中图法分类号： **文献标识码：** A **文章编号：** 1006-8961(XXXX)XX-0001-16

论文引用格式： Han Xu, Wang Qi. Efficient scene text detection with intra-scale distribution-aware modeling and cross-semantic collaborative reasoning[J/OL]. Journal of Image and Graphics, XXXX:1-16. DOI: 10.11834/jig.260116. (韩旭, 王琦. 尺度内分布感知与跨语义协同推理的高效场景文本检测[J/OL]. 中国图象图形学报, XXXX:1-16. DOI: 10.11834/jig.260116.)[DOI:10.11834/jig.260116]

尺度内分布感知与跨语义协同推理的高效场景文本检测

韩旭[1,2], 王琦[2]

1. 西北工业大学计算机学院, 西安市, 710072; 2. 西北工业大学光电与智能研究院, 西安市, 710072

摘　要： **目的**　现有基于分割的场景文本检测方法多默认不同尺度特征可在同一语义空间中直接融合, 采用统一监督信号驱动多尺度特征学习, 忽略了跨层特征在语义层级上的本质差异, 易导致低层像素噪声与高层语义约束相互干扰, 从而影响检测性能。提出了一种基于尺度内分布感知与跨语义协同推理的高效场景文本检测方法。**方法**　将像素级文本标注提升为多层次分布感知监督, 引导不同尺度特征分支自主学习其对应感受野下的文本分布语义; 在此基础上, 引入跨语义全局知识集成机制, 对多层次特征进行尺度内增强与跨层次协同融合, 从而提升模型对复杂文本结构的整体建模能力。所引入的分支自主分布感知建模仅在训练阶段启用, 测试阶段无需额外计算, 保证了检测精度与推理效率之间的良好平衡。**结果**　在多个公开数据集上, 与现有10余种先进方法进行对比, 本文方法均取得显著提升。相较于先进方法DBNet++(differentiable binarization network++), 提出方法的F值在Total-Text、MSRA-TD500(MSRA text detection 500 database)、CTW(Curve Text in the Wild)1500数据集上分别提升了4.2%、5.0%和2.6%。消融实验进一步验证了所提出模块的有效性。**结论**　实验结果表明, 提出方法在多种场景下均具备良好的检测性能, 同时保持较高的推理效率, 验证了提出方法在高效场景文本检测任务中的可行性。

关键词： 场景文本; 目标检测; 文本检测; 语义分割; 卷积神经网络; 特征感知

Efficient scene text detection with intra-scale distribution-aware modeling and cross-semantic collaborative reasoning

Han Xu[1,2], Wang Qi[2]

1. School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China; 2. School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, 710072, China

Abstract: Objective Scene text detection (STD) is a fundamental task in scene text reading and understanding, and plays an important role in enabling intelligent systems to perceive high-level semantic information from natural scenes. It provides essential technical support for various applications, such as autonomous driving, image retrieval, unmanned systems, and intelligent scene analysis. In recent years, with the rapid development of deep learning and visual representation modeling, STD has achieved substantial progress and attracted increasing research attention. Existing deep learning-based methods can generally be divided into regression-based, connected-component-based, and segmentation-based approaches. Among them, segmentation-based methods have become a mainstream solution due to their flexibility in pixel-

收稿日期： 2026-03-04; **修回日期：** 2026-04-29

* **通信作者：** 王琦, 通信作者, 男, 教授, 主要研究方向为计算机视觉、模型识别和遥感图像处理。E-mail: crabwq@nwpu.edu.cn

基金项目： 国家自然科学基金项目(U21B2041; 62471394)

Supported by: Supported by the National Natural Science Foundation of China under Grant U21B2041 and 62471394.

© 中国图象图形学报版权所有

level prediction and strong capability in detecting arbitrarily shaped text instances. However, most existing segmentation-based methods still implicitly assume that multi-scale features can be optimized under a unified supervision signal and fused within a shared semantic space. Such a strategy overlooks the intrinsic semantic heterogeneity across feature hierarchies. Specifically, low-level features contain rich spatial details but are vulnerable to pixel-level noise, whereas high-level features encode stronger semantic information but may lose fine-grained structural cues. Directly supervising and fusing these heterogeneous representations may lead to interference between low-level pixel noise and high-level semantic constraints, thereby weakening feature fusion effectiveness and reducing inference stability. From the perspective of representation learning, multi-scale features are not merely homogeneous representations at different spatial resolutions, but heterogeneous representations associated with different semantic granularities. Therefore, effective STD requires explicit modeling, alignment, and coordination of semantic information across different feature levels. **Method** To address the above issues, we propose an efficient and effective STD framework, which consists mainly of a branch-wise distribution-aware modeling (BDM) module and a cross-semantic global knowledge integration (CGKI) module. Considering that conventional multi-scale text detection methods often ignore the semantic discrepancies among different feature levels at the supervision stage, the BDM module is designed from the perspective of label modeling. Specifically, it transforms pixel-level binary segmentation annotations into hierarchical distribution-aware supervision signals, enabling feature branches at different scales to independently learn text distribution semantics that are consistent with their corresponding receptive fields. In this way, the semantic interference among heterogeneous multi-scale features can be alleviated, and semantically aligned feature representations can be provided for subsequent feature fusion. Notably, the BDM module is only employed during the training stage and removed during inference, thus improving detection accuracy without introducing additional computational overhead. On the basis of intra-scale distribution-aware semantic alignment, we further design the CGKI module to explicitly model the collaborative relationships among different semantic levels. This module first enhances the representation of each scale within its own semantic space, and then performs controlled cross-scale interaction through adaptive scale reweighting and adjacent-scale information injection. By selectively recalibrating the importance of different scales and introducing complementary contextual information from neighboring levels, the CGKI module achieves global coordination and stable fusion of multi-scale semantics while maintaining a controllable computational cost. The ResNet equipped with deformable convolutions and feature pyramid network (FPN) is adopted as the backbone. For the training stage, the model is either directly trained on public datasets for ablation studies or pre-trained on Synth150k for 10 epochs and then fine-tuned on real-world datasets for comparison experiments. SGD with an initial learning rate of 0.001 and a poly learning rate schedule is used for optimization, together with data augmentation strategies including random rotation, cropping, and flipping. **Result** The proposed method is extensively evaluated against more than ten advanced methods on five widely used public text detection benchmarks, including MSRA-TD500, CTW1500, Total-Text, ICDAR2015, and MPSC. Precision (P), recall (R), and F-measure (F) are adopted as the evaluation metrics, where higher values indicate better detection performance. All inference tests are conducted on a single NVIDIA GTX 1080Ti GPU with an Intel i7-6800K CPU to ensure a consistent evaluation environment. Experimental results show that the proposed method consistently outperforms existing efficient STD methods on the above datasets while maintaining competitive inference speed. Specifically, on Total-Text, the proposed method improves the F-measure by 4.2% and 2.7% compared with DBNet++ and FEPE, respectively. On MSRA-TD500, it achieves F-measure improvements of 5.0% and 4.1% over DBNet++ and FEPE, respectively. On CTW1500, it gains 2.6% and 1.0% in F-measure against DBNet++ and FEPE, respectively. On ICDAR2015, it achieves F-measure gains of 2.8% and 2.7% relative to DBNet++ and FEPE, respectively. On the industrial scene text dataset MPSC, the proposed method surpasses existing advanced methods ISTD-DLA, ODM, and RT-DETR by 1.0%, 3.8%, and 1.3% in F-measure, respectively. Ablation studies on MSRA-TD500 further demonstrate the effectiveness of the proposed modules, confirming that BDM and CGKI can enhance multi-scale feature representation and fusion. In addition, visualization results on these datasets show that the proposed method can generate complete and accurate text boundaries in different scenes. Cross-dataset experiments further verify the generalization ability of the proposed method, where it achieves superior performance over existing representative methods such as ZTD, MTD, and CM-Net under both line-level and word-level annotation settings. **Conclusion** This work presents an efficient and effective scene text detection method.

© 中国图象图形学报版权所有

By integrating BDM and CGKI, the proposed method enhances the semantic consistency and collaborative fusion of multi-scale text features, thereby improving the detection of complex text. Experimental results on multiple public benchmarks demonstrate that the proposed method achieves competitive detection accuracy and inference speed, outperforming existing efficient scene text detection methods. In future work, we will explore the integration of the proposed detection model with efficient text recognition models to establish an end-to-end efficient framework for text spotting.

