(8): 2276 - 2297) [DOI: 10.11834/jig.220904]

Liu C L, Jin L W, Bai X, Li X H and Yin F. 2023. Frontiers of intelligent document analysis and recognition: review and prospects. Journal of Image and Graphics, 28(8): 2223 - 2252 (刘成林, 金连文, 白翔, 李晓辉, 殷飞. 2023. 文档智能分析与识别前沿:回顾与展望. 中国图象图形学报, 28(8): 2223 - 2252) [DOI: 10.11834/jig.221112]

Ma K, Shu Z, Bai X, Wang J and Samaras D. 2018. DocUNet: document image unwarping via a stacked U-Net//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE: 4700 - 4709 [DOI: 10.1109/CVPR.2018.00494]

Ma K, Das S, Shu Z and Samaras D. 2022. Learning from documents in the wild to improve document unwarping//Proceedings of the ACM SIGGRAPH 2022 Conference. Vancouver, Canada: Association for Computing Machinery: 34: 1 - 34: 9 [DOI: 10.1145/3528233.3530756]

Markovitz A, Lavi I, Perel O, Mazor S and Litman R. 2020. Can you read me now? Content aware rectification using angle supervision//Computer Vision - ECCV 2020. Cham: Springer: 208 - 223 [DOI: 10.1007/978-3-030-58610-2_13]

Meng G, Pan C, Xiang S, Duan J and Zheng N. 2012. Metric rectification of curved document images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4): 707 - 722 [DOI: 10.

© 中国图象图形学报版权所有

1109/TPAMI.2011.151]

Morris A C, Maier V and Green P D. 2004. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition//Proceedings of Interspeech 2004. Jeju Island, South Korea: International Speech Communication Association: 2765 - 2768 [DOI: 10.21437/Interspeech.2004-668]

Qin H, Li Y J, Liang Q K and Wang Y N. 2023. Asymmetric geometric correction network for document images. Journal of Image and Graphics, 28(8): 2314 - 2329 (秦海, 李艺杰, 梁桥康, 王耀南. 2023. 针对文档图像的非对称式几何校正网络. 中国图象图形学报, 28(8): 2314 - 2329) [DOI: 10.11834/jig.220426]

Tsoi Y C and Brown M S. 2007. Multi-view document rectification using boundary//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Minneapolis, MN, USA: IEEE: 1 - 8 [DOI: 10.1109/CVPR.2007.383251]

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. 2017. Attention is all you need//Advances in Neural Information Processing Systems 30. Red Hook, NY, USA: Curran Associates: 5998 - 6008

Verhoeven F, Magne T and Sorkine-Hornung O. 2023. UVDoc: neural grid-based document unwarping//SIGGRAPH Asia 2023 Conference Papers. Sydney, Australia: Association for Computing Machinery: 1 - 11 [DOI: 10.1145/3610548.3618174]

Wang Z, Simoncelli E P and Bovik A C. 2003. Multiscale structural similarity for image quality assessment//Proceedings of the 37th Asilomar Conference on Signals, Systems and Computers. Pacific Grove, CA, USA: IEEE, 2: 1398 - 1402 [DOI: 10.1109/ACSSC.2003.1292216]

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4): 600 - 612 [DOI: 10.1109/TIP.2003.819861]

Xie G W, Yin F, Zhang X Y and Liu C L. 2020. Dewarping document image by displacement flow estimation with fully convolutional network//Document Analysis Systems. Cham: Springer: 131 - 144 [DOI: 10.1007/978-3-030-57058-3_10]

Xu Y D, Xiong Y P, Zhang Z, Wu G B, Zhang X and Wang W. 2023. Warped document rectification based on image edge detection. Journal of Nanjing University (Natural Sciences), 59(4): 660 - 668 (徐远东, 熊永平, 张铮, 伍贵宾, 张兴, 王伟. 2023. 基于图像边缘检测的扭曲文档矫正. 南京大学学报(自然科学), 59(4): 660 - 668) [DOI: 10.13232/j.cnki.jnju.2023.04.012]

You S, Matsushita Y, Sinha S, Bou Y and Ikeuchi K. 2018. Multiview rectification of folded documents. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(2): 505 - 511 [DOI: 10.1109/TPAMI.2017.2675980]

Yuan K, Guo S, Liu Z, Zhou A, Yu F and Wu W. 2021. Incorporating convolution designs into visual transformers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 579 - 588 [DOI: 10.1109/ICCV48922.2021.00062]

Zhang J, Peng D, Liu C, Zhang P and Jin L. 2024. DocRes: a generalist model toward unifying document image restoration tasks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 15654 - 15664 [DOI: 10.1109/CVPR52733.2024.01482]

Zhang L, Zhang Y and Tan C. 2008. An improved physically-based method for geometric restoration of distorted document images. IEEE Transactions on Pattern Analysis and Machine Intelligence, 30(4): 728 - 734 [DOI: 10.1109/TPAMI.2007.70831]

Zhang X, Liu C, Yang D G, Song T T, Ye Y C, Li K, et al. 2024. RFAConv: innovating spatial attention and standard convolutional operation[EB/OL]. [2026-10-08]. https://arxiv. org/abs/2304.03198v6 [DOI: 10.48550/arXiv. 2304.03198]

Zhou S H, Chen D S, Pan J S, Shi J L and Yang J F. 2024. Adapt or perish: adaptive sparse transformer with attentive feature refinement for image restoration//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 2952 - 2963 [DOI: 10.1109/CVPR52733.2024.00285]

作者简介

曾健烘，男，硕士研究生，主要研究方向为文档校正。E-mail: 18030302604@163.com

周欣玥，女，硕士研究生，主要研究方向为文档矫正。E-mail: zxyyy010918@163.com

江楠峰，通信作者，男，副教授，主要研究方向为图像增强。E-mail: jnfrock@gmail.com

王大寒，男，教授，主要研究方向为模式识别、人工智能。E-mail: Wangdh AT xmut.edu.cn

© 中国图象图形学报版权所有