=5

智能视觉系统在工业自动化中起着至关重要的作用，特别是在工业面板监测领域。工业面板包括示波器、温度控制器、老化测试平台等，可以提供实时数据。然而，有限的数据获取限制了有效管理，使后续验证变得复杂。目前，工业面板数据监测主要依赖人工定期观察和记录。这种方法不仅无法捕获产品制造和测试的完整生命周期数据，还由于需要高强度的人工劳动，导致关键数据丢失和性能评估不准确。此外，随着设备的快速发展和电子行业的持续增长，传统的人工监测已经无法满足现代工业生产对于效率、质量和可靠性日益增长的需求。持续（24/7）监测的需求增加，突出显示了对面板设备智能监测解决方案的需求。因此，设计一种具有高精度、高效率和高鲁棒性的智能视觉系统用于工业面板监测任务已成为迫切的优先事项。基于深度学习的文本检测和识别技术已作为工业自动化监测的创新解决方案出现。由于其在各个领域的成功应用，基于级联的文本识别方法已被广泛验证。它们遵循先检测后识别的策略，如图所示。然而，级联结构固有地导致从检测网络到识别网络的性能下降。此外，在这种范式下，文本检测和识别之间缺乏协同作用。这些限制阻碍了文本检测技术在复杂工业监控场景中的更广泛应用，例如涉及密集文本、多尺度变化和强光照条件的场景。近年来，随着 Transformer 在计算机视觉中的广泛应用，研究人员越来越多地将检测和识别网络集成到 Transformer 框架中以增强特征交互。这些基于 Transformer 的端到端（E2E）文本检测方法因其能够利用文本检测和识别之间的协同作用而受到广泛关注。然而，尽管它们在一般场景中表现令人满意，这些方法常常忽视了由于文本尺度变化、密集区域和复杂的工业面板监控条件所带来的独特挑战。此外，与基于级联的方法相比，专门针对工业应用的研究仍然有限。此外，这些方法的实际有效性尚未在真实工业环境中得到彻底验证，其较高的计算需求为边缘部署带来了挑战。因此，开发一种轻量级且有效、鲁棒的基于 Transformer 的文本检测器，旨在解决工业面板文本检测的特定挑战，变得至关重要。由于面板中文本对象在尺度上的显著差异，基于单层特征的 Transformer 难以有效表示多尺度信息。为了解决这个问题，这项工作设计了一种高效的多级特征混合器，以学习不同层级特征之间的相互依赖关系，从而适应不同尺度的文本。此外，为了解决在密集文本区域中不准确识别的问题，本文引入了基于 Catmull-rom 样条的特征采样方法，该方法明确编码了文本的形状、位置和语义信息。总体而言，提出了一种新颖的多尺度密集文本识别器（EdgeSpotter），用于边缘 AI 视觉系统，实现精确和稳健的工业面板监测。本文的主要贡献如下：

- 在自建的边缘 AI 系统中设计和实现了一种基于边缘智能的工业面板监控检测框架，与繁重、劳动密集且短期的人工监控相比，该框架能够实现面板信息的连续（24/7）监控。
- 提出了一种高效的多级特征混合器，以学习不同特征之间的相互依赖性和空间信息，从而提高任意尺度上的文本检测性能。
- 引入基于 Catmull-Rom 样条的特征采样，以减轻由密集文本区域导致的边界模糊和识别错误。
- 一个专为工业面板监控情景（IPM）量身打造的文本数据集被精心构建。在该数据集上的大量实验展示了 EdgeSpotter 的卓越性能和鲁棒性。此外，实际部署应用进一步验证了其在实际场景中的有效性。