Key words: scene text; object detection; text detection; semantic segmentation; convolutional neural network; feature awareness

论文引用格式:[DOI:10.11834/jig.260116]

0 引　言

场景文本阅读有助于智能设备理解深层语义场景信息,并为诸多应用(如自动驾驶、图像检索、无人系统等)提供技术支持,显著提升生产效率。作为场景理解的基础任务,场景文本检测在文本阅读中至关重要。近年来,随着深度学习与视觉建模技术的不断突破,场景文本检测方法呈现出快速发展的趋势,吸引了大量研究者的关注。

随着基于深度学习的目标检测与分割技术的迅速发展,场景文本检测(王紫霄 等,2023;陈博伟 等,2024;吕学强 等,2024;师广琛 等,2021)也取得了显著进展。现有研究方法大致可分为三类:基于回归的方法、基于组件合成的方法以及基于分割的方法。基于回归的方法通过一系列参数构建文本轮廓表达模型;基于组件合成的方法则首先检测文本的子单元,再利用后处理手段将同属一个实例的子单元聚合,以重建完整的文本轮廓;而基于分割的方法凭借其灵活的像素级预测能力以及相对简单的后处理,在众多方法中占据重要地位。然而,这种灵活性同时也带来了实例粘连的问题。渐进尺度扩展网络(progressive scale expansion network, PSENet)(Wang 等,2019a)通过将文本实例表示为多尺度内核,并采用逐步扩张的后处理策略来重建文本轮廓,从而缓解了相邻文本粘连的问题。在此基础上,像素聚合网络(pixel aggregation network, PAN)(Wang 等,2019b)基于轻量级骨干网络和高效的轮廓重建机制,实现了实时的场景文本检测。DBNet(Liao 等,2020)创新性地将二值化过程嵌入到训练过程中,使其可微并可端到端优化。其改进版本 DBNet++(Liao 等,2023)进一步引入注意力机制与自适应尺度融合模块,提升了模型对多尺度文本的鲁棒性。Qu 等人(2023)在 DBNet 的基础上动态调整扩张系数,以更精确地完成文本轮廓重建。Yang 等人(2023)则从另一视角改进轮廓重建机制,将原本依赖几何先验的扩张距离改为由模型预测,从而更准确地恢复文本实例形状。同心掩膜网络(concentric mask network, CM-Net)(Yang 等,2022)指出,基于全局特征计算收缩距离的方法可能导致某些文本实例被错误地收缩成多个内核。为此,该网络依据文本实例中最薄弱的区域确定收缩距离,有效缓解了上述问题。在不同文本语义表征方面,为了缓解像素级优化方法和实例级建模任务的不平衡,Han 等人(2026)将实例映射为高斯核,追求实例的等权建模,并在预测、特征融合以及后处理阶段构建了像素与实例的深度交互。与此同时,Han 等人(2025c)基于文本内核的强实例区分性进行轮廓建模,并引入语义完整的文本特征进行监督引导,实现复杂场景下的高精度文本建模。尽管现有方法从多维度改进了基于分割的文本检测模型,但大多仍默认不同尺度特征可在同一语义空间中直接叠加,采用统一的监督信号直接驱动不同尺度特征学习,并在解码阶段进行早期或简单融合。这种策略忽略了跨层次特征在语义层级上的本质差异,容易导致低层像素噪声与高层语义约束相互干扰,从而削弱融合效能并影响推理稳定性。从表示学习的视角分析,多尺度特征并非仅分辨率不同的同质表示,而是对应于不同语义粒度的异质表征。因此,在解码阶段,若缺乏对各层级特征语义表达的显式建模与对齐,直接融合往往难以保证跨尺度语义的一致性。基于这一观察,并结合人类“双目独立建模、中枢信息融合”的感知模式,本文认为解码过程应当遵循“先尺度内独立建模,再跨尺度协同推理”的原则,各层级特征应首先在各自的语义维度内完成表征强化,随后再进行全局层面的跨尺度融合,从而确保异质表征在融合过程中的语义一致性与互补性,并最终输出精准文

© 中国图象图形学报版权所有

本预测。

基于上述动机,提出了一种基于尺度内分布感知与跨语义协同推理的高效场景文本检测方法,该方法由分支自主分布感知建模与跨语义全局知识集成模块共同组成。针对传统多尺度方法在监督层面忽略不同尺度语义差异的问题,本文从标签建模角度出发,将像素级二值标注转换为跨层次的分布感知监督信号,并引入分支自主建模机制,使不同尺度特征在其对应感受野下完成分布语义对齐。该模块有效缓解了跨尺度语义干扰问题,为后续融合提供了语义一致且可解释的特征表示。在尺度内分布语义对齐的基础上,本文进一步设计了跨语义全局知识集成模块,通过尺度级自适应重标定与相邻尺度信息注入,显式建模不同语义层级之间的协同关系。该机制在保持计算开销可控的同时,实现了多尺度语义的全局协调与稳定融合。本文的主要贡献可概括如下:1)提出分支自主分布感知建模模块,通过构建跨层次分布感知监督,实现不同尺度特征的独立语义对齐建模;2)提出跨语义全局知识集成机制,通过尺度级重标定与相邻尺度交互,实现多尺度语义的协同融合;3)基于上述模块设计了一种高效多场景文本检测框架,在保证检测精度的同时具备较高的推理速度,在多个常用公开数据集上的实验结果表明,该方法性能优于现有主流文本检测算法。

1 相关工作

1.1 基于回归的检测方法

早期基于回归的文本检测方法主要基于通用的物体检测方法。TextBoxes(Liao 等,2017)基于单发检测器(single shot detector,SSD)(Liu 等,2016)改进了默认锚框,以适应场景文本的不同尺度和长宽比。在此基础上,TextBoxes++(Liao 等,2018)引入了一个角度参数来处理不同方向的场景文本。多方向场景文本检测器(multi-oriented scene text detector,MOST)(He 等,2021)提出了一个文本特征对齐模块,根据初始预测调整感受野。虽然上述方法取得了较好的效果,但它们无法处理场景中常见的任意形状的文本。为了解决这个问题,Textray(Wang 等,2020)、傅里叶轮廓嵌入网络(Fourier contour embedding network,FCENet)(Zhu 等,2021)和自适应贝塞尔曲线网络(adaptive Bezier-curve network,ABCNet)(Liu 等,2020)分别利用极坐标系、傅里叶向量和贝塞尔曲线来表示文本轮廓。Su 等人(2024)采用离散余弦变换来建模文本实例,可以有效地表示不规则形状的实例。EdgeText(Yang 等,2025)则提出以二次多项式来拟合不规则文本的两条长边。LeafText(Yang 等,2023)仿照叶脉生长的模式来有效重建不规则文本的轮廓。Su 等人(2024)提出了一种新颖的基于奇异值分解的文本轮廓表示方法。Zhang 等人(2021)直接预测文本的轮廓点并迭代更新它们。尽管上述方法可以有效地建模不规则形状的文本,但复杂的网络结构仍然限制了它们的效率。

1.2 基于分割的检测方法

基于分割的方法对输入图像执行像素级预测并引入相应的后处理来重建文本轮廓。PixelLink(Deng 等,2018)预测像素的概率和像素之间的关系以区分像素属于某个实例。Wang 等人(2023)通过预测文本区域和文本间隔来区分不同实例。PSENet(Wang 等,2019a)将文本向内收缩以生成不同尺度的文本核,并通过渐进扩展方法重建实例轮廓。在此基础上,PAN(Wang 等,2019b)利用相似性向量细化轮廓重建方法并采用轻量级主干。相较于PAN,Sheng 等人(2021)提出了无需迭代的像素聚合方法,提高了后处理效率。DBNet(Liao 等,2020)进一步简化后处理方法并将二值化方法引入训练。DBNet++(Liao 等,2023)提出了一种轻量级注意力机制来提取多尺度语义特征。Qu 等人(2023)使用卷积神经网络(convolutional neural networks,CNN)预测扩张系数来解决扩张距离不准确的问题。Yang 等人(2023)则是通过直接预测扩张距离以准确重建文本轮廓。上述方法在生成文本内核时仅考虑了文本实例的全局特征,忽略了轮廓的局部几何形状。CM-Net(Yang 等,2022)根据木桶原理,沿长边中线均匀采样,综合考虑不同区域的局部形状生成文本内核,同时设计了多特征提取模块以辅助训练。与CM-Net类似,Han 等人(2024)基于短边的长度以计算收缩距离。尽管上述方法在多个公开数据集上取得了良好的性能表现,但是这些方法过于关注低层次像素特征,忽略了文本实例的不同层次语义特征建模。

1.3 基于组件合成的检测方法

这些方法将文本分割成多个组件,然后尝试检测组件并合并它们。TextSnake(Long 等,2018)用圆

© 中国图象图形学报版权所有

表示文本组件，通过预测圆的中心线、角度和半径重建文本轮廓。Shi 等人(2017)通过文本片段和文本链接对文本实例进行建模，根据链接将前者合并以重建文本。Tian 等人(2016)检测固定尺寸的文本组件并通过组件连接的方法重建文本实例。上述方法巧妙地表示了任意形状的文本，但同样存在复杂的后处理问题。为了缓解该问题，Su 等人(2025)创新性地将文本检测转换为目标跟踪任务，从而无需后处理，大幅提升了基于片段连接的方法的性能。

2 方　法

2.1 整体结构

提出方法的整体框架如图 1 所示，主要由特征提取网络(由主干网络与特征金字塔网络构成)、分支自主分布感知建模(branch-wise distribution-aware modeling，BDM)模块、跨语义全局知识集成(cross-semantic global knowledge integration，CGKI)模块以及文本内核分割模块组成。训练阶段，首先将输入图像送入特征提取网络，获得多尺度特征表示。随后，分支自主分布感知建模模块将单一的像素级文本分割标注转换为跨层次的文本分布感知标签，并以尺度匹配的方式对各层特征施加监督约束，使不同尺度特征能够在其对应感受野下分支化地学习文本分布表征。在此基础上，跨语义全局知识集成模块对多尺度特征进行进一步处理：一方面执行尺度内特征增强以提升表征质量，另一方面通过跨层

