doi:10.6041/j.issn.1000-1298.2026.17.036

基于多尺度增强与几何自适应融合的食品包装文本检测模型

田　萱[1]　肖文昊[2]

（1.北京林业大学信息学院，北京 100083；2.林业智能信息处理工程技术研究中心，北京 100083）

摘要：食品包装文本包含产品名称、成分、厂商信息和保质期等关键信息，对食品安全与消费者权益具有重要意义。将现有的场景文本检测方法应用于食品包装文本检测任务中，普遍存在模糊文本、小目标以及艺术字体的漏检和误检等问题，针对这些问题，提出了一种基于多尺度增强与几何自适应融合的食品包装文本检测模型 MESA－Text。具体而言，在 MESA－Text 中首先提出多尺度注意力特征增强融合模块，通过注意力奇偶层交替融合策略，集成门控双分支注意力机制和特征平滑方法，增强融合不同尺度文本特征能力；然后针对文本高宽比失调问题，提出几何稳定模块，融合文本几何特征进行自适应改进，以精准捕捉扭曲、模糊及微小文本特征；最后辅以动态可变形卷积增强不同尺度特征提取，以提升食品包装文本检测效果。消融实验结果表明，引入多尺度注意力特征增强融合模块、几何稳定模块及动态可变形多尺度特征提取模块后，召回率分别提升 2.18、0.90、1.14 个百分点。采用本文模型方法在食品包装图像上的召回率、精确率和 F1 值分别达 92.12%、91.76% 和 91.94%，有效缓解漏检误检问题。

关键词：食品包装文本检测；多尺度增强；注意力机制；几何特征；动态可变形卷积

中图分类号：TP391.4　　**文献标识码**：A　　**文章编号**：1000-1298(2026)17-0392-12　　**OSID**：

Text Detection Model for Food Packaging Based on Multi-scale Enhancement and Geometric Adaptive Fusion

Tian Xuan[1]　Xiao Wenhao[2]

(1. *School of Information Science and Technology*, *Beijing Forestry University*, *Beijing* 100083, *China*
2. *Engineering Research Center for Forestry-oriented Intelligent Information Processing of National Forestry and Grassland Administration*, *Beijing* 100083, *China*)

Abstract: Text on food packaging contains key information such as product names, ingredients, manufacturer details, and expiration dates, which are crucial for food safety and consumer rights. However, applying existing scene text detection methods to food packaging often results in missed or incorrect detections, especially for blurred text, small targets, and artistic fonts. To address these challenges, a food packaging text detection model was proposed based on multi-scale enhancement and geometric fusion, named MESA－Text. MESA－Text introduced a multi-scale attention feature enhancement and fusion module, which leveraged an alternating fusion strategy of attention-based even-odd layers, integrating a gated dual-branch attention mechanism and a feature smoothing method to effectively enhance the ability to fuse features from different text scales. To address the problem of imbalanced text aspect ratios, a geometric feature stabilization module was proposed to adaptively refine features by incorporating geometric information, allowing more accurate and robust detection of distorted, blurred, and small-scale text. Additionally, a dynamic deformable convolution module was used to enhance feature extraction across multiple scales, further improving the representation capability of complex text regions and enhancing overall detection performance. The proposed framework was specifically designed to address the diverse appearance characteristics of food packaging text, enabling more reliable localization under complex backgrounds and varying imaging conditions. Ablation experiments demonstrated that incorporating the multi-scale attention enhancement module, geometric stabilization module, and dynamic multi-scale feature extraction module improved recall by 2.18, 0.90,

收稿日期：2025－05－07　修回日期：2025－07－22

基金项目：数字赋能国家植物园建设及种质资源收集保护关键技术研究与示范项目(Z221100005222018)

作者简介：田萱(1976—)，女，副教授，博士，主要从事智能信息处理与文本挖掘研究，E-mail：tianxuan@bjfu.edu.cn

and 1.14 percentage points, respectively. Using the proposed model, MESA－Text achieved 92.12% recall, 91.76% precision, and F1-score of 91.94% on food packaging images, demonstrating superior effectiveness and robustness in challenging food packaging scenarios, while effectively reducing missed and false detections.

Key words: food packaging text detection; multi-scale enhancement; attention mechanism; geometric features; dynamic deformable convolution

0　引言

食品包装文本检测是计算机视觉在自然场景文本检测技术中的重要应用，广泛应用于食品溯源、质量监管与自动结算等领域。然而，食品包装上的文本通常具有复杂多样的视觉特征，如艺术字、不同大小字体、复杂背景以及图文混合，这对文本检测算法提出了更高的挑战。目前深度学习方法在处理复杂视觉特征、提升检测精度以及适应多样化文本形态方面展现出显著优势，已成为当前该领域研究的主流方向[1]。

深度学习文本检测主要包括基于边框回归的方法和基于语义分割[2-3]的方法。边框回归方法，例如 CTPN[4]（Connectionist text proposal network）以及 EAST[5]（Efficient and accurate scene text detector），通过回归来预测文本区域，但在处理弯曲文本和不规则文本时存在局限性，面对有艺术字的食品包装文本检测中难以保证检测效果。而语义分割作为一种精细的分割技术旨在准确地区分目标前景和背景，其中许多研究致力于提升特征提取与融合模块的性能。Liu 等[6]提出了一种基于递归特征金字塔模块的自然场景文本检测方法，将特征金字塔网络中的补充反馈连接引入至自下而上的骨干网络中，但是易在语义提取过程中丢失细节特征。随后，Liao 等[7]在 DBNet（Differentiable binarization network）基础上引入了自适应尺度融合模块，该模块能够根据输入图像的尺度与位置差异，自主调整特征融合的权重，但却忽略了文本边界的细粒度分析与建模。为解决此问题，Zhang 等[8]提出了 TextBPN + +（Text boundary perception network + +）模型，专注于任意形状文本的检测，采用特征金字塔（Feature pyramid network，FPN）进行多层次特征融合后分割出距离场和分类图，引入了迭代边界优化模块，通过从粗到精的方式对文本边界显式建模。然而，TextBPN + +在特征融合阶段仍存在对多尺度信息利用不足的问题，限制其在小尺度或复杂背景文本下检测的性能表现。当前多数文本检测方法如 SRFormer[9]（Segmentation and regression former）、LRANet[10]（Low-rank approximation network）和 PSENet[11]（Progressive scale expansion network）等大都通过语义特征分析方式注重于特征提取、边界拟合、尺度融合等方面的提升，对文本本身所具备的几何结构特征关注不足。事实上，在实际食品包装图像中，仅依靠语义分割难以准确捕捉其形状边界。

与通用自然场景文本检测任务相比，食品包装文本检测仍然存在以下问题：受限于感受野大小，易发生艺术字等变形文本检测的不充分；特征利用不充分，模糊文本、小尺度文本容易被漏检；包装文本呈现出高宽失调的几何比例，在检测不同形状文本时容易发生误检。

针对上述问题，本文提出一种改进 TextBPN + +的基于多尺度增强和几何自适应融合食品包装文本检测方法（Multi-scale enhancement and shape-adaptive fusion－text detection，MESA－Text）。

