Computer Science and Application 计算机科学与应用, 2025, 15(2), 169-178
Published Online February 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.152044

Hans汉斯

语言驱动的语义边缘检测

余　斌[1]，张笑钦[2*]，邓若曦[1]

[1]温州大学计算机与人工智能学院，浙江　温州
[2]浙江工业大学计算机科学与技术学院、软件学院，浙江　杭州

收稿日期：2025年1月20日；录用日期：2025年2月20日；发布日期：2025年2月28日

摘　要

语义边缘检测致力于精确描绘对象边界并为各个像素分配类别标签，这对实现准确定位和分类提出了双重挑战。本研究介绍了语言驱动语义边缘检测，这是一个简单的框架，可增强语义轮廓检测模型。语言驱动语义边缘检测旨在利用嵌入在文本表示中的语义信息来重新校准边缘检测器的注意力，从而增强高级图像特征的判别能力。为了实现这一点，我们引入了文本特征信息，使用跨模态融合方式增强了边缘检测器的定位和分类。在SBD和CityScapes数据集上的实验结果表明，模型性能得到显著提升。例如，在CASENet中加入文本特征信息可将SBD数据集上的平均ODS得分从70.4提高到72.6。最终，语言驱动语义边缘检测实现了领先的平均ODS 77.0，超越了竞争对手。我们将展示更多额外的结合方法、主干网络的效果。

关键词

语义边缘检测，跨模态融合，卷积神经网络，CLIP

Language-Driven Semantic Edge Detection

Bin Yu[1], Xiaoqin Zhang[2*], Ruoxi Deng[1]

[1]School of Computing and Artificial Intelligence, Wenzhou University, Wenzhou Zhejiang
[2]School of Computer Science and Technology, School of Software, Zhejiang University of Technology, Hangzhou Zhejiang

Received: Jan. 20th, 2025; accepted: Feb. 20th, 2025; published: Feb. 28th, 2025

Abstract

Semantic edge detection strives to accurately delineate object boundaries and assign category labels to individual pixels, which poses a dual challenge to achieve accurate localization and classification. This study introduces language-driven semantic edge detection, a simple framework that

*通讯作者。

文章引用: 余斌, 张笑钦, 邓若曦. 语言驱动的语义边缘检测[J]. 计算机科学与应用, 2025, 15(2): 169-178.
DOI: 10.12677/csa.2025.152044

enhances semantic contour detection models. Language-driven semantic edge detection aims to leverage the semantic information embedded in text representations to recalibrate the attention of edge detectors, thereby enhancing the discriminative ability of high-level image features. To achieve this, we introduce text feature information and use cross-modal fusion to enhance the localization and classification of edge detectors. Experimental results on SBD and CityScapes datasets show that model performance is significantly improved. For example, adding text feature information to CASENet improves the average ODS score on the SBD dataset from 70.4 to 72.6. Ultimately, language-driven semantic edge detection achieves a leading average ODS of 77.0, surpassing the competition. We will show the effects of more additional combining methods and backbone networks.

Keywords

Semantic Edge Detection, Cross-Modal Fusion, Convolutional Neural Network, CLIP

1. 引言

经典的轮廓检测不依赖于物体类别，通常使用边缘检测算法，如 Canny 边缘检测、Sobel 算子等，来识别图像中的边缘。相比之下，语义边缘检测则更进一步，它不仅识别边缘，还为这些边缘分配语义标签。这意味着它能够理解哪些边缘对应于特定物体的形状或特征，从而实现更高级别的图像理解。然而，由于图像边缘和背景像素分布不均衡，导致语义边缘检测优化困难。先前的语义边缘检测工作[1]-[3]主要建立在 CASENet [4]的基础上，利用上下文信息和多尺度特征，提高了语义边缘检测的性能。尽管这些方法取得了良好的准确率，但它们受到主干网络能力的限制，同时，低质量的语义边缘标签[5]则进一步限制了它们的性能。

为此，我们对和语义边缘检测相似的语义分割任务进行了积极探索，其中，LSeg [6]脱颖而出。它通过结合 CLIP [7]的文本编码器和基于 Transformer 的图像编码器来整合语言信息，从而增强了语义分割能力。LSeg 方法的核心是计算文本和图像嵌入之间的相关性，通过有效地融合文本和图像特征来生成多类别的掩码。这种计算策略对于融合跨模态语义信息并因此增强模型的分类能力至关重要。