过去，文本检测通常被分为两个不同的子任务：检测和识别，每个任务独立研究。然而，这些方法表现出一些局限性，例如密集字符的低效推断、误差累积和次优性能。为了解决这些问题，文本检测方法逐渐从两阶段的浅合并到检测器中，隐式地引导识别头。虽然这些方法消除了对启发式后处理的需求，但仍面临与任务协作和训练效率相关的挑战。为了解决这些问题，DeepSolo 和 DNtextSpotter 显式地建模了一组可学习的点序列，从单个解码器中派生文本的中心线、边界、实例和置信度。此外，现有的大多数工作主要集中在自然场景中的文本定位，对处理复杂的文本信息缺乏鲁棒性，特别是在工业面板中存在显著尺度变化的情况下。此外，这些方法的巨大计算需求导致实时性能问题，这对于工业应用至关重要。为了解决这些挑战，这项工作设计了一种新颖的高效混合 Transformer，并通过尺度自适应向量实现多级特征混合，有效地聚合多尺度文本信息以实现精确和稳健的面板文本定位。最初，在基于级联的方法中，特征采样作为连接检测器和识别器的重要机制，为识别操作提供来自潜在文本区域的原始输入。RoI Pooling 首次被引入用于特征采样，并且自那时以来被广泛采用。RoIAlign 采用双线性插值进行加权特征采样，并首次被扩展用于采样非轴对齐（即旋转）的 RoIs。值得注意的是，ABCNet 引入贝塞尔参数化以定位曲线文本，通过贝塞尔对齐算子进行特征采样，有效解决了检测任意形状文本的问题。更复杂地，GLASS 集成了直接从标准化词块计算的附加信息，使特征采样能够从全局到局部尺度。随着后续端到端 Transformer 方法的出现，特征采样通常出现在编码器和解码器之间，以提取文本特征。例如，TESTR 通过指导生成器获取原始控制点序列，然后利用从框到多边形策略获取后续解码器的参考特征。尽管它使用带有边界框位置的点查询，但检测和识别的查询是不同的。同样，受 ABCNet 启发，DeepSolo 使用中心贝塞尔曲线采样来明确编码文本特征。然而，上述方法中的控制点与采样曲线高度间接，使其不太适合具有相对规则形状的文本，尤其是在处理面板文字的情况下。

EdgeSpotter 的工作流程如图 **??** 所示。它可以分为四个模块：骨干网络、带有效混合器 Transformer 的编码器、带 Catmull-Rom 样条的特征采样、Transformer 解码器。为了确保与其他 SOTA 方法一致，本研究采用 ResNet50 作为特征提取的骨干网络。

为了充分探索多层特征之间的相互依赖性，EdgeSpotter 利用了主干的最后三个阶段的输出特征。具体来说，最后的三个特征图被输入到高效混合 Transformer（EMT）中。通过高效混合器（EM），实现了尺度内交互和跨尺度融合，从而获得具有丰富信息的多尺度特征。简单的网络结构保证了操作效率。为清晰起见，以下介绍中最后三个输出特征图统一用 $\mathcal{F}_l \in \mathbb{R}^{W\times H\times C}$ 表示（C 、W 和 H 分别表示特征图的通道、宽度和高度，且 $l \in \{3,4,5\}$）。然后，$\mathcal{F}_5$ 经过一个步长为 2 的 3×3 卷积以获得 $\mathcal{F}_6$。最后，将 $\mathcal{F}_3$ 、$\mathcal{F}_4$ 、$\mathcal{F}_5$ 和 $\mathcal{F}_6$ 重塑并拼接以获得最终特征 $\mathcal{F} \in \mathbb{R}^{N\times C}$ ，其中 N 是 token 长度。然后，EMT 可以表达为：

$$\mathcal{X} = \mathrm{EM}(\mathrm{LN}(\mathcal{F})) + \mathcal{F} \quad , \tag{1}$$

$$\mathrm{EMT}(\mathcal{F}) = \mathrm{MLP}(\mathrm{LN}(\mathcal{X})) + \mathcal{X} \quad , \tag{2}$$

，其中 $\mathrm{LN}(\cdot)$ 是层归一化，$\mathrm{MLP}(\cdot)$ 是多层感知，$\mathrm{EM}(\cdot)$ 表示高效混合器。