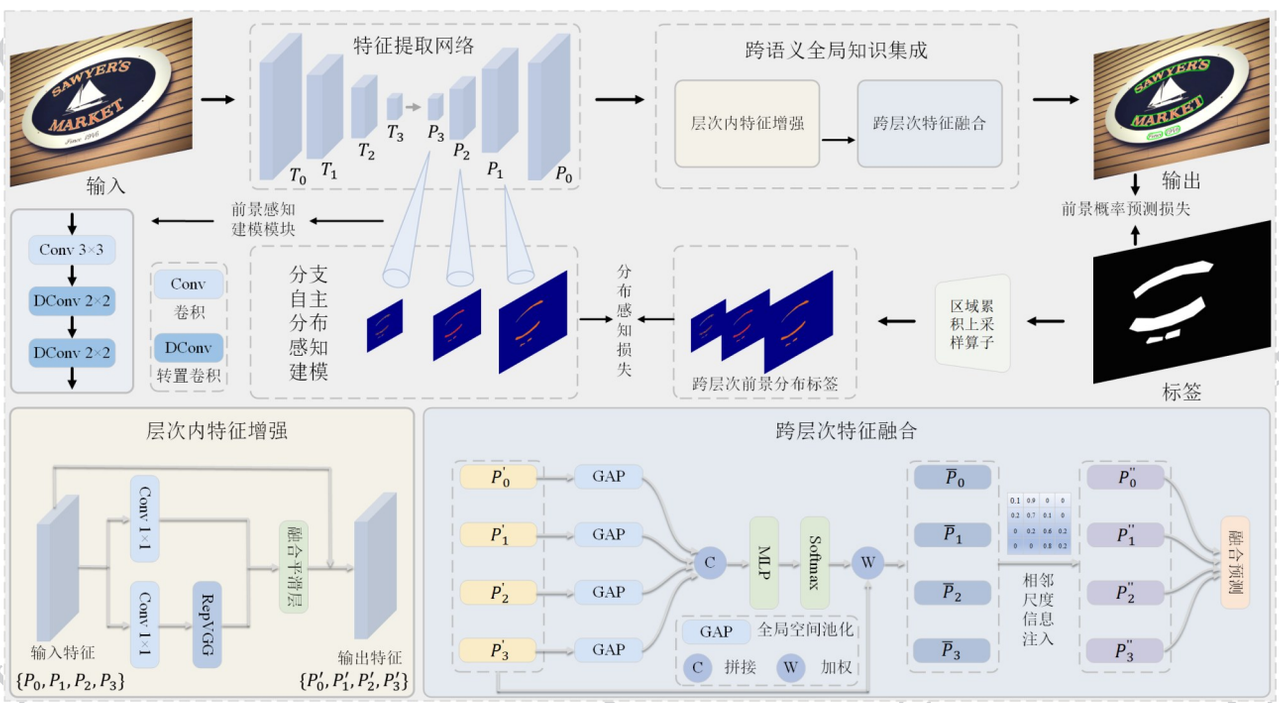


图 1　模型整体流程图

Fig. 1　The overall structure of the proposed model

次的受控融合实现全局语义交互，从而强化模型对前景文本结构的捕捉能力。最终，文本内核分割头输出像素级文本内核概率图，对其结果进行二值化与扩展可获得完整的文本实例检测结果。具体而言，特征提取网络采用带有可变形卷积(Zhu 等，2019)的残差网络(residual network，ResNet)(He 等，2016)作为骨干网络，生成四个空间尺寸分别为输入图像的 1/4、1/8、1/16 和 1/32 的特征图。测试阶段，为降低推理开销，可移除分支自主分布感知建模模块，仅保留主干解码路径进行端到端预测。

2.2 分支自主分布感知建模

现有基于分割的场景文本检测方法依托像素级预测所提供的灵活表征能力，在任意形状文本建模与推理效率方面展现出显著优势。然而，这类方法通常对底层像素信息依赖较强，不同层级特征之间所蕴含的语义差异尚未得到充分重视。事实上，主

© 中国图象图形学报版权所有

干网络通过多层卷积逐级提取特征，不同尺度特征在感受野范围、语义抽象程度以及信息表达形式等方面均存在显著差异。从表示学习的角度来看，多尺度特征并非仅是分辨率不同的同质表示，而是对应于不同语义粒度的异质表征。尽管多尺度融合策略已被广泛应用于现有方法中，但在语义层级尚未对齐的情况下直接进行融合，往往容易引发跨尺度语义干扰，从而影响模型的整体建模稳定性。受人类视觉系统中双目独立感知与协同处理机制的启发，本文从建模范式层面提出分支自主分布感知思想，强调在融合之前应充分刻画不同层级特征各自的语义分布特性。通过鼓励模型在像素信息建模的基础上形成更具层次性的语义认知，该思想为后续的跨尺度协同推理提供了更加一致且可靠的语义基础，有助于提升模型对复杂文本结构的整体理解能力，并缓解噪声信息对检测性能的干扰。具体而言，

$$A(\boldsymbol{Q})_{ij}=\sum_{u=0}^{1}\sum_{v=0}^{1}\boldsymbol{Q}_{2i+u,2j+v}\#(1)$$

式中，$A(\cdot)$表示2×2区域累计算子，$\boldsymbol{Q}$表示给定输入。接着对给定二值标签执行三次2×2区域累计算子以生成不同尺度前景语义分布感知标签：

$$\boldsymbol{S}_0=\boldsymbol{Y}\#(2)$$

$$\boldsymbol{S}_l=A(\boldsymbol{S}_{l-1}),\ l=1,2,3\#(3)$$

式中，$\boldsymbol{Y}\in\{0,1\}^{H\times W}$表示像素级二值标注，$\boldsymbol{S}_0,\boldsymbol{S}_1,\boldsymbol{S}_2,\boldsymbol{S}_3$表示不同层次的样本累计分布表示，$H$和$W$分别表示输入图像的高度和宽度。接着，在解码阶段对每一尺度特征引入独立的语义建模分支。各尺度特征首先通过独立的卷积预测头，生成与其感受野相匹配的中间表征，从而在尺度内完成语义聚合与增强。这一设计使得每一层级特征能够专注于其所擅长的语义粒度。具体而言，给定输入的图像经由主干网络处理后获得多尺度特征图，经由特征金字塔网络（feature pyramid network，FPN）（Lin等，2017）、特征平滑以及通道压缩，获取不同尺度的特征图$\boldsymbol{P}_0\in\mathbf{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$，$\boldsymbol{P}_1\in\mathbf{R}^{\frac{H}{8}\times\frac{W}{8}\times C}$，$\boldsymbol{P}_2\in\mathbf{R}^{\frac{H}{16}\times\frac{W}{16}\times C}$，$\boldsymbol{P}_3\in\mathbf{R}^{\frac{H}{32}\times\frac{W}{32}\times C}$，其中$C$表示特征通道数。不同尺度的前景感知建模分支可表述为：

$$\boldsymbol{Z}_l^1=\rho\left(\mathrm{K}_l(\boldsymbol{P}_l)\right),\ l=1,2,3\#(4)$$

$$\boldsymbol{Z}_l^2=\rho\left(\mathrm{D}_l(\boldsymbol{Z}_l^1)\right),\ l=1,2,3\#(5)$$

$$\boldsymbol{M}_l=\mathrm{ReLU}\left(\mathrm{D}_l(\boldsymbol{Z}_l^2)\right),\ l=1,2,3\#(6)$$

式中，$\rho(x)=\mathrm{ReLU}(\mathrm{BN}(x))$，BN表示批归一化（batch normalization），ReLU表示修正线性单元（rectified linear unit），K_l表示3×3卷积，D_l为核大小2×2、步长为2的转置卷积，$\boldsymbol{M}_l$为分支l在该层级的分布感知预测，$\boldsymbol{Z}_l^1$和$\boldsymbol{Z}_l^2$为中间层特征。最后，将各分支预测结果与对应尺度分布标签进行损失计算以优化模型，提升各分支独立建模文本语义的能力。

2.3 跨语义全局知识集成

分支自主分布感知建模鼓励不同层次特征图进行分布语义对齐以获取尺度匹配的语义表示。然而，仅依赖尺度内感知建模难以充分提取文本的跨尺度一致性特征。据此，本文进一步提出跨语义全局知识集成模块，将独立建模后的多尺度语义表示进行融合表征，以提高模型的综合感知能力。

具体而言，为提升分支自主分布感知建模后的尺度内表征质量，引入层次内特征增强算子$\mathrm{E}(\cdot)$，对不同尺度特征进行独立增强：

$$(\boldsymbol{P}_0',\boldsymbol{P}_1',\boldsymbol{P}_2',\boldsymbol{P}_3')=\mathrm{E}(\boldsymbol{P}_0,\boldsymbol{P}_1,\boldsymbol{P}_2,\boldsymbol{P}_3)\#(7)$$

式中，$\mathrm{E}(\cdot)$逐尺度作用，为不同尺度特征施加同构非线性变换与残差式重组，可表述为：

$$\boldsymbol{P}_k'=\boldsymbol{P}_k+r_k(\boldsymbol{P}_k),\ k\in\{0,1,2,3\}\#(8)$$

式中，$r_k(\cdot)$表示尺度内的增强映射，进一步表述为：

$$r_k(\boldsymbol{P}_k)=\varphi_k\left(\left[\mathrm{a}_k(\boldsymbol{P}_k),\mathrm{b}_k(\boldsymbol{P}_k)\right]\right)\#(9)$$