在此基础上，我们提出了一种创新且有效的方法，称为语言驱动语义边缘检测。我们的研究引入了一种跨模态特征融合方法，旨在从文本嵌入中提取语义信息并利用它来指导语义边缘检测。为了实现这一点，我们使用了 CLIP 中的文本编码器，而图像编码器可以是任意类型的神经网络。我们的方法使边缘检测器能够从文本嵌入中学习语义知识并利用它来重新校准注意力，从而提高性能。总之，我们的工作做出了以下贡献：

1、我们利用语言信息来解决语义边缘检测中优化困难问题。据我们所知，这是利用文本特征来增强边缘检测器性能的初步尝试。

2、我们引入了一种新颖的跨模态特征融合方法，该方法既灵活又有效地整合了语义信息。这种方法重新校准了边缘检测器的注意力，从而进一步增强了模型的性能。

3、我们的语义边缘检测方法在 SBD 和 CityScapes 数据集上取得了显著效果，超越了 DDS 和 STEAL [1]等检测器。同时，我们展示了跨模态特征融合模块在实验中显著提高边缘检测器准确性的能力。

2. 相关工作

2.1. 边缘检测

边缘检测已发展了四十年，最初是依靠图像梯度的 Sobel 和 Canny 检测器[8]-[11]。后来出现了基于学习的方法，使用低级特征来获取对象级轮廓，并应用于图像分割等任务[12]-[14]。然而，这些方法通常依赖于手工制作的特征，限制了它们的增强潜力。

近年来，最先进的边缘检测器主要利用了深度卷积神经网络[15]-[18]。值得注意的例子包括 HED、RCF [19]、LPCB [20]和 DSCD [21]。这些检测器可分为两类。第一类[22]专注于研究模型结构以提高提取图像特征的能力，从而提高整体模型性能。第二类还\\探索了损失函数以解决不平衡分布问题。该领域仍然很活跃，并不断引入新方法来解决该任务。

2.2. 语义边缘检测

CASENet 是一种专注于语义边缘检测的深度学习框架。利用上下文信息来增强边缘特征的表达能力，使得模型在检测边缘时能够考虑周围环境，提升检测的准确性。该模型还通过融合不同尺度的特征，能够更有效地捕捉到图像中各种物体的边缘。这种多尺度处理能够适应不同尺寸和形状的物体。采用多层次的网络结构，逐层提取和组合特征，以实现细粒度的边缘检测。这种层次化的特征提取能够帮助模型捕获更丰富的信息。

Seal 提出了同时进行边缘对齐和学习的框架来应对语义边缘检测的优化困难问题。通过制定一个概率模型，将边缘对齐视为潜在变量优化，并在网络训练期间进行端到端学习。

Steal 在训练期间推理注释噪声来学习清晰而精确的语义边界。提出了一个简单的新层和损失，在训练期间使用公式推理真实的对象边界，使网络能够以端到端的方式从未对齐的标签中学习。学习到的网络可用于显著改善粗分割标签，成为标记新数据的有效方法。

DDS 提出了一种全卷积神经网络，它在多任务框架内使用多样化的深度监督，其中底层旨在生成与类别无关的边缘，而顶层负责检测类别感知的语义边缘。并且，引入了一种新颖的信息转换单元，提高了定位精度，为不准确的边缘注释提供了全面的解决方案。

与其他方法不同，我们的方法从语言描述中提取语义信息，以增强图像编码器的语义理解，从而提高整体模型性能。下文将详细说明。

3. 方法

我们方法的核心概念是提高模型提取优质语义信息的能力，生成更好的注意力特征并增强图像特征的整体判别能力。与以前的语言驱动分割方法如 LSeg 和 CLIPSeg 不同，这些方法主要使用文本分支来实现零样本或一次性能力，而我们的方法则利用 CLIP 文本编码器生成的文本嵌入中的潜在语义信息。在本研究中，我们旨在证明隐藏在文本嵌入中的语义信息可以提高图像特征的质量。这种增强随后会微调边缘检测器的注意力，最终提高模型性能。我们的模型整体结构如图 1 所示。