在沿用 TextBPN + +边界提议模块和迭代优化模块基础上，MESA－Text 提出了动态可变形多尺度特征提取模块（Dynamic deformable multi-scale feature extraction，DDMFE）、多尺度注意力特征增强融合模块（Multi-scale attention feature enhancement and fusion，MAFEF）和几何特征稳定模块（Geometric feature stability，GFS），进一步提升食品包装文本检测的精确性和召回率。具体而言，本文针对受限于感受野大小，艺术字等变形文本检测不充分问题，提出动态可变形多尺度特征提取模块，提高不同尺度文本特征提取能力，以减少误检和漏检问题；针对特征利用不充分、小尺度容易被漏检问题，提出多尺度注意力特征增强融合模块，增强不同尺度特征的融合，以提高模糊文本的检测精度、提高小文本的检测召回率；针对包装文本高宽比失调的几何特征检测效果不高问题，提出几何特征稳定模块，增强对不同高宽比包装文本的自适应检测能力。

1　基于多尺度注意力特征融合的包装文本检测

1.1　网络架构

MESA－Text 网络结构如图 1 所示，主要由三大改进模块构成：动态可变形多尺度特征提取模块、多尺度注意力特征增强融合模块、几何特征稳定模块。

MESA－Text 首先将图像输入动态可变形多尺

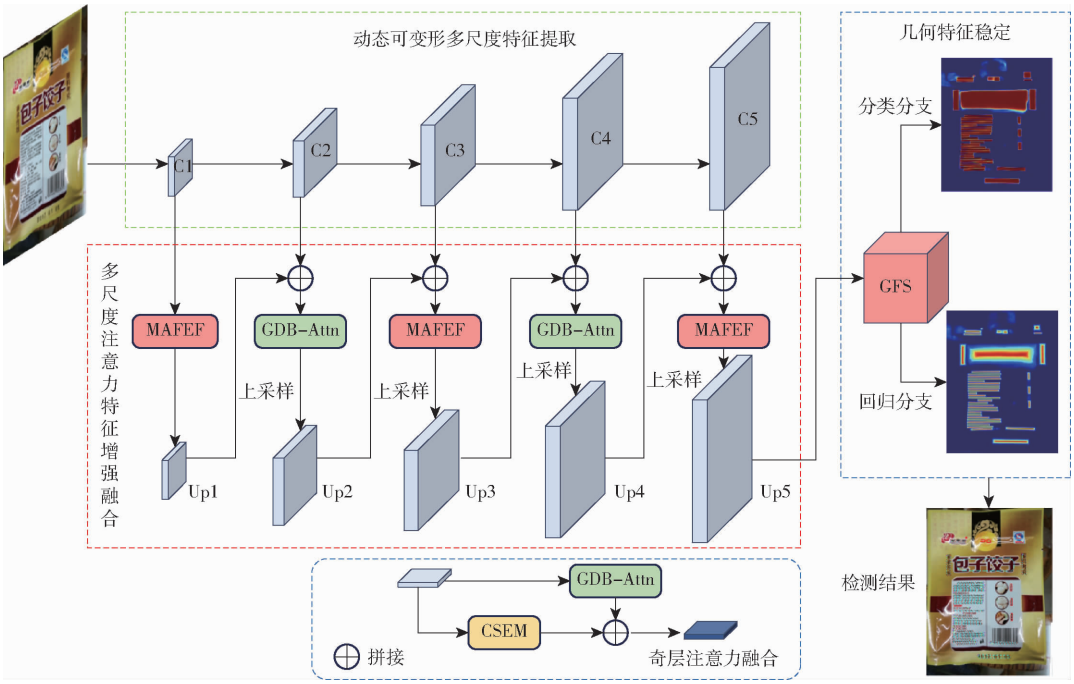

图 1 MESA - Text 模型算法架构

Fig. 1 MESA - Text model architecture

度特征提取模块骨干网络提取特征,通过动态可变形卷积操作调节感受野,生成多尺度特征图 {C1 ~ C5},随后,对提取的特征进行多尺度融合,多尺度特征融合增强模块采用注意力奇偶层交替融合策略,对模块中门控双分支注意力机制(Gated dual-branch attention enhancement, GDB - Attn)与特征平滑方法(Cross-smooth enhancement module, CSEM)处理,生成尺度匹配特征 {A1 ~ A5},再将各尺度特征融合,得到最终特征 {Up1 ~ Up5}。接着几何特征稳定模块进一步对融合后的特征进行处理,结合几何特征进行边界分类与回归,得到可以表示距离场与分类图的特征。最后,将几何特征稳定模块的输出特征送入基线模型的边界提议模块和迭代边界优化模块进行文本边界检测。

1.2 动态可变形多尺度特征提取模块

TextBPN + + 在特征提取时,使用可变形卷积[12]直接提取不同尺度层语义特征,导致信息损失,使得受限于感受野大小的边界弯曲文本、艺术字等变形文本检测不充分。

针对这一问题,本文在 MESA - Text 模型中设计了动态可变形多尺度特征提取模块,在 TextBPN + + 所用的可变形卷积的基础上,设计动态分组可变形卷积。相比于 TextBPN + + 使用固定分组数的传统方法,动态分组可变形卷积可依据输入特征通道大小来调整分组数,调节感受野大小,以适应食品包装上不同尺度的文本形态。若动态分组数为 K,则这 K 组输出的增强特征表示可通过可变形卷积计算,即

$$Y = \sum_{k=1}^{K} \mathrm{DefConv}(X; \Delta P^k) \tag{1}$$

其中,DefConv 表示第 k 组的可变形卷积算子,X 为每层输入特征,ΔP^k 代表第 k 个分组中描述感受野扩大程度的偏移量,其计算完全参照文献[12]中的过程,即先通过对输入特征计算 3 × 3 卷积获得其压缩特征,再通过双线性插值取整求出视野动态扩展的偏移量,模块算法表示为:

算法 1

Input: Feature map $X \in \mathbf{R}^{B\times C\times H\times W}$, K

Output: Enhanced feature map $Y \in \mathbf{R}^{B\times C\times H\times W}$

Define group set $K \in \{1, 4, 8\}$

    // Initialize output

    $Y \leftarrow 0$

    // Split X into K channel groups:

    $\{X[1], X[2], \cdots, X[K]\} \leftarrow X$

    for $k = 1$ to K do

        $\Delta P^k \leftarrow$ Offset($X[k]$) // 参照文献[12]

        // Apply Deformable convolution:

        $Y[k] \leftarrow$ DefConv($X[k]$, ΔP^k)

    end for

    // Merge all sub-group:

    $Y[k] \leftarrow$ Concat($\{Y[1], \cdots, Y[K]\}$)

    // Accumulate result with Resnet:

    $Y \leftarrow X + Y[k]$

end for

16：Return Y

如算法 1 所示，动态可变形多尺度特征提取模块预先定义了 3 种可变形卷积分组策略，即 $K \in \{1, 4, 8\}$，对应不同的感受野配置。当输入 X 特征通道数较小的时候（如大面积艺术字），模块采用较小的组数，如 $K=1$，能使所有通道共享同一偏移量 ΔP^1，统一调整卷积核位置，实现大感受野范围内的一致感知，增强对整体艺术字形变的捕捉能力。相反，当通道数较多时（如复杂边缘细节的艺术字），DDMFE 会采用更大的分组数（如 $K=4$、$K=8$），将通道划分为多个子组，每组学习独立的偏移量 ΔP^k，使卷积核能贴合细节的形变，如艺术字的笔画弯曲、边缘角点。

通过动态分组可变形卷积，DDMFE 有效缓解了传统固定分组方式在特征提取阶段因感受野受限而导致的艺术字检测不充分的问题，根据输入特征自适应调整感受野大小，在大范围语义和局部细节之间取得平衡。

1.3　多尺度注意力特征增强融合模块

在 TextBPN + + 特征金字塔中，不同层级特征采用简单的求和拼接操作来进行特征融合，这样的特征融合方式丢失了部分不同尺度特征的空间和全局信息。

针对这一问题，提出多尺度注意力特征增强融合模块，下含门控双分支注意力增强模块、特征平滑模块、特征融合模块（Feature fusion module，FFM）3 个细分模块，模块整体算法表示为：

算法 2

Input：Multi-scale feature maps $\{X[1], X[2], X[3], X[4], X[5]\}$ from DDMFE

Output：Enhanced feature maps $\{\mathrm{Up}[1], \mathrm{Up}[2], \mathrm{Up}[3], \mathrm{Up}[4], \mathrm{Up}[5]\}$

for $i=1$ to 5 do
　$z \leftarrow$ Global AvgPool($X[i]$)
　$[W_{channel}, W_{spatial}] \leftarrow Gate(z)$，
　　Where $W_{channel} + W_{spatial} = 1$
　ca $\leftarrow$ CA($X[i]$)
　sa $\leftarrow$ SA($X[i]$)
　AttnOut $\leftarrow X[i] \odot (W_{channel} \odot \mathrm{ca} + W_{spatial} \odot \mathrm{sa})$
　if $i \in \{1, 3, 5\}$ then
　　CSEMOut $\leftarrow$ CSEM($X[i]$)
　　$A[i] \leftarrow$ AttnOut + CSEMOut
　else
　　$A[i] \leftarrow$ AttnOut
　end if
　if $i == 1$ then
　　$\mathrm{Up}[i] \leftarrow \mathrm{A}[i]$
　else
　　$\mathrm{Up}[i] \leftarrow \mathrm{FFM}(A[i], A[i-1])$
　end if
end for
return $\{\mathrm{Up}[1], \mathrm{Up}[2], \mathrm{Up}[3], \mathrm{Up}[4], \mathrm{Up}[5]\}$

如算法 2 和图 1 所示，为了使不同感受野提取的特征能更精准地融合，提升 MESA - Text 在复杂背景下对模糊和小目标文本的检测性能，在多尺度注意力特征增强融合模块中设计了一种注意力奇偶层交替融合策略：在 1、3、5 特征层使用 GDB - Attn + CSEM 模块以增强多尺度语义表达能力，在 2、4 层则不采用 CSEM 平滑模块，以减少对细节信息的过度平滑，从而保证特征的一致性，最后使用特征融合模块融合增强后不同尺度的特征。

1.3.1　门控双分支注意力模块（GDB - Attn）

为了增强不同尺度特征层的语义表达能力，设计了门控双分支注意力增强模块，采用自适应门控机制来融合通道注意力（Channel atteniton，CA）[13] 和空间注意力（Spatial attention，SA），自适应门控机制通过计算权重平衡因子，实现对两种注意力的动态加权。

双分支注意力如图 2 所示，对于骨干网络每层提取的特征图 $C_i \in \mathbf{R}^{C \times H \times W}$，GDB - Attn 首先在通道维度加入注意力，生成两个通道描述向量，形状均为 $\mathbf{R}^{C \times 1 \times 1}$，输入共享权重的逐点卷积网络并加和后，经 Sigmoid 激活得到通道注意力 Ca，在空间维度，同样将两个特征图拼接后通过一个 Conv2d（7 × 7）卷积层提取空间注意力 Sa。

如图 2 所示，相比传统注意力机制的卷积块注意模块（Convolutional block attention module，CBAM）[14]，GDB - Attn 的自适应门控机制能够根据输入特征的语义需求在融合通道与空间注意力时，计算权重平衡因子来赋予模型根据图像内容自动调节注意力比重的能力。

权重平衡因子求解公式为

$$z = \mathrm{GAP}(X_i) \quad (z \in \mathbf{R}^{C \times 1 \times 1}) \tag{2}$$

$$[W_{channel}, W_{spatial}] = \sigma(\mathrm{Conv}_{1 \times 1}(z)) \quad (W_{channel} + W_{spatial} = 1) \tag{3}$$

其中，GAP 为自适应全局平均池化（Global average pooling，GAP），X_i 为输入第 i 层特征，z 为 $\mathbf{R}^{C \times 1 \times 1}$ 的语义压缩表示张量，保留特征通道级上下文信息，σ 为 ReLU 激活函数，使用 1 × 1 卷积操作，将通道数从 C 映射为 2，以生成动态权重平衡因子 $W_{channel}$ 和 $W_{spatial}$，二者和为 1。再将权重平衡因子对双分支注

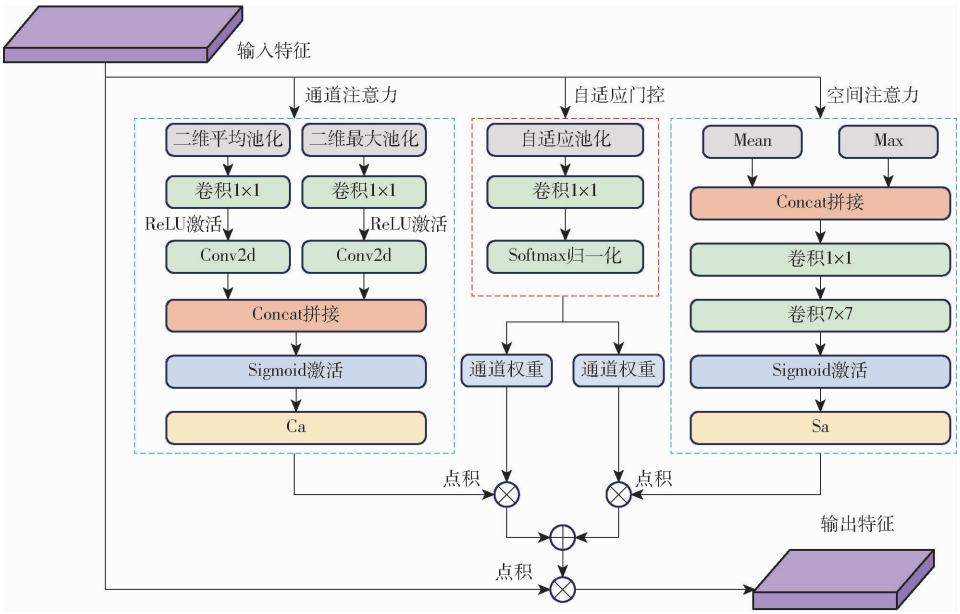

图 2 门控双分支注意力增强模块

Fig. 2 GDB – Attn module

意力进行调节,得到增强后的特征,计算过程为

$$Out = X \odot (W_{channel} \odot Ca + W_{spatial} \odot Sa) \quad (4)$$

式中 Ca——通道注意力

Sa——空间注意力

将权重平衡因子 $W_{channel}$、$W_{spatial}$ 与通道和空间注意力分别进行加权,并将加权后的结果与输入特征 X 融合,生成最终的增强特征表示。