式中，$\varphi_k(\cdot)$表示特征融合算子，$\mathrm{a}_k(\cdot)$和$\mathrm{b}_k(\cdot)$分别表示1×1卷积模块与重参数化视觉几何组（reparameterized visual geometry group，RepVGG）（Ding等，2021）模块。在获得尺度内增强特征后，接下来显式建模跨尺度语义协同关系。首先利用全局空间池化计算不同尺度特征的全局统计表示：

$$\boldsymbol{g}_k=\mathrm{GAP}(\boldsymbol{P}_k')\in\mathbf{R}^{1\times C},\ k\in\{0,1,2,3\}\#(10)$$

式中，$\mathrm{GAP}(\cdot)$表示全局池化算子，$\boldsymbol{g}_k$表示第k个尺度特征的全局统计表示。经由特征拼接得到尺度向量：

$$\boldsymbol{g}=\mathrm{concat}(\boldsymbol{g}_0,\boldsymbol{g}_1,\boldsymbol{g}_2,\boldsymbol{g}_3)\in\mathbf{R}^{4\times C}\#(11)$$

$\mathrm{concat}(\cdot)$表示特征拼接。对尺度向量执行轻量映射：

$$\boldsymbol{w}=\mathrm{softmax}\left(\frac{(\boldsymbol{W}_2(\boldsymbol{W}_1\boldsymbol{g}))}{\tau}\right)\in\mathbf{R}^{4\times C}\#(12)$$

式中，$\boldsymbol{w}$为权重，$\boldsymbol{W}_1$和$\boldsymbol{W}_2$为可学习参数，τ为温度系数。接着对各尺度特征执行重标定：

© 中国图象图形学报版权所有

$$\bar{\boldsymbol{P}}_k = \boldsymbol{P}'_k \odot \boldsymbol{w}_k,\ k \in \{0,1,2,3\} \#(13)$$

式中，$\odot$表示逐元素乘法。在重标定特征$\{\bar{\boldsymbol{P}}_k\}$基础上引入相邻尺度信息注入以建立局部交互关系：

$$\boldsymbol{P}''_0 = \bar{\boldsymbol{P}}_0 + \alpha_{01}\mathrm{T}_{1\to 0}(\bar{\boldsymbol{P}}_1) \#(14)$$

$$\boldsymbol{P}''_1 = \bar{\boldsymbol{P}}_1 + \alpha_{10}\mathrm{T}_{0\to 1}(\bar{\boldsymbol{P}}_0) + \alpha_{12}\mathrm{T}_{2\to 1}(\bar{\boldsymbol{P}}_2) \#(15)$$

$$\boldsymbol{P}''_2 = \bar{\boldsymbol{P}}_2 + \alpha_{21}\mathrm{T}_{1\to 2}(\bar{\boldsymbol{P}}_1) + \alpha_{23}\mathrm{T}_{3\to 2}(\bar{\boldsymbol{P}}_3) \#(16)$$

$$\boldsymbol{P}''_3 = \bar{\boldsymbol{P}}_3 + \alpha_{32}\mathrm{T}_{2\to 3}(\bar{\boldsymbol{P}}_2) \#(17)$$

式中，$\mathrm{T}_{j\to i}(\cdot)$表示将第$j$个尺度的特征重采样并对齐至第$i$个尺度的特征的空间分辨率，$\alpha_{ij}\in\mathbf{R}$为可学习标量系数。引入可学习系数是为了自适应地调节相邻层级间语义流动的强度，增强模型对不同尺度目标的鲁棒性。随后，为了消除不同深度特征层间的空间分辨率差异，以最高分辨率特征$\boldsymbol{P}''_0$为基准，将不同尺度特征映射至同一尺度空间：

$$\boldsymbol{P}'''_k = \mathrm{T}_{k\to 0}(\boldsymbol{P}''_k),\ k \in \{0,1,2,3\} \#(18)$$

执行通道维拼接得到融合表征并通过映射执行语义平滑，实现前景像素概率的预测。

$$\boldsymbol{F} = \mathrm{concat}(\boldsymbol{P}'''_0, \boldsymbol{P}'''_1, \boldsymbol{P}'''_2, \boldsymbol{P}'''_3) \#(19)$$

$$\boldsymbol{X} = \sigma(\mathrm{h}(\boldsymbol{F})) \#(20)$$

式中，$\boldsymbol{F}$表示多层次拼接特征，$\boldsymbol{X}$表示前景概率图，$\sigma(\cdot)$表示sigmoid算子，$\mathrm{h}(\cdot)$由一层3×3卷积层与两层转置卷积层级联而成，确保输出预测图与输入图像的分辨率严格对齐。

2.4 损失函数

为了优化提出方法，本文共对两项任务进行约束，分别是文本内核预测和分支自主分布感知。总损失函数由两种损失函数加权组成，定义如下：

$$L_{\mathrm{t}} = \mu \times L_{\mathrm{m}} + L_{\mathrm{d}} \#(21)$$

式中，L_{t}，L_{m}和L_{d}分别表示总损失，文本内核预测损失和分支自主分布感知损失，μ表示权重系数。前者旨在约束生成像素级预测图任务，该任务的预测结果通过后处理算法以提取实例轮廓。后者约束的任务仅作为辅助监督参与训练，在推理阶段可进行移除。这一策略在有效增强模型特征表征能力的同时，实现了推理阶段零额外计算开销。

2.4.1 文本内核损失函数

采用二元交叉熵（binary cross-entropy，BCE）损失来优化文本内核预测。文本内核区域通常只占图像的小部分，为了缓解正负样本的不平衡，本文引入了困难负样本挖掘，其公式如下：

$$L_{\mathrm{m}} = \frac{\sum_{p\in U} -x^*_p \times \log(x_p) - (1-x^*_p) \times \log(1-x_p)}{|\boldsymbol{U}|} \#(22)$$

式中，$\boldsymbol{U}$和$|\boldsymbol{U}|$表示选定的像素集合和其中的像素数量，x^*_p和x_p表示像素p位置的文本内核标签和预测。

2.4.2 分支自主分布感知损失

采用比率损失L_{r}以优化分支自主分布感知建模任务，具体表述为：

$$L_{\mathrm{r}}(\boldsymbol{G},\boldsymbol{I}) = \log\left(\frac{\max(\boldsymbol{G},\boldsymbol{I})}{\min(\boldsymbol{G},\boldsymbol{I})}\right) \#(23)$$

式中，$\boldsymbol{G}$和$\boldsymbol{I}$分别表示输入标签和预测。分支自主分布感知建模损失可以表述为：

$$L^l_{\mathrm{d}} = L_{\mathrm{r}}(\boldsymbol{S}_l,\boldsymbol{M}_l),\ l \in \{1,2,3\} \#(24)$$

$$L_{\mathrm{d}} = \lambda_1 \times L^1_{\mathrm{d}} + \lambda_2 \times L^2_{\mathrm{d}} + \lambda_3 \times L^3_{\mathrm{d}} \#(25)$$

式中，$\boldsymbol{S}_1$，$\boldsymbol{S}_2$，$\boldsymbol{S}_3$和$\boldsymbol{M}_1$，$\boldsymbol{M}_2$，$\boldsymbol{M}_3$分别表示不同层次文本分布的标签和预测值，λ_1，λ_2和λ_3表示权重系数。

2.5 标签与后处理

文本内核标签参照DBNet生成。在后处理阶段，首先对预测进行二值化处理并提取实例轮廓，其次过滤置信度较低的实例，最后采用Vatti clipping（Vatti，1992）算法扩张文本内核生成文本轮廓。

3 实验结果与分析

3.1 数据集

Synth150k（Liu 等，2020）是一个包含15万张合成图像的大规模数据集，其数据规模显著大于真实场景文本数据集，因此本文主要利用该数据集对模型进行预训练，以提升模型的基础表征能力和鲁棒性。MSRA-TD500（Yao等，2012）是一个行级标注的多方向文本数据集，训练集和测试集分别包含300张和200张图像。由于其训练样本数量较少，本文参考已有工作引入HUST-TR400（Huazhong University of Science and Technology text recognition）（Yao等，2014）对训练集进行补充。CTW1500（Liu 等，2017）同样是行级标注数据集，包含1000张训练图像和500张测试图像，并引入了弯曲文本实例。Total-Text（Ch'ng和Chan，2017）同样包含水平、多方向和弯曲等多种几何形式的文本，但采用词级标注方式，共包含1255张训练图像和300张测试图像。ICDAR2015（International Conference on Document

© 中国图象图形学报版权所有

Analysis and Recognition)(Karatzas等,2015)为词级标注数据集,主要包含多方向文本实例,由1000张训练图像和500张测试图像组成。MPSC(metal part surface character)(Guan等,2022)是一个工业场景文本数据集,主要面临光照不均、对比度较低等挑战,包含2555张训练图像和639张测试图像。上述数据集在样本规模、标注粒度、场景类型及文本形态等方面存在一定差异。为减轻其对训练的影响,本文采用合成数据预训练、真实数据集微调的实验设置,并在部分数据集引入辅助数据。

3.2 评价指标

本文沿用场景文本检测任务中常用指标精确率(precision,P)、召回率(recall,R),F值(f-measure,F)和每秒处理帧数(frames per second,FPS)对方法性能进行评估。其中,P表示检测结果中真实文本实例所占的比例,反映方法抑制误检的能力;R表示真实文本实例中被正确检测出的比例,反映方法发现文本目标的能力;F值则综合衡量精确率与召回率,用于评价方法的整体检测性能;FPS表示模型在单位时间内能够处理的图像帧数,用于衡量方法的推理速度,数值越高表示模型的检测效率越高。