3.1. 模型结构

所提出的方法整合了视觉和文本模态信息，以增强语义边缘检测性能。该模型采用图像 $I \in \mathbb{R}^{H\times W\times 3}$ 作为输入，并附有相应的文本描述 $\{T_i \mid i = 0,1,..N-1\}$，其中 N 表示数据集中的标签数量。模型的输出是一个掩码 $P \in \mathbb{R}^{H\times W\times N}$，用总体框架方程来表达：

$$P = Model(I,T) \tag{3-1}$$

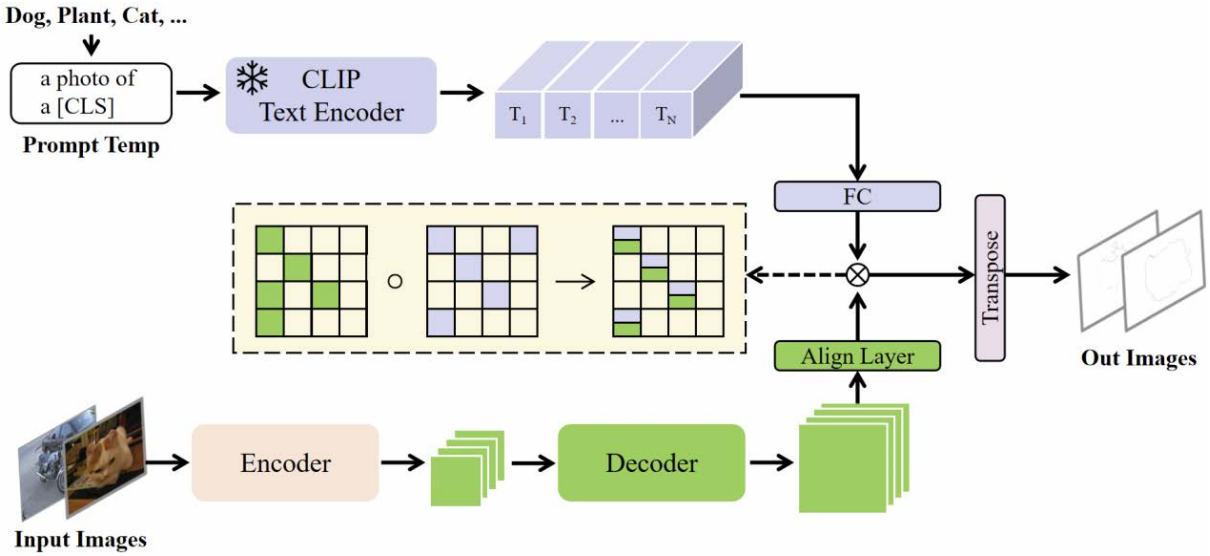

Figure 1. Overall structure of language-driven semantic edge detection method
图 1. 语言驱动的语义边缘检测方法整体结构

边缘检测器分解为编码器 E 和解码器 D。在我们的具体框架中，编码器包括两个组件：图像编码器 E_I 和一个文本编码器 E_T。因此，总体公式(3-1)可以写成：

$$P = F\left(D\left(E_I\left(I\right)\right), E_T\left(T\right)\right) \tag{3-2}$$

这里，F 代表我们提出的跨模态融合模块，用于完成上述步骤。后续内容将详细介绍编码器和我们的解码器。

3.1.1. 图像编码器

图像编码器 E_I 的目的是为特征融合和上采样提供多尺度特征。在本研究中，我们在实验部分使用了各种主干网络来评估所提出方法的有效性，包括 VGG-16、ResNet-101 和 Efficient-b7。

3.1.2. 文本编码器

文本编码器 E_T 将标签文本转换为数值特征，将 N 个标签词与 N 个连续向量关联起来。例如，在具有 20 个标签类别的 SBD 数据集中，我们将这 20 个标签文本输入到文本编码器中，从而产生相应的文本编码。与图像编码器类似，可以使用各种文本编码器来实现此目的。在本研究中，我们采用预训练的 CLIP 的文本编码器，特别是 ViT B/32 变体。值得注意的是，文本编码器在所有实验中都保持冻结状态，这表明它在训练过程中不会更新。

3.1.3. 解码器

解码器 D 在细化路径中将掩码编码与主干特征相结合。使用单独的卷积层来平滑特征和掩码编码对。然后使用反卷积层进行分辨率上采样。解码器重复此过程，直到特征的分辨率与输入图像匹配。最后，卷积层调整输出特征以与数据集的标签计数对齐。

3.2. 跨模态特征融合

我们贡献的重点在于融合模块 F，我们将对这一关键元素提供更详细的解释。如前所述，LSeg 的工作展示了对图像和文本特征融合的探索，这启发了我们的方法。这些特征可以成功融合，意味着它们的语义信息可能会相互改进。为了验证这一观念，我们引入了跨模态特征融合方法来实现这一目标。跨模

