中图法分类号:TP18;TP391　　**文献标识码**:A　　**文章编号**:1006-8961(2026)01-0289-14

论文引用格式:Cao F, Zhang X W, Yue Z J, Li L and Shi M J. 2026. Vision-language model driven object counting. Journal of Image and Graphics, 31(1):0289-0302(曹锋, 张孝文, 岳子杰, 李莉, 史淼晶. 2026. 视觉语言模型驱动的目标计数. 中国图象图形学报, 31(1):0289-0302)[DOI:10.11834/jig.250119]

视觉语言模型驱动的目标计数

曹锋[1],张孝文[2],岳子杰[2],李莉[2],史淼晶[2*]

1. 浙江省轨道交通运营管理集团有限公司,杭州　310000; 2. 同济大学电子与信息工程学院,上海　201804

摘　要:**目的**　大型视觉语言模型的进展给解决基于文本提示的目标计数问题带来新的思路。然而,现有方法仍面临类别语义错位与解码器架构局限两大挑战。前者导致模型易将相似背景或无关类别误检为目标,后者依赖单一卷积神经网络(convolutional neural network,CNN)架构的局部特征提取,可能引发全局语义与局部细节的割裂,严重制约复杂场景下的计数鲁棒性。针对上述问题,提出跨分支协作对齐网络(cross-branch cooperative alignment network,CANet)。**方法**　其核心包括:1)双分支解码器架构:通过并行Transformer分支(建模全局上下文依赖)与CNN分支(提取细粒度局部特征),结合信息互馈模块实现跨分支的特征交互和密度图预测;2)视觉—文本类别对齐损失:通过约束图像与文本特征的跨模态对齐,迫使模型区分目标与干扰语义,实现对类别的准确检测。**结果**　在5个基准数据集上与先进的4种基于文本的目标计数方法进行比较实验。在FSC-147(few-shot counting-147)数据集上,CANet相较于性能第2的模型,在测试集上的平均绝对误差(mean absolute error,MAE)和均方根误差(root mean squared error,RMSE)分别降低1.22和8.45;在CARPK(car parking lot dataset)和PUCPR+(Pontifical Catholic University of Parana+ dataset)数据集的交叉验证实验上,相较于性能第2的模型,MAE分别降低0.08和3.58;在SHA(ShanghaiTech part-A)和SHB(ShanghaiTech part-B)数据集的交叉验证实验上,相较于性能第2的模型,MAE分别降低了47.0和9.8。同时也在FSC-147数据集上进行丰富的消融实验以验证算法的有效性,消融实验结果表明提出的方法针对两个问题做出了有效改进。**结论**　本文方法能够解决现有方法所面临的两个问题,使计数结果更加准确。本文方法在4个数据集的交叉验证实验均取得SOTA(state-of-the-art)的性能,表明了CANet在零样本目标计数任务中的强大泛化能力。

关键词:目标计数;视觉语言模型(VLM);文本提示;双分支解码器;信息互馈

Vision-language model driven object counting

Cao Feng[1], Zhang Xiaowen[2], Yue Zijie[2], Li Li[2], Shi Miaojing[2*]

1. Zhejiang Rail Transit Operation Management Group CO., LTD., Hangzhou　310000, China;

2. School of Electronic and Information Engineering, Tongji University, Shanghai　201804, China

Abstract: **Objective** In recent years, pre-trained large vision-language models (VLMs) have been widely used in numerous vision-language tasks. These models typically incorporate a vision encoder and a text encoder to facilitate image alignment with their corresponding text prompts in the embedding space. Inspired by their success, interest in solving the class-agnostic counting problem based on VLMs has surged, leveraging text prompts instead of exemplars to specify the target category of interest. However, existing methods still encounter two critical challenges: category semantic misalignment and decoder architecture limitations. The former leads to false detections, where models often mistakenly identify background regions or irrelevant categories as targets. The latter arises from the over-reliance on single-CNN architectures for local fea-