3.3 补充细节

训练的批次大小设置为16。采用了两种训练策略:模型直接在公开数据集上进行训练;模型在Synth150k数据集上预训练10轮,然后在真实数据集上进行微调。优化器和初始学习率分别设置为随机梯度下降(stochastic gradient descent,SGD)和0.001。损失函数中的系数分别设定为:μ=6,λ_1=0.5,λ_2=0.25,λ_3=0.125。在训练阶段遵循"poly"策略调整学习率。数据增强包括随机旋转、裁剪和翻转。测试阶段,同一数据集的图像会调整短边至相同大小。所有测试均在单张GTX 1080Ti GPU与Intel i7-6800K CPU的硬件环境下进行。

3.4 消融实验

为了验证提出的各模块的有效性,本文共设置四组实验方案:1)仅预测文本内核的基线方案;2)在基线方案上引入分支自主分布感知建模机制;3)在基线方案上引入跨语义全局知识集成模块;4)同时引入上述两个模块。所有实验均在MSRA-TD500数据集上进行,并统一采用ResNet18作为特征提取骨干网络,以保证比较的公平性。

表1 不同模块组合在MSRA-TD500数据集上的结果对比

Table 1 Performance comparison of different module combinations on the MSRA-TD500 dataset.

内核预测	分支自主建模	全局知识集成	精确率(%)	召回率(%)	F值(%)
√	×	×	87.0	80.2	83.5
√	√	×	89.2	82.1	<u>85.5</u>
√	×	√	88.1	82.8	85.4
√	√	√	90.8	84.9	**87.8**

注:黑色字体表示最优结果,下划线表示次优结果。"√"和"×"分别表示是否启用该模块。

表2 不同跨尺度特征交互方案在MSRA-TD500上的结果

Table 2 Performance of various cross-scale feature interaction schemes on MSRA-TD500.

方案	精确率(%)	召回率(%)	F值(%)
固定数值相加	88.2	85.1	<u>86.6</u>
可学习相邻尺度信息注入	90.8	84.9	**87.8**

注:黑色字体表示最优结果,下划线表示次优结果。

表3 不同推理方案在MSRA-TD500上的结果

Table 3 Performance of different inference schemes on MSRA-TD500.

各分支预测是否参与推理	精确率(%)	召回率(%)	F值(%)
是	87.6	84.7	<u>86.1</u>
否	90.8	84.9	**87.8**

注:黑色字体表示最优结果,下划线表示次优结果。

3.4.1 分支自主分布感知建模的有效性

表1给出了分支自主分布感知建模模块的消融实验结果。与仅进行文本内核预测的基线方案相比,引入分支自主分布感知建模后,模型在MSRA-TD500数据集上的精确率、召回率和F值分别提升了2.2%、1.9%和2.0%。这一结果表明,对不同尺度分支进行针对性的分布建模,有助于缓解统一监督信号下不同层级特征之间的语义干扰,使各尺度特征能够在其对应感受野范围内学习更加匹配的文本分布表示,从而提升多层次文本特征的表征质量。

© 中国图象图形学报版权所有

在此基础上，进一步结合跨语义全局知识集成模块后，模型的F值相较于仅引入分支自主分布感知建模的方案继续提升了2.4%。这一现象说明，分支自主分布感知建模不仅能够改善各尺度特征的局部语义表达，还能够为后续跨层语义交互提供更加稳定的特征基础。因此，该模块的作用并非仅体现在性能增强上，还体现在其对后续跨尺度融合过程的支撑作用上，二者具有明显的协同增益关系。

3.4.2　跨语义全局知识集成模块的有效性

为进一步验证跨语义全局知识集成模块在多尺度特征协同建模中的作用，本文开展了相应的消融实验。如表1所示，在不引入分支自主分布感知建模的情况下，仅引入跨语义全局知识集成模块即可使模型在MSRA-TD500数据集上的F值较基线方案提升1.9%，表明该模块能够有效地融合不同尺度的文本特征。当跨语义全局知识集成模块与分支自主分布感知建模模块联合使用时，模型的F值在该数据集上进一步提升2.3%。此外，为进一步验证跨语义全局知识集成模块在跨尺度融合过程中的稳定性，本文将其与固定权重的跨尺度交互方案进行了对比。如表2所示，得益于可学习参数对不同尺度语义贡献的自适应调节，所提方法在F值上实现了1.2%的提升。该结果表明，相较于静态融合方式，所提模块能够提高跨尺度语义对齐的充分性与融合过程的稳健性。此外，本文对各分支预测是否参与推理的结果进行了进一步验证。如表3所示，当各分支预测结果直接参与最终推理时，模型性能未得到提升，F值反而下降至86.1%。原因在于，不同分支预测来源于异构尺度特征，若在推理阶段强行进行统一对齐与融合，容易引入插值误差问题，特别是在文本边界区域会带来定位偏差。同时，各分支预测参与推理还会增加额外推理开销。因此，本文采用各分支预测仅参与训练、不参与推理的设计。

3.4.3　讨论

上述实验结果表明，本文两个模块的性能提升并非简单叠加所得，而是分别针对多尺度文本建模中的不同关键问题发挥作用。分支自主分布感知建模侧重于提升各尺度特征的语义表征质量，跨语义全局知识集成模块则进一步促进不同层级特征之间的协同与互补。二者联合后能够在保持轻量骨干网络的前提下，有效增强模型对复杂背景、尺度变化及形态多样文本的适应能力。

3.5　与现有先进方法的对比

为了验证提出方法的有效性，本文在MSRA-TD500、CTW1500、Total-Text、ICDAR2015和MPSC数据集上与现有的先进高效方法进行了对比。结合本文在多个数据集上的实验结果可以进一步说明，所提方法的改进并不依赖于特定场景分布，而是对高效场景文本检测任务具有较好的泛化潜力。

3.5.1　在MSRA-TD500数据集上的对比

MSRA-TD500是一个包含多方向长文本的行级标注数据集，广泛用于评估方法对多方向布局和长文本实例的检测能力。如表4所示，本文方法在该数据集上的精确率、召回率和F值分别达到92.3%、88.0%和90.1%。与现有代表性高效方法ZTD（zoom text detector）和FEPE（focus entirety and perceive environment）相比，本文方法在F值上分别提升了3.3%和4.1%。上述结果表明，所提出方法在多方向长文本场景下具有较强的检测能力。其原因在于，本文方法能够有效协调不同尺度特征间的语义差异，使模型在面对方向变化较大、跨度较长的文本实例时，仍能保持较好的文本区域完整性与定位准确性。

进一步地，在获得较高检测精度的同时，本文方法仍保持了较快的推理速度，说明其性能提升并非依赖于更复杂的模型结构，而是通过优化多尺度语义建模与特征协同实现的。

3.5.2　在CTW1500数据集上的对比

为验证本文方法在任意形状行级文本检测任务中的性能，本文在CTW1500数据集上与多种现有先进方法进行了对比实验。如表4所示，RSMTD（reinforcement shrink-mask for text detection）（Yang 等，2023）通过自适应预测文本扩张距离来重建文本轮廓，在该数据集上取得了87.8%的精确率、80.3%的召回率和83.9%的F值。相比之下，本文方法在上述三项指标上分别提升了0.7%、4.2%和2.6%，在检测精度和召回能力上均表现出更优结果。这一结果说明，本文方法对于弯曲形变明显、边界变化复杂的任意形状文本同样具有较好的适应能力。究其原因，分支内的分布感知建模能够增强不同尺度特征对局部文本结构的表征能力，而跨层语义协同进一步提高了多尺度特征融合时的稳定性，从而有助于提升复杂形状文本区域的完整检测能力。值得注意的是，在采用ResNet18作为主干网络的轻量配置

© 中国图象图形学报版权所有

表4　现有高效方法在MSRA-TD500、CTW1500和Total-Text数据集上的性能对比

Table 4　Performance comparison of existing efficient methods on the MSRA-TD500, CTW1500, and Total-Text datasets.