通过添加权重平衡因子作为门控,突破了通道注意力与空间注意力以固定比例融合的局限,使得注意力的融合方式由静态转变为基于食品包装文本数据的动态融合,更契合食品包装文本检测任务中多变的视觉特性。权重平衡因子可根据食品包装图像中不同大小的文字和背景干扰程度,灵活地分配注意力比重。例如,对于小尺度、模糊或边缘模糊的文本区域,模型可以增强空间注意力以更好定位文本位置;而对于字体复杂、语义差异显著的区域,则更侧重通道注意力以强化语义区分,从而增强了特征语义表达能力。

1.3.2 特征平滑模块(CSEM)

GDB – Attn 能有效提升通道和空间维度的特征表达能力,但在每个尺度层堆叠时可能会导致特征的失调,掩盖一些关键的局部细节。尤其是在食品包装图像中,存在大量低分辨率甚至十分细微的文字区域,若不更加关注这些区域,GDB – Attn 注意力机制可能强化了主干区域的响应而忽略了微小文本的存在。

针对此问题,在多尺度注意力特征增强融合模块中进一步引入了特征平滑模块以提升特征的细节保留能力。如图 3 所示,其本质是一个高效的注意力机制模块[15],主要在组级别进行特征之间的内部语义学习,利用自适应池化和组归一化等方法,将输入 $X \in \mathbf{R}^{C \times H \times W}$ 分为 G 个子特征组,每个组新的输入特征 $X \in \mathbf{R}^{\frac{C}{G} \times H \times W}$,并在每个子组内独立学习语义特征。

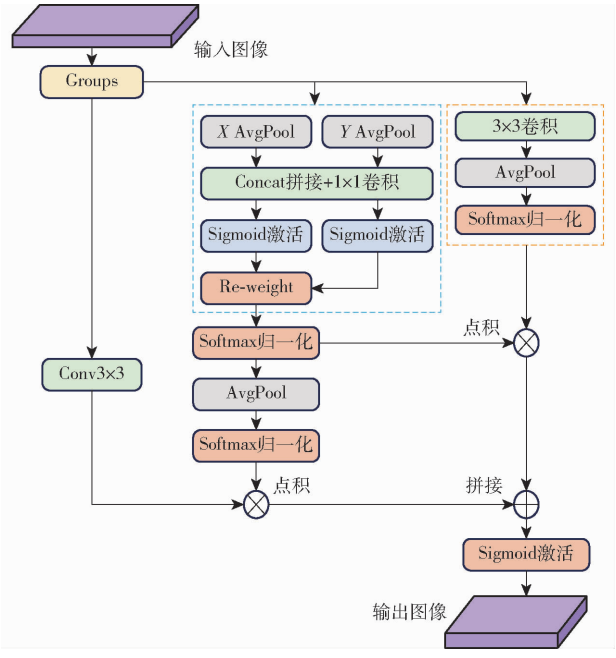

图 3 特征平滑模块

Fig. 3 CSEM module

在引入特征平滑模块的基础上,为提升 CSEM 与 GDB – Attn 在多尺度融合过程的协同效果,设计了注意力奇偶层交替融合的策略,其数学表达为

$$Out = \begin{cases} \text{GDB} - \text{Attn}(X_i) + \text{CSEM}(X_i) & (i = 1,3,5) \\ \text{GDB} - \text{Attn}(X_i) & (i = 2,4) \end{cases} \quad (5)$$

该策略的性能优势，主要体现在以下两个方面：

（1）采用并行融合，避免特征干扰：不同注意力机制采用串行结构时，易因注意力关注域重叠而导致过拟合。针对门控双分支注意力增强与特征平滑分别侧重通道-空间增强与细节平滑的优势特点，多尺度注意力特征增强融合模块采用并行融合策略，分别对输入特征X_i采用 GDB－Attn 和 CSEM 操作后再进行融合，有效发挥了两者的互补优势。

（2）采用奇偶层交替融合结构，防止过度平滑：CSEM 模块虽在细节建模中表现优异，但若在所有层级中频繁应用，可能导致过度平滑，抑制 GDB－Attn 突出的显著语义区域。注意力奇偶层交替融合策略在第 1、3、5 层采用 GDB－Attn＋CSEM，强化关键层级的多尺度语义融合；在第 2、4 层仅保留 GDB－Attn，保持特征的敏感性和多样性，降低注意力堆叠的复杂度。

1.3.3　特征融合模块（FFM）

在 TextBPN＋＋中，特征金字塔通过简单的逐层求和与拼接方式来融合不同尺度的特征，造成了特征语义混淆，尤其对模糊和小文本的区域时易发生漏检。

针对这一问题，本文设计了特征融合模块，旨在充分利用 GDB－Attn 与 CSEM 模块增强后的特征$\{A_1,A_2,A_3,A_4,A_5\}$，采用自顶向下的路径设计，逐层传递语义信息，并在每一层融合当前尺度与上一层的特征图，从而生成尺度对齐的融合特征，FFM 模块如图 4 所示。

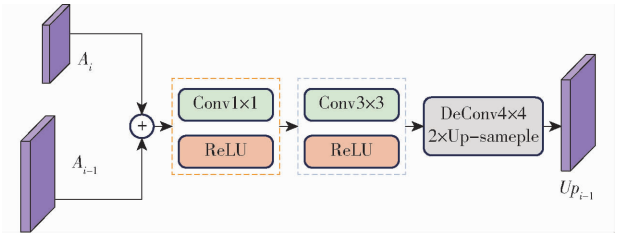

图 4　特征融合模块

Fig. 4　FFM module

A_i为第 i 层经注意力增强得到的特征图，A_{i-1}为前一层提取到的特征图，⊕为逐元素加法，将当前层与前一层特征图进行拼接，并通过Conv 1×1 卷积进行通道对齐，ReLU 为激活函数，再使用 DiConv 3×3 空洞卷积对拼接后的特征提取局部信息，DeConv 4×4 表示反卷积操作（上采样）[16]，采用 4×4 卷积核得到统一尺度下的融合结果Up_i，Up_i表示每相邻两层输出的融合特征图，不仅保留了高层的全局语义信息，也整合了低层的局部细节。

1.4　几何特征稳定模块（GFS）

在 TextBPN＋＋中，分割距离场和分类图存在未注意食品包装文本几何结构上的高宽比失衡的问题，如宽大字体、超长文本行或竖排排列等，仅依赖主干网络提取的特征难以表示文本的几何形态。