收稿日期:2025-04-03;**修回日期**:2025-06-06;**预印本日期**:2025-06-13

*通信作者:史淼晶　mshi@tongji.edu.cn

ture extraction, resulting in a lack of understanding of global semantic information, which severely compromises the capability of the model to accurately estimate object counts in complex scenes. This limitation also compromises the robustness of counting performance in complex real-world scenarios, such as scenes with dense occlusions, perspective distortions, or heterogeneous object scales. **Method** Aiming to address these issues, a cross-branch cooperative alignment network (CANet) is proposed. This study introduces a dual-branch decoder architecture: The Transformer branch employs self-attention mechanisms to model global contextual dependencies (e. g., spatial relationships and density distribution), while the CNN branch uses convolutional operations to extract fine-grained local features (e. g., edges and textures of small). A bidirectional information mutual feedback module dynamically bridges the two branches, facilitating cross-scale feature interaction. This module transfers global priors from the Transformer branch to realize refined local predictions in the CNN branch, while injecting local details from the CNN branch into the Transformer branch to enhance spatial coherence. Moreover, a vision-text category alignment loss is proposed: this loss enforces cross-modal alignment between global image embeddings and text prompts via contrastive learning. Specifically, visual embeddings are pulled closer to their corresponding text embeddings (anchor text prompts) while being pushed away from unrelated text embeddings using counterfactual prompts (shift text prompts). By optimizing semantic consistency between the visual and textual modalities, the model learns to distinguish target semantics from interfering information, substantially reducing false detections caused by category ambiguity. **Result** CANet is compared with four state-of-the-art text-promptable object counting methods on five benchmark datasets. On the FSC-147 dataset, CANet surpassed the second-best model, reducing the MAE and RMSE on the test set by 1. 22 and 8. 45, respectively. In cross-dataset evaluations, CANet achieved MAE reductions of 0. 08 and 3. 58 on CARPK and PUCPR+, respectively, compared to the second-best model. On the SHA and SHB datasets, CANet reduced the MAE by 47. 0 and 9. 8, respectively. Additionally, comprehensive ablation studies on the FSC-147 dataset demonstrated the effectiveness of the proposed method. For instance, the performance of two single-branch and dual-branch variants was compared using only CNN or Transformer to validate the effectiveness of the dual-branch decoder. The results showed that homogeneous dual-branch variants substantially outperform their corresponding single-branch counterparts, but none surpass the proposed dual-branch decoder. This finding indicates that single-architecture decoders have inherent limitations, and leveraging the complementary characteristics of both architectures can help simultaneously focus on global- and local-scale predictions, thereby improving overall performance. Aiming to verify the effectiveness of the information mutual feedback module, ablation studies were conducted, identifying a performance drop with the removal of the module. Visualization results were provided to validate the effectiveness of the vision-text category alignment loss, revealing that this loss substantially reduces false detections. Furthermore, this study demonstrated the effectiveness of CANet using different encoders and training strategies. **Conclusion** The proposed CANet effectively addresses the challenges of category semantic misalignment and limitations in decoder architecture commonly encountered in text-prompted object counting. By leveraging global-local feature interaction and precise cross-modal alignment, CANet achieves state-of-the-art performance in cross-dataset evaluations, demonstrating strong generalization capabilities in zero-shot counting tasks. This work presents a robust and scalable framework for open-world scenarios, where objects notably vary in scale, density, and contextual complexity.

Key words: object counting; visual-language model (VLM); text-promptable; dual-branch decoder; information mutual feedback

0 引 言

目标计数(object counting)任务旨在估计图像或视频中特定目标的数量。传统方法通过在预定义类别(如人群(Liang等,2022)、车辆(Hsieh等,2017))的标注数据上训练深度神经网络来实现计数,这种类别特定目标计数(class-specific object counting)方法依赖封闭世界的类别假设,导致其面对未见类别缺乏良好的泛化能力。因此Lu等人(2019)提出类别无关目标计数(class-agnostic object counting),使用示例指定计数类别,并利用示例特征与图像特征的相似度匹配实现开放类别的目标计数。但在动态多变的场景中,依赖人工标注示例不仅成本高昂且