方法	MSRA-TD500				CTW1500				Total-Text			
	精确率(%)	召回率(%)	F值(%)	FPS	精确率(%)	召回率(%)	F值(%)	FPS	精确率(%)	召回率(%)	F值(%)	FPS
PAN(Wang等,2019b)	84.4	83.8	84.1	30.2	86.4	81.2	83.7	39.8	89.3	81.0	85.0	39.6
DBNet(Liao等,2020)	90.4	76.3	82.8	62	84.8	77.5	81.0	55	88.3	77.9	82.8	50
CT(Sheng等,2021)	90.0	82.5	86.1	34.8	88.3	79.9	83.9	40.8	90.5	82.5	86.3	40
PAN++(Wang等,2022)	85.3	84.0	84.7	32.5	87.1	81.1	84.0	36.0	89.9	81.0	85.3	38.3
CMNet(Yang等,2022)	89.9	80.6	85.0	41.7	86.0	82.2	84.1	50.3	88.5	81.4	84.8	49.8
HFENet(Liang等,2023)	89.7	81.1	85.2	40.9	85.1	81.2	83.1	32.2	85.7	81.7	83.7	22.0
FS(Wang等,2023)	90.4	81.6	85.3	25.4	84.6	77.7	81.0	35.2	85.8	77.0	81.1	33.5
RSMTD(Yang等,2023)	89.8	83.1	86.3	62.5	87.8	80.3	83.9	72.1	88.5	83.8	86.1	70.9
DBNet++(Liao等,2023)	87.9	82.5	85.1	55	86.7	81.3	83.9	40	87.4	79.6	83.3	48
ZTD(Yang等,2024)	91.6	82.4	86.8	59.2	88.4	80.2	84.1	76.9	90.1	82.3	86.0	75.2
FEPE(Han等,2025a)	89.4	82.8	86.0	62	88.8	83.0	85.5	55	90.8	79.5	84.8	50
STD(Han等,2025d)	92.2	83.2	87.4	33.4	88.7	84.1	86.3	30.6	–	–	–	–
CIC(Han等,2025c)	91.6	86.3	88.8	40.9	88.2	85.2	**86.5**	50.1	88.8	84.6	86.6	40.9
本文	92.3	88.0	**90.1**	50.1	88.5	84.5	**86.5**	51.2	88.5	86.7	**87.5**	48.5

注:黑色字体表示最优结果,下划线表示次优结果。“–”表示对应方法没有报告该数据集的结果。

下,本文方法的检测性能仍优于大多数非高效方法,表明该方法的性能增益并非主要来源于模型规模扩大,而是来自更有效的特征建模与跨层交互机制。因此,本文方法在保持较低计算复杂度的同时,仍表现出良好的检测鲁棒性和实际应用价值。

3.5.3　在Total-Text数据集上的对比

Total-Text是一个词级标注的包含任意形状文本的数据集,对检测方法的几何建模能力和特征拟合能力提出了较高要求。如表4所示,本文方法在F值上分别较现有高效方法FEPE(Han等,2025a)和DBNet++(Liao等,2023)提升了2.7%和4.2%。上述结果表明,本文方法在词级标注任意形状文本检测任务中具有较强的适应能力。其原因在于,所提出方法能够通过分支内分布感知建模增强不同尺度特征对局部文本结构的表达能力,并通过跨层语义协同进一步提升多尺度特征融合的稳定性,从而更好地拟合弯曲文本和复杂形变文本的几何边界。

3.5.4　在ICDAR2015数据集上的对比

ICDAR2015数据集中的图像通常具有背景复杂、文本模糊等特点,对文本检测方法的鲁棒性提出了较高要求。如表5所示,本文提出的尺度内独立建模策略通过对各层级特征语义的显式表征增强,有效抑制了背景噪声对文本区域的干扰,使精确率达到89.0%。与代表性方法ZTD(Yang等,2024)、DBNet++(Liao等,2023)和CMNet(Yang等,2022)相比,本文方法的F值分别提升了2.8%、2.8%和2.0%,验证了其在复杂场景文本检测任务中的有效性与竞争力。受益于该策略推理阶段可移除的特性,在保持高推理精度的同时,仍能保持41FPS的推理速度,实现了精度和效率的平衡。

3.5.5　在MPSC数据集上的对比

工业场景中的文本通常存在对比度低、表面腐蚀严重及纹理复杂等问题,给文本检测带来较大挑战。为验证所提出方法在工业场景文本检测任务中的适用性与鲁棒性,本文在MPSC数据集上进行了实验。如表6所示,现有先进方法中,ODM(OCR-text destylization modeling)(Duan等,2024)和RT-DETR(real-time detection transformer)(Zhao等,2024)的F值分别为83.4%和85.9%。得益于本文方法对不同层级特征语义的显式建模以及高效融合

© 中国图象图形学报版权所有

表5　现有高效方法在ICDAR2015数据集上的性能对比

Table 5　Performance comparison of existing efficient methods on the ICDAR2015 dataset.

方法	精确率(%)	召回率(%)	F值(%)	FPS
PAN(Wang等,2019b)	84.0	81.9	82.9	26.1
DBNet(Liao等,2020)	86.8	78.4	82.3	48
PAN++(Wang等,2022)	85.9	80.4	83.1	28.2
CMNet(Yang等,2022)	86.7	81.3	83.9	34.5
FS(Wang等,2023)	88.1	78.8	83.2	15.3
DBNet++(Liao等,2023)	90.1	77.2	83.1	44
ZTD(Yang等,2024)	87.5	79.0	83.1	48.3
FEPE(Han等,2025a)	87.3	79.4	83.2	48
STD(Han等,2025d)	89.3	81.1	<u>85.0</u>	23.7
本文	89.0	82.9	**85.9**	41.0

注:黑色字体表示最优结果,下划线表示次优结果。

表6　现有方法在MPSC数据集上的性能对比

Table 6　Performance comparison of existing methods on the MPSC dataset.

方法	精确率(%)	召回率(%)	F值(%)	FPS
PSENet(Wang等,2019a)	85.4	78.4	81.8	–
PAN(Wang等,2019b)	87.1	81.6	84.2	–
FCENet(Zhu等,2021)	87.1	81.6	84.3	–
RFN(Guan等,2022)	89.3	83.3	<u>86.2</u>	–
DBNet++(Liao等,2023)	86.1	81.2	83.3	–
RITD(Yang等,2025)	87.9	82.3	85.4	32
ODM(Duan等,2024)	86.2	81.7	83.4	–
RT-DETR(Zhao等,2024)	84.5	87.3	85.9	–
ISTD-DLA(Hu等,2025)	88.6	84.2	<u>86.2</u>	32
本文	90.8	83.8	**87.2**	46.7

注:黑色字体表示最优结果,下划线表示次优结果。

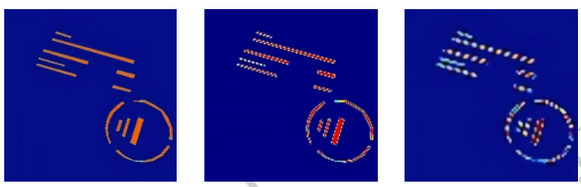
策略,能够有效应对复杂的工业背景,其不仅在性能上分别超过上述方法3.8%和1.3%,在推理效率上也大幅领先,进一步验证了所提方法在复杂工业场景中的优势。

3.6　可视化

为了更清晰地展示本文提出的跨层次联合分布感知方法,图2对所使用的标签进行了可视化。从左至右、自上而下依次为:(a)输入图、(b)标签二分类图、(c)实例轮廓图以及三种不同层次的正样本分布图(d)-(f)。图3给出了本文方法与现有先进方法的可视化对比结果,FEPE(Han等,2025),KPN(kernel proposal network)(Zhang等,2023)和DBNet++(Liao等,2023)错误地将某些纹理相似的花纹误判为文本,KPN出现了部分漏检现象。此外,上述方法未能完整地建模长文本实例。受益于分支自主分布感知建模与跨语义全局知识集成模块对不同层次文本特征的有效融合,相较于上述方法,提出的方法能够很好地应对上述问题。图4展示了在MSRA-TD500、CTW1500、Total-Text、ICDAR2015和MPSC数据集上的部分预测结果。可以得出,本文方法能够较好地适应行级和词级任意形状文本的检测,并同时适用于中英文混合场景。

© 中国图象图形学报版权所有

（a） 输入图　（b） 标签二分类图　（c） 文本轮廓图



（d） 二倍累计分布图　（e） 四倍累计分布图　（f） 八倍累计分布图

((a) input image;(b) binary label map;(c) text contour map;(d) 2× cumulative distribution map;(e) 4× cumulative distribution map;(f) 8× cumulative distribution map))

图2 标签可视化

Fig. 2 Visualization of the generated labels.

图3 本文方法与其他先进方法的可视化对比

Fig. 3 Visual comparison between the proposed method and other state-of-the-art methods.

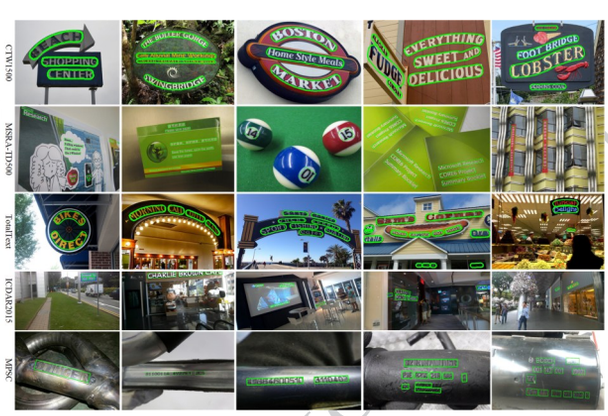


图4 本文模型在不同数据集中的可视化结果

Fig. 4 Visualization results of the proposed model across different datasets.

表7 跨数据集交叉验证

Table 7 Cross-dataset validation results.