备注 1：EMT 强调多级特征之间的交互，促进构建层间依赖关系，并实现多级特征互补。随后，使用可变形 Transformer 应用了多级编码器。

受 [?] 的启发，EM 避免了昂贵的矩阵乘法操作，并将计算复杂度从平方降低到线性。同时，引入了一个可学习的参数向量 $\mathcal{W}_m \in \mathbb{R}^C$ 来表示多层次注意力权重。由于获得的特征 $\mathcal{F}$ 包含丰富的多层次特征，使用 $\mathcal{W}_m$ 来重新塑造每层中特定类型的文本信息（如不同的尺度）。这不仅避免了不同层次之间交互造成的信息混淆，还便于动态探索多层次特征信息。具体来说，输入嵌入矩阵 X_n 通过两个矩阵 W_k , W_v （其中 K 等于 Q ）被转换为 Q , K 和 $V \in \mathbb{R}^{N\times C}$ 。接下来，矩阵 K 与 $\mathcal{W}_m$ 相乘以学习查询的注意力权重，产生全局注意查询向量 $\mathcal{W}_{attn} \in \mathbb{R}^N$ 。随后，矩阵 V 与广播的 $\mathcal{W}_{attn}$ 进行元素级相乘以产生全局上下文表示。该操作将全局信息整合到矩阵的每个元素中，从而增强模型对多尺度特征的敏感性。在经过一个线性层以后，所得到的多尺度注意力矩阵 $G \in \mathbb{R}^{N\times C}$ 与 Q 进行元素级相加。如图 **??** 所示，EM 可以总结为：

$$\mathcal{W}_{attn} = K\mathcal{W}_m \quad , \tag{3}$$

$$\mathrm{EM}(X_n) = \varphi(\varphi(V \star \mathcal{W}_{attn}/\sqrt{\mathrm{D}}) \oplus Q) \quad , \tag{4}$$

其中 X_n 是输入特征，$\star$ 表示点积，$\varphi(\cdot)$ 表示线性层。所提出的矩阵操作捕捉了来自每个标记的信息，并学习了输入序列中的相关性。

注释 2：EM 使用元素级乘法来实现线性复杂度，这大大减少了计算成本。此外，$\mathcal{W}_m$ 自适应地学习多层次特征的注意力权重，因此该模型在处理任意规模文本时具有很强的鲁棒性。

常见的特征采样方法，如框选择建议和 Bezier 曲线建议，要么在信息表示上存在限制，要么由于曲线形状与其控制点之间的间接关系而表现出平滑度差。相比之下，本文提出了一种简单的基于 catmull-rom 样条的方案，称为 catmull-rom 样条特征采样（FSCRS），它是从文本中心线的角度设计的。FSCRS 能够有效地拟合面板文本并区分不同的文本实例。然后，这种特征采样方法用于指导后续的特征解码，有效缓解因密集面板文本引起的边界模糊。给定特征提取网络得到的图像特征，在特征的每个像素上，使用一个 MLP 来预测四个 catmull-rom 控制点的偏移，从而确定表示一个文本实例的曲线。令 i 索引特征中的一个像素，其二维归一化坐标为 $\hat{p}_i = (\hat{p}_{ix}, \hat{p}_{iy}) \in [0,1]^2$ 。对于每个像素 i ，预测相应的 catmull-rom 控制点 $CRP_i = \{crp_{i_0}, crp_{i_1}, crp_{i_2}, crp_{i_3}\}$ 。这些控制点的坐标使用 sigmoid 函数计算如下：控制点中得分最高的 K 个被选择为建议，基于线性层的评分结果。所描述的初始控制点查询记为 $\{P_c^{(k)}\}_{k=1}^K$ 。然后，对每条曲线上的 n 个点进行均匀采样 [?] ，最终得到点坐标为 $P_s \in \mathbb{R}^{K\times n\times 2}$ 。基于 P_s ，我们使用一个简单的 MLP 进一步投影以获得位置查询 $P_q \in \mathbb{R}^{K\times n\times C}$ 。P_q 的计算可表示为：其中 CatRom $(\cdot)$ 表示 catmull-rom 样条采样，PE$(\cdot)$ 表示正弦位置编码函数。CatRom $(\cdot)$ 可表示为：

$$\mathrm{CatRom}(P_c^{(k)}) = U(u)M(\tau)P_c^{(k)} \quad , \tag{5}$$

，其中 $u \in [0,1]$ 表示曲线参数，τ 是用于控制曲线平滑度的张力参数（$\tau = 0.5$ 是通过经验设置的）。矩阵 M 的系数决定了 catmull-rom 样条的形状。即使在具有挑战性的面板文本监控场景中，例如密集和多尺度文本，FSCRS 仍然可以准确定位所有文本实例并提取关键特征，如图 1 所示。

Fig. 1. Density maps of control points for Top-K scores. Text instances with lower scores tend to have a greater number of control points. Zoom in for better visualization.