易引入偏差，限制了应用。

近年来，预训练的视觉语言模型（Radford 等，2021；Li 等，2022）在多个计算机视觉任务中取得了显著成果。这些模型通常包括一个视觉编码器和一个文本编码器，通过海量图文对齐数据构建了视觉—语言共享的特征空间。受其成功的启发，基于视觉语言模型解决类别无关的计数问题的研究逐渐增多。这些方法利用文本描述替代示例，因此称为基于文本提示的目标计数方法（text-promptable object counting）。具体而言，通过视觉编码器提取图像特征，同时利用文本编码器将描述语义的文本提示映射为特征向量，最终通过跨模态的相似度匹配实现计数。为增强模型对复杂场景的适应性，现有方法（Jiang 等，2023；Kang 等，2024）引入了多模态对比学习策略：在特征空间中，将包含目标类别的图像块特征向对应文本提示对齐，同时强制不包含目标类别的图像块特征远离文本提示。文本提示的计数方法具备良好的灵活性，能够通过简单调整转化为类别特定计数方法。例如，当固定提示词为“people”时，模型可直接应用于人群计数任务。

基于文本提示的目标计数方法在开放场景中展现了较强的适应性，但在应用中仍面临类别语义错位与解码器架构局限的双重挑战。

首先，现有方法的本质更接近于无参考目标计数（Ranjan 和 Nguyen，2023）的扩展：计数网络仅依靠预训练模型学习到的类别差异区分物体，其训练没有强调类别之间的差异，因此未能充分学习到类别的语义特征。如图 1 所示，现有方法（Jiang 等，2023）在预测一些类别时难以与颜色、形状相似的背景区分开来，导致预测的密度峰值偏离目标中心，存在语义错位问题。具体而言，语义错位是指计数模型未能将文本提示的语义与视觉特征中的目标区域对齐，导致模型错误聚焦于背景或相似类别。本文中提及的语义错位问题指的是模型跨模态对齐能力的不足。例如，即使使用准确的提示词如“圆形细胞”，模型仍可能因跨模态特征偏差而将背景区域识别为“圆形细胞”。针对此问题，本文提出视觉—文本类别对齐损失，通过构建描述目标的锚点文本提示与描述干扰类别的偏移文本提示，显式约束图像特征逼近目标语义并远离干扰信息。如图 1 所示，本文方法预测的密度峰值精准聚焦于目标中心。

其次，现有方法（Jiang 等，2023；Kang 等，2024；Amini-Naieni 等，2023；Xu 等，2023）仍普遍采用卷积神经网络（convolutional neural network，CNN）构建密度解码器，而基于 Transformer 的解码器架构尚未得到充分研究。本文首次系统化探索了 Transformer 架构作为解码器的可行性，实验表明，Transformer 解码器在主流数据集上可达到与 CNN 相当的计数性能。进一步分析发现，二者在特征捕获能力上呈现显著互补性：CNN 凭借局部感受野优势，在局部区域的计数精度更高；而 Transformer 通过自注意力机制建模长距离依赖，在全局预测方面表现更优。这一发现与两类模型的基础特性一致，同时也揭示了单一架构的固有局限——局部细节与全局语义的割裂可能会降低复杂场景的计数鲁棒性。基于上述发现，本文提出双分支解码器，通过并行化 CNN 分支与 Transformer 分支实现特征协同建模。具体而言，CNN 分支聚焦局部纹理与细节特征提取，而 Transformer 分支学习全局上下文语义表示。进一步地，本文设计了信息互馈模块实现两个分支之间的特征交互：一方面将 CNN 的局部细节特征注入 Transformer 分支以增强细粒度感知；另一方面将 Transformer 的全局信息反馈至 CNN 分支以提升空间一致性。在双分支特征交互后，分别通过 1 × 1 卷积将双分支的特征解码为密度图，随后通过平均融合生成兼具局部细节与全局一致性的密度图。本文首次在