训练集	测试集	方法	精确率(%)	召回率(%)	F值(%)
MSRA	CTW	CMNet	77.2	69.7	72.8
		ZTD	84.1	73.4	78.4
		MTD	82.7	72.3	77.2
		本文	83.3	76.0	**79.5**
CTW	MSRA	CMNet	85.8	77.1	81.2
		ZTD	86.8	77.9	82.1
		MTD	88.6	79.7	84.1
		本文	86.3	84.4	**85.3**
IC15	Total	CMNet	75.8	64.5	69.7
		ZTD	78.5	64.1	70.6
		MTD	78.7	74.2	76.4
		本文	82.7	74.4	**78.3**
Total	IC15	CMNet	76.5	68.1	72.1
		ZTD	79.8	69.3	74.2
		MTD	82.0	71.6	76.5
		本文	83.7	74.6	**78.9**

注:黑色字体表示最优结果,下划线表示次优结果。

3.7 跨数据集交叉验证

为进一步评估所提出方法的泛化能力及其对数据分布变化的鲁棒性,本文参考ZTD的实验设置,开展了跨数据集评测。具体而言,分别在行级标注和词级标注两种设定下进行跨数据集实验,以验证模型在不同标注粒度和不同场景分布下的适应能力。

在行级标注场景中,首先在MSRA-TD500数据集上训练模型,并在CTW1500数据集上进行测试。如表7所示,本文方法在精确率、召回率和F值上分别达到83.3%、76.0%和79.5%,优于CM-Net(Yang等,2022)、ZTD(Yang等,2024)和MTD(magnetic text detector)(Han等,2025b)。随后交换训练与测试数据集,即在CTW1500上训练并在MSRA-TD500上测试,所提出方法的F值达到85.3%。该结果不仅超过了CM-Net和ZTD等方法,而且优于部分直接在MSRA-TD500上训练的先进方法,如DBNet(Liao等,2022)和PAN++(Wang等,2022)。进一步地,在词级标注场景下,本文继续开展跨数据集实验。当模型在ICDAR2015数据集上训练并在Total-Text数

© 中国图象图形学报版权所有

（a） 实例截断

（b） 文本褪色

（c） 字符离散化

((a) instance truncation; (b) color fading; (c): character discretization)

图5 本文方法的一些不足之处

Fig. 5 Illustrations of the limitations of the proposed method.

据集上测试时，取得了78.3%的F值；交换训练与测试数据集后，取得了78.9%的F值，整体性能优于现有先进方法MTD（Han 等，2025b）、CM-Net（Yang 等，2022）以及ZTD（Yang 等，2024）。综合上述结果可以得出，所提出方法在训练集与测试集分布存在明显差异的情况下，仍能够保持较为稳定且具有竞争力的检测性能，表明该方法并不依赖于特定数据分布，而具备较好的跨场景泛化能力。进一步分析发现，当模型在以弯曲文本为主的数据集上训练并在多方向文本数据集上测试时，其性能通常优于反向设置。这主要是因为多方向文本数据集对弯曲文本形态的覆盖相对有限，而弯曲文本数据集在文本形态上具有更强的多样性，因此更有利于模型学习到具有迁移性的文本表示。需要指出的是，本文方法的泛化优势并非来源于更复杂的推理结构或更大的模型容量，而是在轻量骨干网络基础上，通过分支内分布建模与跨层语义协同，有效提升了多尺度文本特征的表征质量与融合稳定性。因此，该方法能够在跨数据集评测中展现出较强的鲁棒性与泛化潜力，这也进一步说明其适用于高效场景文本检测任务。

3.8 不足之处分析

图5（a）展示了由于遮挡造成文本实例局部截断，从而被错误识别为两个独立实例的情况，表明当前模型在区分实例内部遮挡区域方面仍存在不足，尚需进一步优化。图5（b）呈现了模型在文本显著褪色文本时的检测结果，反映出在极端外观变化条件下模型判别能力仍受到一定限制。此外，如图5（c）所示，对于空间分布高度离散的文本实例，内核特征受字符间距干扰呈现边界模糊化，难以实现精准定位。这导致在实例扩张后处理阶段，预测区域超出文本轮廓边界，部分背景区域被误判为文本前景。上述失败案例揭示了模型在当前阶段的局限性，也为后续研究中进一步提升模型鲁棒性与判别能力提供了重要改进方向。

4 结 论

本文提出了一种高效场景文本检测方法。该方法通过分支自主分布感知建模，将像素级标注转化为多层次分布监督，引导不同尺度特征学习匹配其感受野的文本分布语义；同时利用跨语义全局知识集成模块进行尺度内增强与跨层次融合，提升复杂文本结构建模能力。实验结果表明，所提出的方法在多个公开场景文本检测基准数据集上均取得了具有竞争力的性能和推理速度，优于现有高效方法。

尽管提出方法取得了一定成果，但在面临实例局部截断、褪色实例以及字符高度分散的实例时，仍存在一定局限性，其主要原因在于上述样本的局部视觉信息不完整、文本与背景区分度较低，或字符间空间连续性较弱。未来工作中，计划将提出的检测模型与高效文本识别模型结合构建端到端的高效文本检测与识别框架。

参考文献（References）

Chen B W, Yi Y H, Tang Z W, Peng J B and Yin A G. 2024. Ship name text detection method with scene priors fusion. Journal of Image and Graphics, 29(10): 3104-3115 (陈博伟，易尧华，汤梓伟，彭继兵，尹爱国. 2024. 融合场景先验的船名文本检测方法.

© 中国图象图形学报版权所有

中国图象图形学报，29(10)：3104-3115)[DOI：10. 11834/jig.230564]

Ch'ng C, and Chan C. 2017. Total-text: A comprehensive dataset for scene text detection and recognition//Proceedings of 2017 14th IAPR International Conference on Document Analysis and Recognition. Kyoto, Japan: IEEE: 935-942 [DOI: 10.1109/ICDAR.2017.157]

Deng D, Liu H F, Li X L, and Cai D. 2018. Pixellink: Detecting scene text via instance segmentation //Proceedings of the AAAI Conference on Artificial Intelligence. [DOI:10.1609/aaai.v32i1.12269]

Ding X, Zhang X, Ma N, Han J, Ding G, and Sun J. 2021. RepVGG: Making VGG-style ConvNets Great Again//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 13728 - 13737[DOI:10.1109/CVPR46437.2021.01352]

Duan C, Fu P, Guo S, Jiang Q, and Wei X. 2024. ODM: A text-image further alignment pre-training approach for scene text detection and spotting//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 15587 - 15597[DOI: 10.1109/CVPR52733.2024.01476]

Guan T, Gu C, Lu C, Tu J, Feng Q, Wu K, and Guan X. 2022. Industrial scene text detection with refined feature-attentive network. IEEE Transactions on Circuits and Systems for Video Technology, 32(9):6073 - 6085[DOI: 10.1109/TCSVT.2022.3156390]

Han X, Gao J Y, Yuan Y, and Wang Q. 2024. Text kernel calculation for arbitrary shape text detection. The Visual Computer, 40(4): 2641 - 2654. [DOI:10.1007/s00371-023-02963-2]

Han X, Gao J Y, Yang C, Yuan Y, and Wang Q. 2025a. Focus entirety and perceive environment for arbitrary-shaped text detection. IEEE Transactions on Multimedia, 27: 287-299. [DOI: 10.1109/TMM.2024.3521797]

Han X, Yang C, and Wang Q. 2025b. Pull pole points to text contour by magnetism: A real-Time scene text detector. IEEE Transactions on Image Processing, 34: 6374-6385. [DOI: 10.1109/TIP. 2025.3609196]

Han X and Wang Q. 2025c. Compensating for the incomplete with the complete: An efficient scene text detector. IEEE Transactions on Circuits and Systems for Video Technology, 35(12):12096-12108. [DOI: 10.1109/TCSVT.2025.3588711]

Han X, Gao J, Yang C, Yuan Y, and Wang Q. 2025d. Spotlight text detector: spotlight on candidate regions like a camera. IEEE Transactions on Multimedia, 27:1937-1949. [DOI:10.1109/TMM.2024.3521824]

Han X, Yang C, Gao J, and Wang Q. 2026. Balancing optimization strategies and practical goals: An efficient scene text detector. IEEE Transactions on Multimedia, 28: 426-438. [DOI: 10.1109/TMM.2025.3623548]

He K M, Zhang X , Ren S Q, and Sun J. 2016. Deep residual learning or image recognition//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 770-778 [DOI: 10.1109/CVPR.2016.90]

He M H, Liao M H, Yang Z B, Zhong H M, Tang J, Cheng W Q, Yao C, Wang Y P, and Bai X. 2021. MOST: A multi-oriented scene text detector with localization refinement//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 8809-8818 [DOI: 10.1109/CVPR46437.2021.00870]

Hu M, Yang Y, Yu H, and Jing B. 2025. ISTD-DLA: Industrial scene text detection method based on dynamic local-aware aggregation network. IEEE Signal Processing Letters, 32:4264 - 4268[DOI: 10.1109/LSP.2025.3627114]

Karatzas D, Gomez-Bigorda L, Nicolaou A, Ghosh S, Bagdanov A, Iwamura M, Matas J, Neumann L, Chandrasekhar V R, Lu S J, Shafait F, Uchida S and Valveny E. 2015. ICDAR 2015 competition on robust reading//Proceedings of the 13th International Conference on Document Analysis and Recognition. Tunis, Tunisia: IEEE:1156-1160[DOI:10.1109/ICDAR.2015.7333942]

Liao M H, Shi B G, and Bai X. 2018. TextBoxes++: A single-shot oriented scene text detector. IEEE Transactions on Image Processing, 27(8):3676-3690. [DOI:10.1109/TIP.2018.2825107]

Liao M H, Shi B G, Bai X, Wang X G, and Liu W Y. 2017. Textboxes: A fast text detector with a single deep neural network//Proceedings of the AAAI Conference on Artificial Intelligence. [DOI: 10.1609/aaai.v31i1.11196]