针对此问题，MESA－Text 设计了两项改进策略：设计几何特征稳定机制，提升模型对高宽比失衡文本区域的几何特征抓取能力；在迭代边界细化模块中融入绝对位置编码（Absolute position encoding，APE）[17]，增强模型对几何位置信息的感知。

1.4.1　几何稳定机制

食品的横向品牌名称、超长文本行或竖排标签等在几何形态上具有极端高宽比。几何特征稳定模块通过在垂直与水平方向上分别提取文本高宽特征，解决失衡问题。

如图 5 所示，DiConv 为一个 3×3 的空洞卷积[18]，在不增加参数量的前提下扩大感受野，从而获取更丰富的文本特征，再使用 BatchNorm 批量归一化和 PReLU 激活函数增强基本几何特征的表达能力。在此基础上，GFS 分别在基本几何特征的高度维度和宽度维度获取文本高宽几何特征，具体的提取公式为

$$Z = X \odot (\sigma(\mathrm{Convh}_{3\times3}(\mathrm{AvgPoolh}(X)))) \odot (\sigma(\mathrm{Convw}_{3\times3}(\mathrm{AvgPoolw}(X)))) \tag{6}$$

式中　AvgPoolh——高度维度的自适应平均池化

AvgPoolw——宽度维度的自适应平均池化

$\mathrm{Convh}_{3\times3}$、$\mathrm{Convw}_{3\times3}$——3×3 卷积操作

Z——输出的特征图

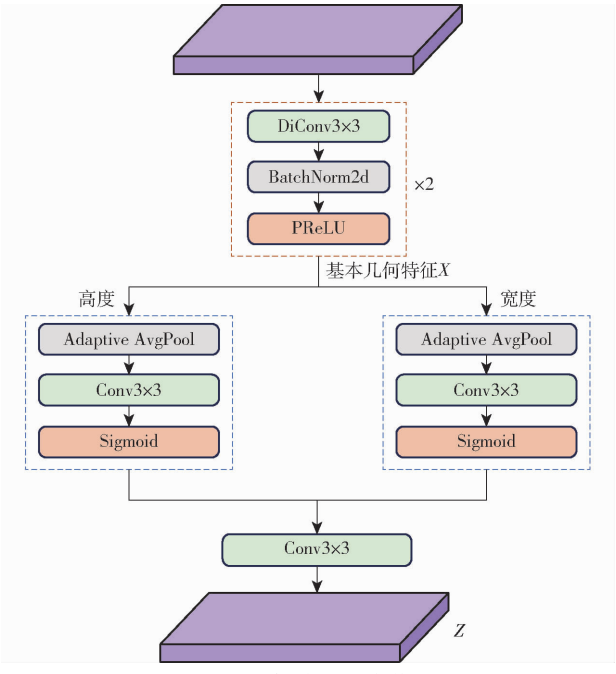

图 5　几何特征稳定模块

Fig. 5　GFS module

几何特征稳定机制通过在高度和宽度两个方向上分别进行自适应平均池化与 3×3 卷积操作，

捕捉文本垂直与水平方向的几何特征,然后将这两条路径的几何信息融合,学习图像中不同几何区域。

输出特征图 Z 的前两个通道被送入分类分支,用于预测像素级的文本前景概率,用于区分图像中属于文本的区域,生成对应的分类图;而后两个通道则用于回归任务,表示文本边界几何结构,即距离场特征,分别回归文本区域中像素点到其水平和垂直方向边界的距离。

经几何特征稳定优化后的输出,能够动态适应不同高宽比例的文本,从而提高对高宽比失调文本几何结构的感知能力。

1.4.2 绝对位置编码

在 TextBPN + + 边界迭代精细化过程中,仅依赖特征的语义信息不足以准确表示文本几何位置。为此,几何特征稳定模块借鉴自然语言处理中的位置建模思想[19-20],引入绝对位置编码,增强对空间位置信息的感知能力。绝对位置编码的具体表达为

$$\begin{cases} PE_{(pos,2i)} = \sin\left(\dfrac{pos}{10\,000^{\frac{2i}{d}}}\right) \\ PE_{(pos,2i+1)} = \cos\left(\dfrac{pos}{10\,000^{\frac{2i}{d}}}\right) \end{cases} \tag{7}$$

式中 $PE_{(pos,2i)}$——位置 pos 在偶数维度上的位置编码值

$PE_{(pos,2i+1)}$——位置 pos 在奇数维度上的位置编码值

i——维度索引

d——编码维度

得到的 PE 为生成的位置编码,在检测食品文本几何边界提供了序列中各个位置的位置信息。

在 TextBPN + + 的边界迭代的前向传播中,将输入的特征张量与位置编码相加,使得输入张量不仅包含文本边界语义信息,还包含其在序列中的位置信息。从而提升模型在边界定位中的判别能力。

本文损失函数沿用 TextBPN + + 的损失函数。

2 实验与结果对比分析

2.1 数据集

本文采用数据集来源于文献[21],通过手机拍摄采集,涵盖了透明包装、彩色塑料袋、纸盒、玻璃瓶、金属罐等多种包装形式,共计收集食品包装图像 509 幅,采用固定划分方案,其中 409 幅用于训练,100 幅用于测试。分割数据集涉及的食品种类丰富,包括茶叶类、肉制品、乳制品、调味品、食用油、干果类及酒类等多个主要食品类别。包含了不同种类与尺寸的食品包装文本,共标注文本框 13 205 个,训练集 10 037 个文本框,测试集 3 168 个文本框,标注内容包括文本位置与对应文本内容,确保数据的一致性与准确性。地址为 https://github.com/xwh-2/data/tree/main/image。

由于数据集偏小,故采用以下数据增广策略提高模型训练效果:对图像亮度和对比度进行中等强度的随机调整取值为 0.5,并且裁剪文本区域并将其尺寸调整为 640 × 640,在高度或宽度方向上,标注框信息保持同步变换,图像会按 0.375 ~ 2.5 之间的随机比例进行缩放,随机旋转图像并填充空白区域,最大旋转角度为 60°,旋转时保留颜色信息,不会对图像进行灰度处理,同时随机水平翻转图像,增强模型的对称性学习能力此外,还将图像按通道内的均值和标准差进行归一化以统一训练数据的分布,增加模型的训练速度。

2.2 实验环境和训练参数

如表 1 所示,本文实验环境为 Ubuntu 22.04 操作系统,CUDA 11.8,显卡型号为 L20,显存容量 48 GB,运行内存容量 100 GB,深度学习框架为 Pytorch。ResNet50[22] 骨干网络使用基于 ImageNet 数据集得到的训练模型。

表 1 实验环境配置

Tab. 1 Experimental environment configuration

名称	配置	型号/版本
操作系统	Ubuntu	22.04
显卡(GPU)	NVIDIA	L20
处理器 CPU	Intel(R) Xeon(R) Platinum	8457C
CUDA	N/A	11.8
深度学习框架	Pytorch	11.2