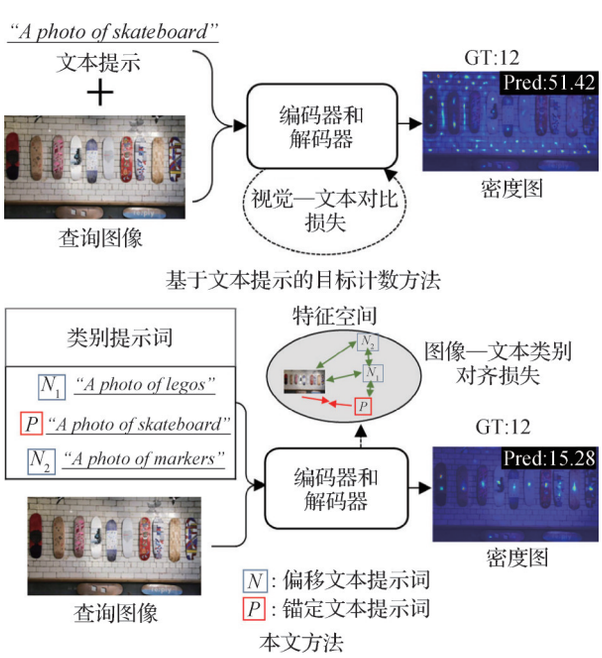


图 1　基于文本提示的目标计数方法与本文方法比较

Fig. 1　Comparison between existing text-promptable counting methods and our method

目标计数任务的解码器上实现了局部与全局特征的协同,实验表明其显著提高密集遮挡、透视失真等复杂场景下的计数精度与模型鲁棒性。

针对现有方法在类别语义错位与解码器架构局限上的不足,本文提出跨分支协作对齐网络(cross-branch cooperative alignment network, CANet)。首先,基于目标类别标签生成锚点提示与偏移提示,并通过文本编码器提取特征;接着,视觉编码器提取查询图像的特征,并通过图像—文本类别对齐损失优化跨模态特征距离;进一步地,通过注意力机制将锚点语义信息注入视觉特征,增强模型对目标区域的感知能力。在解码阶段,CANet采用双分支架构:CNN分支聚焦局部纹理细节,Transformer分支关注全局信息处理,二者通过信息互馈模块实现特征增强。两个分支的特征分别经过卷积层预测为密度图,最终通过平均融合得到预测结果。在FSC-147(few-shot counting-147)(Ranjan 等,2021)、CARPK(car parking lot dataset)(Hsieh 等,2017)、PUCPR+(Pontifical Catholic University of Parana+ dataset)(Hsieh等,2017)及ShanghaiTech(Zhang等,2016)等数据集上的实验结果表明,CANet显著优于最新的方法。

1 相关工作

1.1 类别特定目标计数

类别特定目标计数方法聚焦于对图像或视频中预定义类别的目标(如人群((Liang等,2022;Zhang等,2016)和车辆(Mundhenk 等,2016;Hsieh 等,2017)等)进行精确数量估计。当前主流的方法可分为基于检测(Liang等,2022;Liu 等,2023a,2019;张芯源 等,2025)与基于回归(曹锋 等,2024;邹敏 等,2023;Ranasinghe 等,2024;Guo 等,2024)的两类范式。基于检测的方法依托目标检测网络生成目标实例的检测框,通过统计检测框的数量实现计数。Liu等人(2023a)提出PET(point query Transformer)网络,基于DETR(detection Transformer)框架(Carion等,2020)预测目标的中心点,能够获取人群分布与密度信息;基于回归的方法通过端到端的网络回归密度图,然后对密度图积分得到预测个数。例如Ranasinghe等人(2024)提出基于扩散模型的密度图生成方法,借助降噪扩散过程提高人群密度预测的鲁棒性。

1.2 类别无关目标计数

类别无关目标计数方法通过引入少量目标示例(如参考框或图像块)表征待计数类别,基于示例特征与图像区域的相似度匹配完成计数(Ranjan等,2021;Liu 等,2023b;Lu 等,2019;Pelhan等,2024;Shi等,2022;Đukić 等,2023)。代表性工作如Lu 等人(2019)提出GMN(generic matching network)网络,首次将类别无关计数任务形式化为图像与示例间的特征匹配问题,通过图像和示例之间的关联度实现计数;Liu 等人(2023b)提出CounTR(counting Transformer)利用基于注意力机制的特征融合模块,优化示例与图像之间的特征距离,提升密度图预测精度。