态特征融合能够利用不同模态(如图像、文本、音频等)携带的信息，最大程度地利用各模态的优势。例如，图像提供视觉信息，而文本提供语义上下文，二者结合可以提升理解的深度和准确性，同时，不同模态之间的潜在关系能够帮助模型更好地捕捉数据的共性和差异，从而增强特征表达能力。融合机制引入了不同来源的信息，能够提升模型的泛化能力。在面对复杂任务时，多种模态信息的结合能够提高模型对未知数据的适应性。当然，跨模态特征融合方法是有多种形式的，不同的融合方式对模型的整体性能会产生一定影响，为此我们进行了一些消融实验来进一步优化融合过程，我们通过将图像和文本特征进行不同方式的运算组合来寻找最优融合方式，详细结果见表 2。

3.2.1. 全局信息嵌入

在融合特征之前，必须确定最佳融合位置。主要目标是增强高级图像特征并提高其判别能力。对于文本特征是直接利用文本编码器的输出。对于图像特征，是利用图像编码器的输出，因为它通常具有最小的分辨率但包含最多的语义信息。

我们通过将特征的全局信息压缩到通道描述符[23]中。对于形状为 $N\times M$ 的文本特征，其中 N 表示标签数量， M 表示文本特征的长度，我们首先使用线性操作将特征压缩为 $N\times M'$ 的形状。

$$f_{T-fc}=FC\left(f_T\right) \tag{3-3}$$

通过这样的操作，我们从文本特征中获得了一种全局信息。我们对形状为 $B\times C\times H\times W$ 的高级图像特征执行对齐操作，将其转换为 $B\times C\times\left(H\times W\right)$。

接下来，我们将同步图像特征和文本特征的通道。这个调整是必要的，因为高级图像特征通常比文本特征具有更多的通道。

$$f_{I-fc}=\mathrm{Re}\,LU\left(FC\left(f_I\right)\right) \tag{3-4}$$

这里， FC 表示全连接层， $\mathrm{Re}\,LU$ 表示整流线性单元。

3.2.2. 图像－文本特征融合

一旦特征描述符的形状对齐，我们就会启动融合过程。特征融合函数必须既灵活又有效。我们的融合函数表示为：

$$f_{fusion}=\left(f_{T-fc}\odot f_{I-fc}\right)\times C \tag{3-5}$$

从公式(3-5)中，我们观察到函数将残差应用于两个特征的元素乘积。C 代表可学习的变量。

4. 实验

在本节中，我们将评估所提出方法的有效性。我们首先概述基本的实验设置。然后通过消融研究证明了所提方法的有效性。我们将该方法与 SBD 和 CityScapes 数据集上的最新语义边缘检测检测器进行了比较。

4.1. 实验设置

4.1.1. 数据集

我们使用两个数据集来评估语义边缘检测：SBD 和 CityScapes [24]。SBD 是来自 PASCAL VOC2011 的 11,355 幅图像，其中 8498 幅用于训练，2857 幅用于测试(20 个类别)。CityScapes 提供了 5000 张街景图像的像素级数据；我们使用 2975 张进行训练，500 张来自验证集的图像作为测试集(19 个类别)。SBD 和 CityScapes 数据集都是根据 SEAL 提供的数据转换代码，将语义分割数据集中的掩码文件转化为语义边缘检测的掩码文件。我们通过 SEAL 的代码生成数据集，然后进行后续的实验。在后续章节将展示在这两个数据集上各项类别的 ODS 得分以及所有类别的平均 ODS。

4.1.2. 实现细节

我们使用了以 VGG-16 [25]、ResNet-101 [26]和 EfficientNet-B7 [27]为骨干的编码器解码器网络来展示方法的有效性。所有编码器都在 ImageNet 预训练权重。CLIP 使用了 ViT-B/32 模型的文本编码器，该编码器在训练期间处于冻结状态。

训练采用 Adam 优化器(训练轮次：200，学习率：1e−4，权重衰减：5e−4)。两个数据集的图像大小均设置为 512 × 512。对于 SBD，通过随机裁剪调整整体图像比例(0.5 到 1.5)来填充和增强小于 512 × 512 的图像。对于 CityScapes，将大小为 2048 × 1024 的图像分成 512 × 512 的部分。两个数据集都进行了数据增强，包括随机翻转。