训练阶段使用 Adam 优化器,batch-size 设置为 8,初始学习率为 0.003。模型参数量为 3.979×10^7(TextBPN + + 为 3.632×10^7),检测速率为 5.93 f/s(TextBPN + + 为 7.78 f/s),单幅推理时间为 3.14 s(TextBPN + + 为 2.65 s)。

2.3 实验结果

为了全面评估 MESA - Text 模型的有效性,本文进行消融实验、模块验证、超参分析和模型对比实验。通过消融实验验证各关键模块对性能的独立贡献;超参数实验则评估置信度阈值、图像分辨率参数设置对性能的影响;此外,还将 MESA - Text 与 TextSnake[23]、TextField[24]、PSENet、DFM[18]、TextBPN + + 等方法在数据集上进行对比,全面验证模型在精确率、召回率与 F1 值等方面的优势。

2.3.1 模块验证实验

与消融实验通过“纵向对比”来验证不同改进

模块效果不同，模块验证实验侧重于“横向比较”的设计分析，即针对某一功能点改进，对比其对应模块采用不同实现方式的差异，以验证模块方法改进和设计的合理有效性。

(1) GDB－Attn 模块验证

为了验证在多尺度特征融合时加入 GDB－Attn 的有效性，分别对未加入注意力的 FPN、加入传统注意力 CBAM 以及加入门控双分支注意力 GDB－Attn 的情况进行评估，实验结果如表 2 所示。

表 2　不同注意力模型效果对比

Tab. 2　Comparison of different attention performances

%

类别	召回率	精确率	F1 值
FPN	89.67	88.06	88.86
FPN＋CBAM	90.83	83.75	87.15
FPN＋GDB－Attn	91.85	89.50	90.66

未加入注意力的 FPN 结果，作为基线对比。加入 CBAM 注意力后，虽然召回率提升至 90.83%，但精确率下降，说明 CBAM 存在对文本的误检。

加入了 GDB－Attn 模块后，较基线召回率提升了 2.18 个百分点，精确率提升了 1.44 个百分点，同时也优于加入 CBAM。结果表明，通过权重平衡因子的自适应门控机制，GDB－Attn 在食品包装文本检测中能够有效捕捉文本特征通道和空间中的关键信息，提升检测食品包装文本中的模糊文本和小文本的能力。

(2) 注意力奇偶层交替融合策略验证

为验证不同 CSEM 模块融合策略在多尺度特征融合中对文本检测性能的影响，本文设计并对比验证了 4 种不同的组合策略：①不引入 CSEM 模块。②每层级串行融合 CSEM。③每层级并行融合 CSEM。④奇偶层交替融合。各策略的实验结果如图 6 所示。

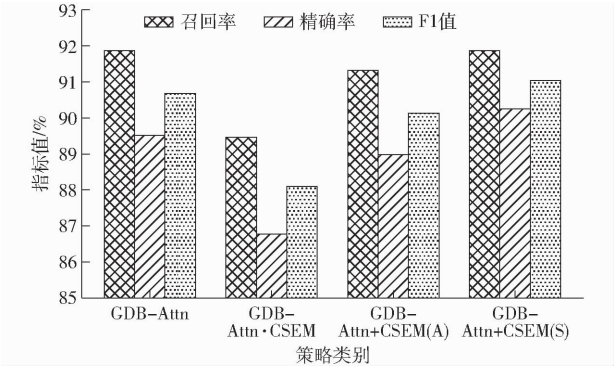

图 6　不同 CSEM 融合策略对比

Fig. 6　Comparison of different CSEM fusion strategies

策略①作为对比基线。策略②将 CSEM 模块与 GDB－Attn 串行叠加于每一层中时，召回率和精确率都下降。这表明单纯地将 CSEM 模块堆叠在门控双分支注意力之后，不仅没有起到添加局部细节的作用，还造成信息干扰。

策略③中(A)表示每一层都采用 GDB－Attn 并行引入 CSEM 模块后，模型性能有所恢复，F1 值提升至 90.12%。策略③增强了各层特征发掘重要细节的能力，但是存在过多的冗余，导致一些原来的关键语义信息被掩盖。

策略④中(S)表示只在奇数层中采用 GDB－Attn 和 CSEM，偶数层只保留 GDB－Attn。模型取得了最佳性能，精确率提高至 90.24%，F1 值提升至 91.04%。在关键层进行平滑，既保证了全局信息获取，又避免了冗余计算和干扰信息，实现了性能和效率的良好平衡。

(3) 不同位置编码验证

在几何特征稳定模块中，位置编码有助于学习文本边界的位置特性，一般位置编码形式有两种，绝对位置编码 APE 和相对位置编码(Relative position encoding，RPE)[25]，为选择更适合的编码实现，对不同的编码方式分别进行了验证，实验结果如图 7、8 和表 3 所示。

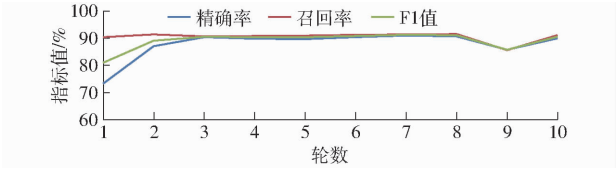

图 7　绝对位置编码结果

Fig. 7　Absolute position encoding results

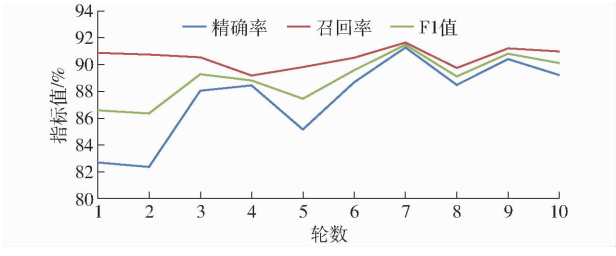

图 8　相对位置编码结果

Fig. 8　Relative position encoding results

表 3　不同编码方式训练的模型对比

Tab. 3　Comparison of model trained with different encoding

%

位置编码	召回率	精确率	F1 值
	91.86	90.24	90.77
APE	91.44	90.95	91.20
RPE	91.64	91.29	91.47

如表 3 所示，尽管相对位置编码在最优结果上表现得比绝对位置编码要更好，但如图 7、8 对比所示，在每隔 20 个 epoch 进行验证的过程中，绝对编

码方式的平均召回率、精确率和 F1 值更高,结果表现得更加稳定,因此本文采用绝对位置编码。

2.3.2 消融实验

为了验证所提出模块的有效性,对各模块进行消融实验,以 TextBPN + + 所采用的 ResNet50 作为基线,逐步验证 DDMFE、GDB - Attn、CSEM、GFS 等模块在食品包装文本数据集上的效果。

如表 4 所示,在仅使用 ResNet50 作为特征提取网络的情况下,模型的精确率、召回率和 F1 值分别为 87.96%、87.57% 和 87.72%。基线模型 ResNet50 本身具备一定的特征提取能力,但仍存在改进空间。