1.3 基于文本提示的目标计数

基于文本提示的目标计数方法通过使用文本提示替代视觉示例指导模型进行计数。当前主流方法采用视觉语言模型分别提取文本特征与图像特征进行特征交互和密度预测。例如,VLCounter(visual-language counter)(Kang等,2024)引入参数高效微调策略,在特征编码阶段将文本提示信息注入视觉特征,接着通过CNN解码器预测密度图。除了基于密度图预测的方法,Paiss等人(2023)基于对比学习的训练范式,从LAION-400M(Schuhmann 等,2021)中筛选158 k含目标数量的图文对构建CountBench数据集训练CLIP(contrastive language- image pre-training),以此赋予CLIP基础的计数能力,但仅支持10以内的样本。

1.4 预训练的视觉语言模型

视觉语言预训练模型通过联合学习大规模图像—文本对齐数据,提升了跨模态语义理解能力,成为推动多模态学习发展的核心技术。其中,CLIP(Radford 等,2021)作为里程碑式工作,采用对比学习策略将图像与文本映射至共享特征空间,通过最大化匹配图文对之间的特征余弦相似度实现了零样本泛化,为图像分类、文本生成等任务提供了通用表征基础。为增强模型适应性,Cherti 等人(2023)提出通过优化训练策略与数据增强方法进一步提高CLIP的性能,并开源了CLIP的训练实现,为后续研究者深入探索CLIP的底层机制奠定了可复现的实验基础。与此同时,BLIP(bootstrapping language-image pre-training)(Li 等,2022)创新性地统一了视觉语言的理解与生成能力,提出Captioner-Filter机

制降低文本噪声干扰，在视觉问答、图像描述等任务表现优异。这些模型为解决人脸伪造检测（王诗雨等，2025）、青瓷跨知识图谱（肖刚 等，2025）构建等任务提供了新的解决思路。

2 方 法

2.1 概述

给定查询图像 $\boldsymbol{I}\in\mathbf{R}^{H\times W\times 3}$ 与目标类别标签 C，本文提出 CANet，旨在输出与类别相关的密度图 $\boldsymbol{D}$。如图 2 所示，CANet 首先基于类别标签 C 生成锚点文本提示（anchor text prompt）P^C，同时构建偏移文本提示（shift text prompt）作为干扰项，二者分别与查询图像 $\boldsymbol{I}$ 组成匹配的正样本对与非匹配的负样本对。通过视觉语言模型的编码器提取文本特征 $\boldsymbol{F}^C$ 与图像特征 $\boldsymbol{F}^I$ 后（详见 2.2 节），本文提出视觉—文本类别对齐损失，约束图像编码器能够更精准地提取与目标类别相关的特征（详见 2.4 节）。接着通过多模态融合机制将锚点文本特征的信息注入图像特征得到能感知目标类别的视觉特征 $\boldsymbol{F}^V$。在特征解码阶段，CANet 采用双分支解码器（详见 2.3 节）预测密度图：其中，Transformer 分支建模全局上下文语义，CNN 分支提取局部细节特征，二者通过层级化的信息互馈模块实现特征交互与聚合，充分发挥局部与全局视觉特征的优势。最终通过卷积层回归得到密度图。