4.1.3. 评估指标

我们采用了最佳数据集大小(ODS)作为评估指标，遵循先前的研究[28]。我们将展示各个类别的 ODS 得分以及所有类别的平均 ODS (均值 ODS)。

4.2. 消融研究

所有实验均在 SBD 上进行，使用主干的编码器–解码器边缘检测器。我们在实验中使用的文本编码器是 CLIP ViT-B/32 变体。

4.2.1. 模型组件

在我们的消融研究中，评估了每个边缘检测器组件的有效性。第一个基线使用 EfficientNet-B7 作为唯一的主干。第二个基线引入了跨模态特征融合模块。结果见表 1。

Table 1. Ablation studies are performed on model components. The results highlight the effectiveness of our cross-modal feature fusion module
表 1. 对模型组件进行消融研究。结果突出了我们的跨模态特征融合模块的有效性

EfficientNet-B7	跨模态特征融合	平均 ODS
√	-	73.8
√	√	77.0

我们的跨模态特征融合模块方法显著提高了边缘检测器的性能，将基线提高 3.2%，实现了 77.0%的 ODS。这超过了之前的检测器，如 CASENet (71.4%)和 DDS (76.0%)。这些结果强调了我们方法的有效性。

4.2.2. 图像–文本融合方法

我们探索了特征融合函数的影响。当将图像特征和文本特征对齐到相同形状时，以下方法 $(f_T \odot f_I)$，$(f_T \odot f_I, \|f_T - f_I\|)$ 和 $(f_T, f_I, \|f_T - f_I\|)$ 对特征融合有效[29]。符号 $\odot$ 表示对两个特征进行连接操作。结果详见表 2。我们注意到我们的方法 $(f_T \odot f_I)$ 优于其他两种方法。这一结果凸显了所提方法的简单性和有效性。

Table 2. Comparison of feature fusion methods. Despite its simplicity, our method outperforms the state-of-the-art methods
表 2. 特征融合方法的比较。尽管我们的方法很简单，但它的表现却优于最好的方法

融合方法	平均 ODS
$(f_T \odot f_I)$	77.0
$(f_T \odot f_I, \|f_T - f_I\|)$	76.0
$(f_T \odot f_I, \|f_T - f_I\|)$	76.4

4.2.3. 基准网络比较

本节探讨了我们语言驱动框架中的不同主干 VGG-16、ResNet-101 和 EfficientNet-B7。表 3 中的定量结果显示了不同基准网络的性能。最显著的增强发生在用 ResNet-101 替换 VGG-16 时，结果增加了 3.2 个百分点。从 ResNet-101 过渡到 EfficientNet-B7 只会带来较为温和的 1.5 个百分点增强。

结果证明了我们的框架与不同先进网络的适应性。根据结果，选择 EfficientNet-B7 进行后续的最新比较。

Table 3. Ablation studies on state-of-the-art backbones show the flexibility of our approach and demonstrate its ability to work with different backbones and achieve good results
表 3. 对先进主干网的消融研究显示了我们方法的灵活性，证明了其与不同主干网配合并取得良好结果的能力

方法	基准网络	平均 ODS
语言辅助的语义边缘检测	VGG-16	72.3
	ResNet-101	75.5
	Efficient-b7	77.0

4.2.4. 与 LSeg 比较

LSeg 任务是应用在语义分割任务中的，而本文对 LSeg 任务进行了积极探索，并提出了一种跨模态特征融合的方法进行语义边缘检测任务。为了验证 LSeg 能否直接应用于语义边缘检测任务，我们进行了以下消融实验，详细结果见表 4，可以看到，LSeg 方法的直接应用不仅不能提升语义边缘检测的效果，而且还降低了。所以语义分割任务的方法，并不一定直接适用于语义边缘检测任务。

为此，我们还额外统计了一项数据分布，对语义分割数据集和语义边缘检测数据集的掩码文件进行了前景像素的占比计算，得到语义分割数据集的掩码文件前景像素占比达到了 30.5%，而语义边缘检测数据集的掩码文件前景像素占比仅 3.2%，前景像素占比的巨大差距更加说明了两个任务的本质区别。

Table 4. Ablation experiments compared with LSeg method
表 4. 与 LSeg 方法比较的消融实验

方法	平均 ODS
LSeg	72.1
Our	77.0

4.3. 最新技术比较

4.3.1. SBD 的结果

我们在 SBD 数据集上进行了实验，比较了 CASENet、STEAL、DDS 等方法，和我们的方法。CASENet 作为先驱，为后续的发展奠定了基础。STEAL、DDS 和相关方法都是从 CASENet 框架发展而来的。所有方法最初都是在 COCO 数据集[30]上进行预训练，然后在 SBD 上进行训练。定量结果见表 5，定性结果见图 2。