注释 3：通过分析 Top-K 控制点的密度分布，观察到得分较低的文本实例往往具有更多的控制点，如图 1 所示。这表明 FSCRS 可以根据不同文本实例的得分动态调整控制点的数量，从而整体提升性能。

A. *Transformer* 解码器

如图 **??** 所示，FSCRS 获取的特征查询被输入到后续解码器中作为参考点，EMT 编码器提取的多尺度特征则作为解码器的输入，形成基础序列。然后，遵循 [?] ，该工作采用四个简单的预测头分别预测实例类、字符类、中心曲线点和边界框。

注释 4：通过 EMT，各级特征的空间和语义信息被充分利用，增强了处理大规模变化时的鲁棒性。同时，FSCRS 被用来从这些多级特征中提取关键信息，这大大有利于最终的特征解码。仅需四个并行的简单预测头，EdgeSpotter 即可取得令人满意的结果，如图 2 所示。

I. 实验

A. 数据集

这项工作在一个新的基准数据集 IPM 上评估我们方法的性能，该数据集专门为工业面板监测设计，包含 2,005 张图像。其中，1,200 张图像用于训练，300 张图像用于验证，剩余的图像根据各种挑战属性进行分类。这些属性包括“侧视”、“强光”、“反射”、“阴影”、“特殊字符”、“多尺度”和“密集字符”。此外，本工作使用以下数据集进行预训练：1）Synth150K [?] ，这是一个包含 94,723 张多方向文本图像和 54,327 张弯曲文本图像的合成数据集；2）IC15 [?] ，其中包括用于四边形场景文本的 1,000 张训练图像和 500 张测试图像；3）CTW1500 [?] ，这是一个用于具有任意形状场景文本的文本行级基准，包含 1,000 张训练图像和 500 张测试图像。

备注 5：为了确保更好的泛化能力和高效的收敛，预训练在多样化的数据集上是必不可少的。然后使用 IPM 进行微调，涵盖各种工业面板类型，并专门为工业面板监测设计。该数据集包含 31,595 个文本目标，平均每张图像 16 个目标。

B. 实现细节和评估指标

在我们的模型中，提案的数量 K 设置为 100，而采样点的数量 n 设置为 25。EdgeSpotter 在 IPM 上预测 96 类。AdamW 优化器用于训练。模型训练和评估在 NVIDIA