表 4 消融实验结果

Tab. 4 Ablation experiment results %

ResNet50	DDMFE	(MAFEF)GDB - Attn	(MAFEF)CSEM	GFS	精确率	召回率	F1 值
√					87.96	87.57	87.72
√	√				88.06	89.67	88.86
√	√	√			89.50	91.85	90.66
√	√	√	√		90.24	91.86	91.04
√	√	√	√	√	91.76	92.12	91.94

注:√表示运行此模块。

加入 DDMFE 后,虽然精确率只提升了 0.1 个百分点,但召回率提升了 2.1 个百分点,F1 值提升了 1.14 个百分点,表明通过动态可变形卷积改变感受野,模型检测到了更多文本区域。

进一步加入 GDB - Attn 后,MESA - Text 精确率提升了 1.44 个百分点,召回率提升了 2.18 个百分点,F1 值提升了 1.8 个百分点。说明 GDB - Attn 使 MESA - Text 检测到了更多细小、模糊的文本区域,同时也抓住了更多语义的细节从而抑制了误检,提高了精确率。

采用注意力奇偶层交替融合策略引入 CSEM 模块后,精确率和召回率都有所上升,F1 值达到 91.04%。CSEM 通过抑制由注意力堆叠引起的噪声,提升了特征一致性与细节表达能力。

最后引入 GFS 模块后,模型的精确率提升至 91.76%,召回率提升至 92.12%,F1 值达到 91.94%,为所有组合中最优。GFS 模块利用文本在空间几何上的分布特性,抓住文本的几何结构,使模型更加关注文本几何形态变化显著的区域,从而提升边界拟合精度和文本检测完整性。

2.3.3 超参数实验

在超参数实验中,由于置信度阈值[26]大小影响模型判断是否为文本区域,图像分辨率大小[27]影响检测的实际效果,因此对置信度阈值和图像分辨率这两个超参数进行了实验,初始实验使用 TextBPN + + 中的超参,图像分辨率为高分辨率(960 像素 × 2 880 像素)和低分辨率(640 像素 × 960 像素),置信度阈值为 0.80。

(1)置信度阈值参数实验

置信度阈值越高表明判别为文本区域的条件越苛刻,在置信度阈值参数实验中,保持图像分辨率为 960 像素 × 2 880 像素,训练轮数为 50 轮,逐步提升置信度阈值来分析其对模型性能的影响,具体如表 5 所示。

表 5 置信度阈值参数效果对比

Tab. 5 Comparison of effects of confidence parameters

训练轮次	图像分辨率/(像素 × 像素)	置信度阈值	精确率/%	召回率/%	F1 值/%
50	960 × 2 880	0.80	90.75	92.38	91.56
50	960 × 2 880	0.85	91.03	92.25	91.63
50	960 × 2 880	0.90	91.76	92.12	91.94
50	960 × 2 880	0.95	92.10	91.64	91.87

如表 5 的实验结果所示,置信度阈值为 0.80 时,F1 值为 91.56%,召回率虽高,但精度相对较低,说明误检较多。

将置信度阈值上调至 0.85,精确率和 F1 值有所提升,误检得到抑制,性能趋于均衡。当置信度阈值提升至 0.90,模型精确率和 F1 值进一步提升,错误区域过滤能力增强,模型精确率和召回率实现平衡,F1 值达到最高(91.94%),表现最佳。置信度阈值 0.95 时,尽管精确率仍高,但召回率下降,F1 值回落至 91.87%,导致漏检增加。

综上所述,实验结果表明,随着置信度阈值的提升,模型误检率逐渐降低,但当置信度阈值过高时召回能力开始下降,因此选择 0.90 为置信度阈值,兼顾误检控制与文本区域完整性。

(2)图像分辨率参数实验

为分析不同图像分辨率对文本检测模型性能的影响,比较低分辨率(640 像素 × 960 像素)与高分辨率(960 像素 × 2 880 像素)在相同置信度阈值和训练轮数下的表现。实验结果如表 6 所示。在置信

度阈值为 0.80 时，低分辨率图像的 F1 值为 91.84%，略优于高分辨率的 91.56%，说明其在降低误检方面具有一定优势，但同时召回率明显下降，存在漏检风险。

表 6　图片分辨率参数效果对比

Tab.6　Comparison of image size parameter effects

训练轮次	图片分辨率/（像素×像素）	置信度阈值	精确率/%	召回率/%	F1 值/%
50	640×960	0.80	92.59	91.09	91.84
50	960×2 880	0.80	90.75	92.38	91.56
50	640×960	0.85	92.79	90.95	91.86
50	960×2 880	0.85	91.03	92.25	91.63
50	640×960	0.90	93.25	90.66	91.94
50	960×2 880	0.90	91.76	92.12	91.94

当置信度阈值设为 0.85，低分辨率的精确率为 92.79%，F1 值略高于高分辨率但整体在召回率和精确率之间尚未取得理想平衡；当置信度阈值升至 0.90 时，高分辨率精确率和召回率都比较高，达到了一个良好的平衡，同时 F1 值也为最高。

综上，虽然低分辨率图像具备更高精度，适合对误检敏感的场景；而高分辨率图像在召回率上略有优势，更适用于食品包装文本中密集文本检测和复杂背景下的检测，因此选择 960 像素×2 880 像素为图像分辨率。

2.3.4　对比实验

为了验证 MESA－Text 模型食品包装文本检测的性能，将模型与基线模型和已有的基于语义分割模型作对比，所有模型在同一食品包装文本数据集上训练与评估，训练集与测试集按照相同比例划分，且运行在同一硬件和软件环境下，结果如表 7 所示。

表 7　不同检测方法效果对比

Tab.7　Comparison of different detection methods

%

方法	召回率	精确率	F1 值
TextSnake	84.25	66.59	74.38
TextField	85.40	66.37	74.70
PSENet	85.42	66.23	74.61
DFM	87.61	76.50	81.68
TextBPN＋＋	87.57	87.96	87.72
MESA－Text	91.76	92.12	91.94

在食品包装文本检测任务中，传统基于语义分割的文本检测方法虽然具备一定的文本区域提取能力，但在实际应用中仍存在明显局限，以 TextSnake 为例，该方法通过将文本区域建模为一系列圆形片段以灵活适应任意形状文本，理论上具有较强的适应能力。但在食品包装场景中，复杂曲线文本占比较低，TextSnake 的建模方式并未完全发挥优势，导致精度较低，如表 7 所示，精确率仅为 66.59%，F1 值为 74.38%。TextField 利用一个方向场来表示文本区域，使得文本像素可以沿方向场流动到其中心，该方法在一定程度上提升了曲线文本的处理能力，但在复杂背景中可能存在误分割，表现与 TextSnake 相近。PSENet 则采用逐步尺度扩张策略，从文本的核心区域逐步向外扩张，提升了对粘连文本的处理能力，但对于很小的文本其中心区域极小，很容易被忽略，召回能力受到限制，F1 值为 74.61%。这 3 种基于语义分割的文本检测方法表现相对接近，其召回率均在 84% ~85% 区间，精确率约 66%，F1 值约为 74%。这表明它们在发现文本区域方面具有一定能力，但由于背景干扰大、形变文本难以精确分割，精度偏低，导致整体性能受限。