在定量结果方面，我们的方法实现了最先进的性能，平均 ODS 为 77.0，超过了 DDS (76.0)、STEAL (75.6)和其他方法。我们的方法简单灵活，利用语言驱动边缘检测，即使在存在噪声标签和缺失边界的情况下，也具有很强的稳健性。

4.3.2. CityScapes 的结果

在 CityScapes 的案例中，我们将我们的方法与 CASENet 和 DDS 进行了比较。两种方法都直接在 CityScapes 上进行训练。结果见表 6 和图 3。

Table 5. State-of-the-art comparisons are performed on the SBD dataset. Our method achieves the best performance
表 5. 在 SBD 数据集上进行最新比较。我们的方法取得了最佳性能

方法	飞机	单车	鸟	船	瓶子	公交	汽车	猫	椅子	牛	桌子	狗	马	机车	人	盆栽	羊	沙发	火车	屏幕	平均
CASENet	83.3	76.0	80.7	63.4	69.2	81.3	74.9	83.2	54.3	74.8	46.4	80.3	80.2	76.6	80.8	53.3	77.2	50.1	75.9	66.8	71.4
SEAL	84.9	78.6	84.6	66.2	71.3	83.0	76.5	87.2	57.6	77.5	53.0	83.5	82.2	78.3	85.1	58.7	78.9	53.1	77.7	69.7	74.4
STEAL	85.8	80.0	85.6	68.4	71.6	85.7	78.1	87.5	59.1	78.5	53.7	84.8	83.4	79.5	85.3	60.2	79.6	53.7	80.3	71.4	75.6
DDS	86.7	79.6	85.6	68.4	74.5	86.5	81.1	85.9	60.5	79.3	53.5	83.2	85.2	78.8	83.9	58.4	80.8	54.4	81.8	72.2	76.0
Our	86.1	78.0	86.3	69.9	73.1	86.5	80.5	86.8	64.2	85.7	53.9	85.6	86.1	79.7	82.5	58.5	84.3	60.7	83.0	69.4	77.0

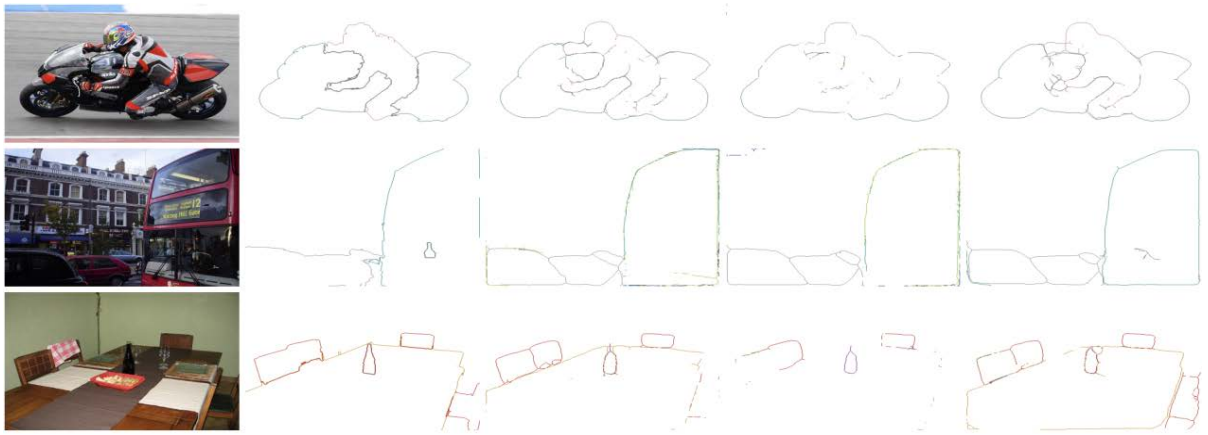

Figure 2. State-of-the-art comparison on the SBD dataset. The first column is the original image, the second column is the mask file, the third column is the detection result of CASENet, the fourth column is the detection result of DDS, and the fifth column is our result. The results show that our method preserves fine object details while presenting fewer artifacts, proving its effectiveness
图 2. 在 SBD 数据集上进行的最先进的比较。第一列为原始图像，第二列为掩码文件，第三列是 CASENet 的检测结果，第四列为 DDS 的检测结果，第五列是我们的结果。结果显示，我们的方法保留了精细的物体细节，同时呈现出更少的伪影，证明了其有效性

Table 6. On the CityScapes dataset, our framework achieves leading performance
表 6. 在 CityScapes 数据集中，我们的框架取得了领先的性能