Liao M H, Wan Z Y, Yao C, Chen K, and Bai X. 2020. Real-time scene text detection with differentiable binarization//Proceedings of the AAAI Conference On Artificial Intelligence AAAI: 11474-11481[DOI:10.1609/aaai.v34i07.6812]

Liang M, Zhu X B, Zhou H Y, Qin J Y, and Yin X C. 2023. HFENet: hybrid feature enhancement network for detecting texts in scenes and traffic panels. IEEE Transactions on Intelligent Transportation Systems, 24 (12) : 14200-14212. [DOI: 10.1109/TITS. 2023.3305686]

Liao M H, Zou Z S, Wan Z Y, Yao C, and Bai X. 2023. Real-time scene text detection with differentiable binarization and adaptive scale fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45 (1) : 919-931 [DOI: 10.1109/TPAMI. 2022.3155612]

Lin T Y, Dollar P, Girshick R, He K M, Harijaran B, and Belongie S. 2017. Feature pyramid networks for object detection//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 936-944 [DOI:10.1109/CVPR.2017.106]

Liu W, Anguelov D, Erhan D, Szegedy C, Read S, Fu C Y, and Berg A C. 2016. SSD: Single shot multibox detector// Proceedings of the European conference on computer vision.[DOI:10.1007/978-3-319-46448-0_2]

Liu Y L, Jin L W, Zhang S T, and Zhang S. 2017. Detecting curve text in the wild: New dataset and new solution. ArXiv[DOI:10.48550/

© 中国图象图形学报版权所有

arXiv.1712.02170]

Liu Y, Chen H, Shen C, He T, Jin L, and Wang L. 2020. Abcnet: Real-time scene text spotting with adaptive bezier-curve network// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, WA, USA: IEEE: 9806-9815 [DOI: 10.1109/CVPR42600.2020.00983]

Long S B, Ruan J Q, Zhang W J, He X, Wu W H and Yao C. 2018. Textsnake: A flexible representation for detecting text of arbitrary shapes// Proceedings of the European conference on computer vision. Springer: 20-36[DOI:10.1007/978-3-030-01216-8_2]

Lyu X Q, Quan W J, Han J, Chen Y Z and Cai Z T. 2024. Text self-training and adversarial learning-relevant domain adaptive industrial scene text detection. Journal of Image and Graphics, 29(10): 3090-3103 (吕学强,权伟杰,韩晶,陈玉忠,才藏太. 2024. 结合文本自训练和对抗学习的领域自适应工业场景文本检测. 中国图象图形学报,29(10):3090-3103)[DOI:10.11834/jig.230519]

Qu Y D, Xie H T, Fang S C, Wang Y X, and Zhang Y D. 2023. ADNet: Rethinking the shrunk polygon-based approach in scene text detection. IEEE Transactions on Multimedia, 25: 6983-6996 [DOI:10.1109/TMM.2022.321672]

Sheng T, Chen J, and Lian Z. 2021. CentripetalText: An efficient text instance representation for scene text detection// Proceedings of the International Conference on Neural Information Processing Systems, 34:335-346.

Shi B G, Bai X, and Belongie S. 2017. Detecting oriented text in natural images by linking segments//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE: 3482-3490 [DOI:10.1109/CVPR.2017.371]

Shi G C and Wu Y R. 2021. Arbitrary shape scene-text detection based on pixel aggregation and feature enhancement. Journal of Image and Graphics,26(07):1614-1624 (师广琛,巫义锐. 2021. 像素聚合和特征增强的任意形状场景文本检测. 中国图象图形学报,26(07):1614-1624)[DOI:10.11834/jig.200522]

Su Y C, Chen Z N, Du Y N, Ji Z L, Hu K, Bai J F, and Gao X P. 2025. Explicit relational reasoning network for scene text detection// Proceedings of the AAAI Conference on Artificial Intelligence. AAAI: 7069-7077 [DOI:10.1609/aaai.v39i7.32759]

Su Y C, Chen Z N, Shao Z W, Du Y N, Ji Z L, Bai J F, Zhou Y, and Jiang Y G. 2024. LRANet: Towards accurate and efficient scene text detection with low-rank approximation network//Proceedings of the AAAI Conference on Artificial Intelligence. AAAI: 4979-4987 [DOI:10.1609/aaai.v38i5.28302]

Su Y C, Shao Z W, Zhou Y, Meng F R, Zhu H C, Liu B, and Yao R. 2023. TextDCT: Arbitrary-shaped text detection via discrete cosine transform mask. IEEE Transactions on Multimedia, 25:5030-5042. [DOI:10.1109/TMM.2022.3186431]

Vatti B R. 1992. A generic solution to polygon clipping. Communications of the ACM, 35(7):56-63. [DOI:10.1145/129902.129906]

Wang F F, Chen Y F, Wu F, and Li X. 2020. Textray: Contour-based geometric modeling for arbitrary-shaped scene text detection// Proceedings of the ACM International Conference on Multimedia. ACM: 111-119 [DOI:10.1145/3394171.3413819]

Wang F F, Xu X G, Chen Y F, and Li X. 2023. Fuzzy semantics for arbitrary-shaped scene text detection. IEEE Transactions on Image Processing, 32:1-12. [DOI:10.1109/TIP.2022.3201467]

Wang W H, Xie E Z, Li X, Hou W B, Lu T, Yu G, and Shao S. 2019a. Shape robust text detection with progressive scale expansion network//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Neach, USA: IEEE: 9328-9337 [DOI:10.1109/CVPR.2019.00956]

Wang W H, Xie E Z, Song X G, Zang Y H, Wang W J, Lu T, Yu G, and Shen C H. 2019b. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network//Proceedings of IEEE/CVF Conference on International Conference on Computer Vision. Seoul, Korea (South): IEEE: 8439-8448 [DOI: 10.1109/ICCV.2019.00853]

Wang Z X, Xie H T, Wang Y X and Zhang Y D. 2023. Hierarchical semantics-fused scene text detection. Journal of Image and Graphics, 28(08):2343-2355 (王紫霄,谢洪涛,王裕鑫,张勇东. 2023. 层级语义融合的场景文本检测. 中国图象图形学报,28(08):2343-2355)[DOI:10.11834/jig.220902]

Yang C, Chen M L, Xiong Z T, Yuan Y, and Wang Q. 2022. CM-Net: Concentric mask based arbitrary-shaped text detection. IEEE Transactions On Image Processing, 31:2864-2877. [DOI:10.1109/TIP.2022.3141844]

Yang C, Chen M L, Yuan Y, and Wang Q. 2023a. Text growing on leaf. IEEE Transactions on Multimedia, 25:9029-9043. [DOI:10.1109/TMM.2023.3244322]

Yang C, Chen M L, Yuan Y, and Wang Q. 2024. Zoom Text Detector. IEEE Transactions on Neural Networks and Learning Systems, 35(11):15745-15757. [DOI:10.1109/TNNLS.2023.3289327]

Yang C, Chen M L, Yuan Y, and Wang Q. 2023b. Reinforcement shrink-mask for text detection. IEEE Transactions on Multimedia, 25:6458-6470. [DOI:10.1109/TMM.2022.320902]

Yang C, Han X, Han T, Han H, Zhao B X, and Wang Q. 2025. Edge approximation text detector. IEEE Transactions on Circuits and Systems for Video Technology. [DOI:10.1109/TCSVT.2025.3558634]

Yang Y, Hu M, Yu J, and Jing B. 2025. RITD: Real-time industrial text detection with boundary- and pixel-aware modules. Displays, 87:102973[DOI:10.1016/j.displa.2025.102973]

Yao C, Bai X, and Liu W Y. 2014. A Unified Framework for Multioriented Text Detection and Recognition. IEEE Transactions on Image Processing, 23 (11): 4737-4749. [DOI: 10.1109/TIP. 2014.2353813]

Yao C, Bai X, Liu W Y, Ma Y, and Tu Z W. 2012. Detecting texts of arbitrary orientations in natural images//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE: 1083-1090 [DOI: 10.1109/CVPR.2012.

© 中国图象图形学报版权所有

6247787]

Zhang S, Zhu X B, Yang C, Wang H F, and Yin X. 2021. Adaptive boundary proposal network for arbitrary shape text detection//Proceedings of IEEE/CVF Conference on International Conference on Computer Vision. Montreal, QC, Canada: IEEE: 1285-1294 [DOI:10.1109/ICCV48922.2021.00134]

Zhang X S, Zhu X B, Hou J B, Yang C, and Yin X C. 2023. Kernel proposal network for arbitrary shape text detection. IEEE Transactions on Neural Networks and Learning Systems, 34(11): 8731-8742. [DOI:10.1109/TNNLS.2022.3152596]

Zhao Y, Lv W, Xu S, Wei J, Wang G, Dang Q, Liu Y, and Chen J. 2024. DETRs beat YOLOs on real-time object detection//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 16965 - 16974 [DOI:10.1109/CVPR52733.2024.01605]

Zhu X Z, Hu H, Lin S and Dai J F. 2019. Deformable convnets v2: More deformable, better results//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE: 9300-9308 [10.1109/CVPR.2019.00953]

Zhu Y Q, Chen J Y, Liang L Y, Kuang Z H, Jin L W, and Zhang W. 2021. Fourier contour embedding for arbitrary-shaped text detection//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, TN, USA: IEEE: 3122-3130 [DOI:10.1109/CVPR46437.2021.00314]

© 中国图象图形学报版权所有