DFM 模型是一种专门针对食品包装文本的模型，结合了注意力机制，如图 9 所示，将召回率提升到 87.61%，精确率至 76.50%，F1 值提升至 81.68%，但是整体性能受限于特征的感受范围和注意力模块的效果，检测的效果一般。

TextBPN＋＋的采用边界迭代模块的精细设计，使得召回率与精确率为 87.57% 和 87.96%，F1 值达到 87.72%，提升了特征表达的一致性和多尺度上下文感知能力，但是尺度融合的不足和对几何特征的忽视使其易发生漏检和误检。

相较之下，MESA－Text 在多尺度增强和几何自适应融合协同作用下，取得了 91.76% 的召回率和 92.12% 的精确率，F1 值为 91.94%，与 TextBPN＋＋相比，MESA－Text 的召回率和精确率均有约 4 个百分点的提升，F1 值提高了 4.22 个百分点，充分证明了多尺度增强和几何自适应融合在应对食品包装中高宽比失调、艺术字、小文本等复杂文本场景时的有效性。

3　结束语

针对食品包装图像中的包装文本检测任务，本文提出了一种基于 TextBPN＋＋的改进模型 MESA－Text，并围绕多尺度特征增强和几何自适应融合进行系统性优化。结果表明，检测效果有了明显提升，具体而言：提出动态可变形多尺度特征提取模块，使模型的感受野更加灵活，提升对艺术字文本的检测能力；提出多尺度注意力特征增强融合模块，通过注意力奇偶层交替融合策略，将门控双分支注意力模块和特征平滑模块集成，解决了模糊、小文本检测不充分的问题；提出几何特征稳定模块，在空间上利用文本的几何结构特征优化了边界回归。解决



(a) 艺术字漏检,真值 (b) 艺术字漏检,TextBPN++ (c) 艺术字漏检,MESA-Text

(d) 模糊、小文本检测,真值 (e) 模糊、小文本检测,TextBPN++ (f) 模糊、小文本检测,MESA-Text

(g) 文本高宽比失调,真值 (h) 文本高宽比失调,TextBPN++ (i) 文本高宽比失调,MESA-Text

图 9 MESA－Text 与 TextBPN＋＋方法效果对比

Fig. 9 Comparison of MESA－Text and TextBPN＋＋

高宽比失调文本的文本检测问题。实验表明,加入动态可变形多尺度特征提取模块、多尺度注意力特征增强融合模块、几何特征稳定模块,模型 F1 值分别提升了 1.14、2.18、0.90 个百分点,相较原始模型总共提升 4.22 个百分点。MESA－Text 在食品包装图像数据集上取得了优异的精确率、召回率与 F1 值,显著优于近年来的文本检测方法,充分验证了其在复杂视觉场景下的有效性与先进性。

参 考 文 献

[1] 王建新, 王子亚, 田萱. 基于深度学习的自然场景文本检测与识别综述[J]. 软件学报, 2020, 31(5): 1465－1496.
Wang Jianxin, Wang Ziya, Tian Xuan. Review of natural scene text detection and recognition based on deep learning[J].

Journal of Software, 2020, 31(5): 1465 – 1496. (in Chinese)
[2] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015.
[3] Liu Y, Jin L, Zhang S. Accurate scene text detection by image pyramid network with local refinement[C]// Proceedings of the International Conference on Document Analysis and Recognition,2019.
[4] Tian Zhiqiang, Shen Chunhua, Chen Hao. Detecting text in natural image with connectionist text proposal network[C]// European Conference on Computer Vision. Amsterdam, 2016.
[5] Zhou Xinyu, Yao Cong, Shi Weiying, et al. EAST: an efficient and accurate scene text detector[C]// IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, 2017.
[6] Liu Yuliang, Chen Hao, Wei Lianwen, et al. Towards robust curve text detection with conditional spatial expansion[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019.
[7] Liao Minghui, Wan Zhaoyi, Yao Cong, et al. Real-time scene text detection with differentiable binarization and adaptive scale fusion[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3): 1034 – 1046.
[8] Zhang Shixue, Yang Chun, Zhu Xiaobin, et al. Arbitrary shape text detection via boundary transformer[J]. IEEE Transactions on Multimedia, 2023, 26: 2771 – 2784.
[9] Bu Qingwen, Park S, Khang M, et al. SRFormer: text detection transformer with incorporated segmentation and regression[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2024.
[10] Su Yuchen, Chen Zhineng, Shao Zhiwen, et al. LRANet: towards accurate and efficient scene text detection with low-rank approximation network[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2024.
[11] Wang Wenhai, Xie Enze, Li Xiang, et al. Shape robust text detection with progressive scale expansion network[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019.
[12] Zhu Xizhou, Hu Han, Lin S, et al. Deformable ConvNets v2: more deformable, better results[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2019.
[13] Hu Jie, Shen Li, Sun Gang, et al. Squeeze-and-excitation networks[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2018.
[14] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]// Proceedings of the European Conference on Computer Vision,2018.
[15] Guo Menghao, Fan Xin, Zhang Rui, et al. EMA – Net: efficient multi-scale attention network for lightweight image super-resolution[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,2021.
[16] Zeiler M D, Fergus R. Visualizing and understanding convolutional networks[C]// Proceedings of the European Conference on Computer Vision, 2014.
[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, 2017.
[18] Yu F, Koltun V. Multi-scale context aggregation by dilated convolutions[C]// International Conference on Learning Representations, 2016.
[19] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]// European Conference on Computer Vision,2020.
[20] Bao Hangbo, Dong Li, Piao Songhao, et al. BEiT: betr pre-training of image transformers[C]// International Conference on Learning Representations,2022.
[21] 田萱,王子亚,王建新. 基于语义分割的食品标签文本检测[J]. 农业机械学报,2020,51(8):336 – 343.
Tian Xuan, Wang Ziya, Wang Jianxin. Text detection of food labels based on semantic segmentation[J]. Transactions of the Chinese Society for Agricultural Machinery,2020,51(8):336 – 343. (in Chinese)
[22] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[23] Long Shangbang, Rhee S, Yao Cong, et al. TextSnake: a flexible representation for detecting text of arbitrary shapes[C]// European Conference on Computer Vision,2018.
[24] Xu Yongchao, Wang Yancheng, Wang Wenhai, et al. TextField: learning a deep direction field for irregular scene text detection[J]. IEEE Transactions on Image Processing, 2019, 28(11): 5566 – 5579.
[25] Shao Zhiwen, Stein J, Giese M A. Relative positional encoding for transformers with linear complexity[C]// International Conference on Machine Learning,2022.
[26] Liu Wei, Anguelov D, Erhan D, et al. SSD: single shot multibox detector[C]// European Conference on Computer Vision, 2016.
[27] Liao Minghui, Bao Yuan, Huang Pengyuan, et al. Real-time scene text detection with differentiable binarization[C]// Proceedings of the AAAI Conference on Artificial Intelligence,2020.