方法	路	步道	建筑	墙壁	栅栏	电杆	交通灯	标志	植被	地形	天空	人	骑手	汽车	卡车	公交车	火车	机车	自行车	平均
CASENet	86.6	78.8	85.1	51.5	58.9	70.1	70.8	74.6	83.5	62.9	79.4	81.5	71.3	86.9	50.4	69.5	52.0	61.3	80.2	71.3
DDS	89.7	79.4	80.4	52.1	53.0	82.4	81.9	80.9	83.9	62.0	89.4	86.0	77.8	92.3	59.8	74.8	55.3	64.4	77.3	74.9
Our	90.0	80.2	80.6	51.2	55.2	82.6	81.7	82.6	84.0	62.8	89.1	85.9	76.2	91.8	59.2	79.6	64.7	63.4	76.3	75.6

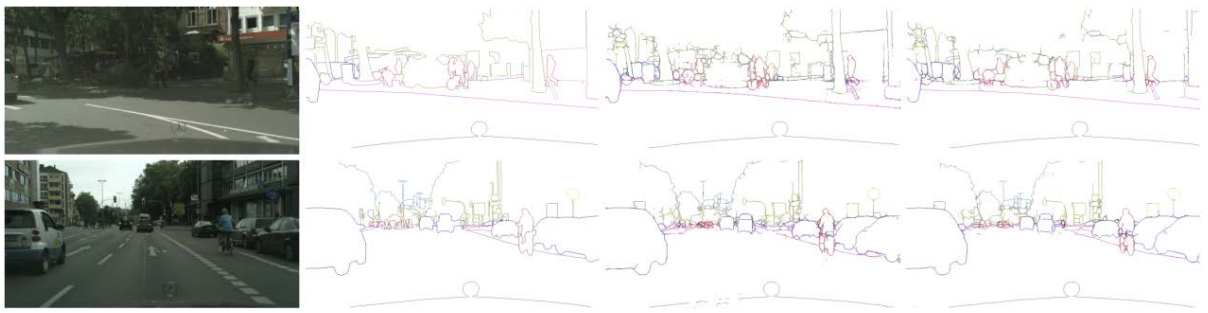

Figure 3. State-of-the-art comparison on the CityScapes dataset. The first column is the original image, the second column is the mask file, the third column is the detection result of CASENet, and the fourth column is our method. The results show that our method performs well in preserving fine details
图 3. 在 CityScapes 数据集上进行的最先进的比较。第一列为原始图像，第二列为掩码文件，第三列是 CASENet 的检测结果，第四列为我们的方法。结果显示在保留精细细节方面我们的方法表现出色

两组结果均表现出一致的优异性能。我们的方法不仅获得了领先的性能，而且在对象边界中表现出更精细的细节，噪声像素更少。结果证明了我们框架的泛化能力。

5. 结论

在本研究中，我们介绍了一种新方法，即语言驱动语义边缘检测，该方法利用来自文本嵌入的语义知识来重新校准边缘检测器的注意力。我们提出的跨模态特征融合模块被证明既简单又高效。我们方法所展示的灵活性和有效性凸显了利用语言增强边缘检测的巨大潜力。未来的研究将侧重于使用我们的方法探索零样本或少样本应用。

参考文献

[1] Acuna, D., Kar, A. and Fidler, S. (2019). Devil Is in the Edges: Learning Semantic Boundaries from Noisy Annotations. 2019 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Long Beach, 15-20 June 2019, 11067-11075. https://doi.org/10.1109/cvpr.2019.01133

[2] Liu, Y., Cheng, M., Fan, D., Zhang, L., Bian, J. and Tao, D. (2021) Semantic Edge Detection with Diverse Deep Supervision. *International Journal of Computer Vision*, **130**, 179-198. https://doi.org/10.1007/s11263-021-01539-8

[3] Yu, Z.D., Liu, W.Y., Zou, Y., Feng, C., *et al*. (2018) Simultaneous Edge Alignment and Learning. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 388-404.

[4] Yu, Z.D., Feng, C., Liu, M.-Y. and Ramalingam, S. (2017) CaseNet: Deep Category-Aware Semantic Edge Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, 21-26 July 2017, 5964-5973.

[5] Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S. and Malik, J. (2011) Semantic Contours from Inverse Detectors. 2011 *International Conference on Computer Vision*, Barcelona, 6-13 November 2011, 991-998. https://doi.org/10.1109/iccv.2011.6126343

[6] Li, B.Y., Weinberger, K.Q., Belongie, S., Koltun, V. and Ranftl, R. (2022) Language-Driven Semantic Segmentation. *International Conference on Learning Representations*, 25-29 April 2022.