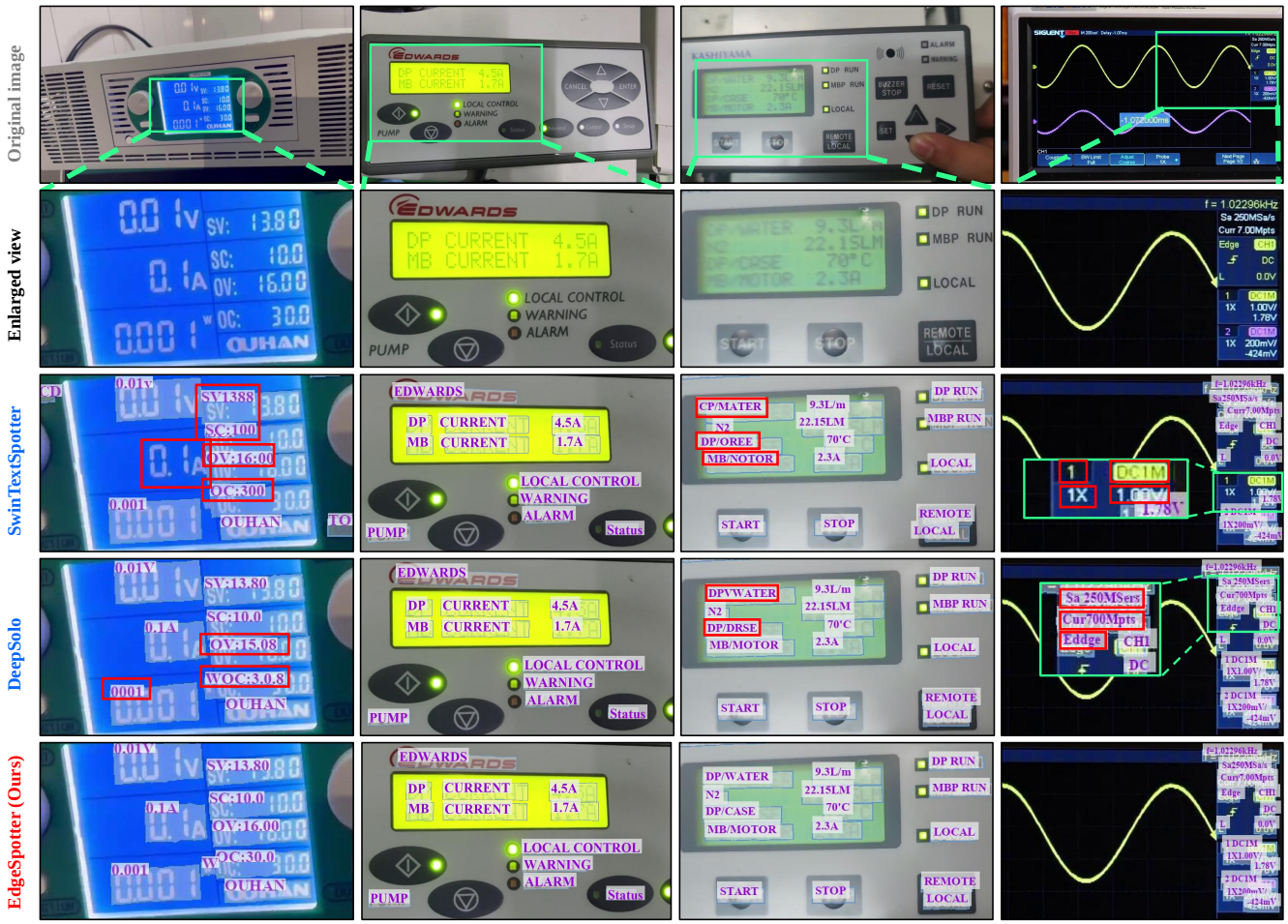

Fig. 2. Comparison of EdgeSpotter spotting results with other SOTA spotting results. Areas with incorrect identification (e.g., incorrect identification, missed detection) are marked with red boxes. The original image and the enlarged view can be found in the second and first rows, respectively.

TITAN RTX GPU 上进行。为了公平起见，所有模型的训练次数均为 12k 次迭代（基于级联的方法可能需要训练更多次以确保结果）。评估协议使用了 ICDAR 鲁棒阅读竞赛中使用的标准来评估检测性能。具体来说，如果一个边界框与任何地面实况的 IoU 超过 0.5，并且识别的单词也匹配，则视为正确。在识别中，本文遵循“端到端”评估协议，这要求图像中的所有单词都要被识别出来，无论字符串是否存在于提供的上下文词库中。

为了全面评估所提出方法的有效性，本研究在 IPM 验证集上将我们的模型与八种 SOTA 文本检测方法进行了比较，包括 ABCNet 系列、Mask TextSpotter v3、PGNet、PAN++、SwinTextSpotter、DeepSolo 和 DNtextSpotter。为保证公平性，所有方法在可能情况下都使用 ResNet50 作为基础网络。如表所示，EdgeSpotter 在整体表现上优