[7] Radford, A., Kim, J.W., Hallacy, C., *et al*. (2021) Learning Transferable Visual Models from Natural Language Supervision. *International Conference on Machine Learning*, Online, 18-24 July 2021, 8748-8763.

[8] Canny, J. (1986) A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **8**, 679-698. https://doi.org/10.1109/tpami.1986.4767851

[9] Fram, J.R. and Deutsch, E.S. (1975) On the Quantitative Evaluation of Edge Detection Schemes and Their Comparison with Human Performance. *IEEE Transactions on Computers*, **24**, 616-628. https://doi.org/10.1109/t-c.1975.224274

[10] Kittler, J. (1983) On the Accuracy of the Sobel Edge Detector. *Image and Vision Computing*, **1**, 37-42. https://doi.org/10.1016/0262-8856(83)90006-9

[11] Perona, P. and Malik, J. (1990) Scale-Space and Edge Detection Using Anisotropic Diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **12**, 629-639. https://doi.org/10.1109/34.56205

[12] Lowe, D.G. (2004) Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, **60**, 91-110. https://doi.org/10.1023/b:visi.0000029664.99615.94

[13] Senthilkumaran, N. and Rajesh, R. (2009) Edge Detection Techniques for Image Segmentation—A Survey of Soft Computing Approaches. *International Journal of Recent Trends in Engineering*, **1**, 250-254.

[14] Siddiqui, M. and Medioni, G. (2010) Human Pose Estimation from a Single View Point, Real-Time Range Sensor. 2010 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, San Francisco, 13-18 June 2010, 1-8. https://doi.org/10.1109/cvprw.2010.5543618

[15] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, **25**, 1-9.

[16] Su, Z., Liu, W.Z., Yu, Z.T., *et al*. (2021) Pixel Difference Networks for Efficient Edge Detection.

[17] Xie, S.N. and Tu, Z.W. (2015) Holistically-Nested Edge Detection. *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, 7-13 December 2015, 1395-1403.

[18] Zhou, C., Huang, Y., Pu, M., Guan, Q., Huang, L. and Ling, H. (2023) The Treasure beneath Multiple Annotations: An Uncertainty-Aware Edge Detector. 2023 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Vancouver, 17-24 June 2023, 15507-15517. https://doi.org/10.1109/cvpr52729.2023.01488

[19] Liu, Y., Cheng, M.-M., Hu, X.W., Wang, K. and Bai, X. (2016) Richer Convolutional Features for Edge Detection.

[20] Deng, R.X., Shen, C.H., Liu, S.J., *et al*. (2018) Learning to Predict Crisp Boundaries. *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, 8-14 September 2018, 562-578.

[21] Deng, R. and Liu, S. (2020) Deep Structural Contour Detection. *Proceedings of the* 28*th ACM International Conference on Multimedia*, Seattle, 12-16 October 2020, 304-312. https://doi.org/10.1145/3394171.3413750

[22] Pu, M.Y., Huang, Y.P., Liu, Y.M., Guan, Q.J. and Ling, H.B. (2022) Edter: Edge Detection with Transformer. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, 18-24 June 2022, 1402-1412.

[23] Hu, J., Shen, L. and Sun, G. (2018) Squeeze-and-Excitation Networks. 2018 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-23 June 2018, 7132-7141. https://doi.org/10.1109/cvpr.2018.00745

[24] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., *et al*. (2016) The Cityscapes Dataset for Semantic Urban Scene Understanding. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 June 2016, 3213-3223. https://doi.org/10.1109/cvpr.2016.350

[25] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition.

[26] He, K.M., Zhang, X.Y., Ren, S.Q. and Sun, J. (2016) Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778.

[27] Tan, M.X. and Le, Q. (2019) EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *International Conference on Machine Learning*, Long Beach, 10-15 June 2019, 6105-6114.

[28] Hu, Y., Chen, Y.P., Li, X. and Feng, J.S. (2019) Dynamic Feature Fusion for Semantic Edge Detection.

[29] Reimers, N. and Gurevych, I. (2019). Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, 3-7 November 2019, 3982-3992. https://doi.org/10.18653/v1/d19-1410

[30] Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., *et al*. (2014) Microsoft COCO: Common Objects in Context. In: *Lecture Notes in Computer Science*, Springer, 740-755. https://doi.org/10.1007/978-3-319-10602-1_48