TABLE I

基准测试上的定量识别结果。所有评估均在一台 NVIDIA TITAN RTX 上进行。红色代表最佳结果。

Method	Years	Backbone	Detection			Recognition		
			P(%)	R(%)	F1(%)	P(%)	R(%)	H(%)
ABCNet [?]	2020	ResNet50	97.29	90.39	93.71	74.43	69.16	71.70
Mask TexSpotter v3 [?]	2020	ResNet50	92.54	89.67	91.08	73.26	69.34	71.25
PGNet [?]	2021	ResNet50	68.00	63.38	65.61	43.60	35.74	39.25
ABCNetv2 [?]	2021	ResNet50	95.02	91.90	93.43	73.09	71.22	72.14
PAN++ [?]	2021	ResNet18	93.26	88.15	90.63	73.15	70.69	71.89
SwinTextSpotter [?]	2022	Swin-T	96.11	90.57	93.26	80.31	75.69	77.93
DeepSolo [?]	2023	ResNet50	97.67	93.78	95.68	78.60	75.47	77.00
DNtextSpotter [?]	2024	ResNet50	93.21	94.23	93.72	78.12	76.35	77.22
EdgeSpotter (Ours)	-	ResNet50	98.16	94.52	96.31	82.53	80.02	81.25

于其他 SOTA 文本检测方法。具体而言，EdgeSpotter 在 F1 和 H 指标上均取得最高分，分别比第二好的方法高出 0.63 和 3.32。图中展示了一些可视化示例。在综合评价中的优异表现表明，EdgeSpotter 是工业面板监控的最佳选择。

为了在各种挑战下对 EdgeSpotter 进行彻底评估，进行了基于属性的比较，如图所示。基于方法学的差异和表中呈现的全面表现数据，本节比较了五种方法。总体而言，EdgeSpotter 在检测和识别表现上均排名第一。具体来说，在多尺度、密集文本、强光和侧视等属性上，它显著优于第二好的方法。其中，检测密集文本的显著改进最为明显，而在识别方面，多尺度文本的显著提升最为突出。令人满意的结果表明，高效的多层次特征混合策略和基于 Catmull-Rom 样条的特征采样方法能够有效提升工业面板文本检测的表现，尤其是在复杂场景中，包括多尺度和密集场景。在本节中，通过在 IPM 上进行的实验分析每个模块的贡献。基础模型定义为仅由特征提取、Transformer 编码和 Transformer 解码组成的网络。如表格 **??** 所示，基础网络的 F1 得分为 90.06 % ，识别精度（H）为 74.34 % ，低于本研究中比较的所有基于 E2E Transformer 的方法。在引入 FSCRS 后，F1 提高到 95.49 % ，提高了 5.43 % ，而 H 上升到 79.04 % ，其性能与其他 SOTA 方法相当甚至更佳。这表明 FSCRS 显著增强了特征提取能力，提升了工业面板中的文本检测性能。此外，仅引入 EMT 就导致 F1 提高了 4.77 % ，H 增加了 2.64 % 。值得注意的是，EMT 在文本定位上的改善比识别上的改善更为显著，突出其在促进跨多层次特征的空间信息深度交互方面的有效性，从而进一步提高了任意尺度文本检测的性能。如图 **??** 所示，为了证明 EdgeSpotter 在实际工业面板监测中的适用性，我们在配备 Intel i9 CPU 和 NVIDIA RTX 3070 GPU 的自建智能视觉系统上进行了广泛的测试。在实际测试中，EdgeSpotter 在保持准确性的同时，达到了超过 25 FPS 的点位速度。为了进一步评估点位质量，计算了每个测试组的平均得分。TEST 1 表明了所提出的方法在多尺度文本场景中有效检测高质量结果。TEST 2 侧重于评估在密集文本和强光干扰中的性能。TEST 3 提出了一个更具挑战性的监控场景。尽管存在这些挑战，但得益于其强大的多尺度和密集文本特征提取能力，EdgeSpotter 取得了出色的点位结果。总之，EdgeSpotter 能够在复杂条件下可靠地识别工业面板中的文本信息，使其成为监控真实工业面板的宝贵工具。

本工作提出了 EdgeSpotter，这是一种新型文本识别器，部署在轻量、便携、低功耗的视觉系统中，用于智能工业面板监控。其目标是降低人工成本，实现工业生产的全周期监控。为解决尺度变化问题并确保效率，EdgeSpotter 引入了 EMT 来捕捉跨尺度的信息并建立层间依赖关系。FSCRS 用于编码密集文本的信息。实验结果表明，EdgeSpotter 在复杂工业面板上有效地进行了文本识别。总之，我们相信这项工作将大大推动工业自动化的发展。

本工作得到了国家自然科学基金项目（U24B20161）的支持。