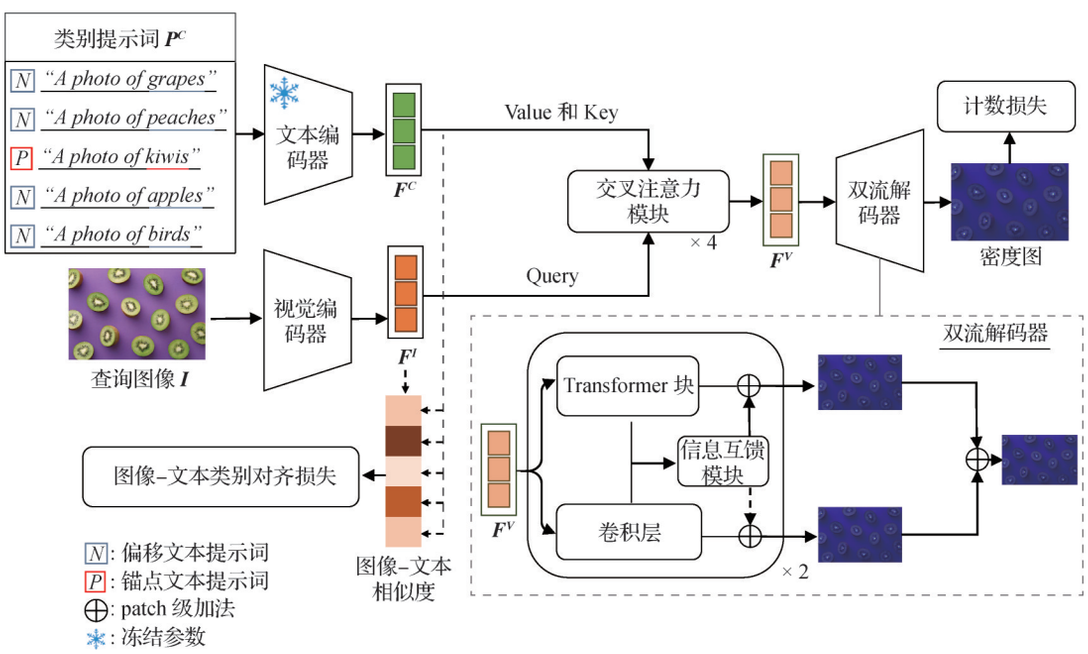


图 2 本文方法的流程图

Fig. 2 Flow chart of the proposed method

2.2 特征提取和融合

类别提示词用于指定查询图像 $\boldsymbol{I}$ 中待计数的目标类别。其生成采用模板 {*a photo of* [class]}，其中 [class] 被替换为类别标签。通过明确目标类别，提示词引导模型聚焦于指定类别的计数任务。具体而言，基于类别标签 C 生成的锚点文本提示 P^C 与查询图像 $\boldsymbol{I}$ 组成了视觉—文本正样本对，同时基于其他类别生成的偏移文本提示与图像组成负样本对。如，有一幅含有 14 个猕猴桃的图像，那么锚点提示为 {*a photo of kiwis*}，偏移提示可以为 {*a photo of grapes/peaches/apples/birds*}。二者通过对比学习的方式以优化模型性能，通过区分正负样本强化模型区分不同类别的能力。在这个过程中，本文发现往往是那些难以区分的样本起到了关键作用，因此在构造偏移提示的时候优先选取与目标类别相似度更高的类别。具体而言，计算每幅训练图像对应的目标类别文本特征与其他类别文本特征的余弦相似度，从中选取前 k 个相似度高的类别生成偏移提示词。这些类别对模型来说更加难以区分，能够进一步让模型学习到类别之间的差异，从而提高计数精度。

生成类别提示词后，通过视觉语言模型的文本

编码器提取对应的文本特征 $\boldsymbol{F}^C \in \mathbf{R}^{1\times d}$。同时，利用图像编码器从查询图像 $\boldsymbol{I}$ 中提取视觉特征 $\boldsymbol{F}^I \in \mathbf{R}^{N\times d}$，其中，$N$ 表示提取的视觉特征的个数，d 表示视觉文本特征的维度。

为了聚焦于对指定类别的计数，需要将文本特征 $\boldsymbol{F}^C$ 的类别信息注入到视觉特征 $\boldsymbol{F}^I$ 中，生成能感知目标类别的融合特征 $\boldsymbol{F}^V$。本文采用交叉注意力机制（cross-attention）实现两种模态的特征融合，表示为

$$\boldsymbol{F}^V = \boldsymbol{F}^I + f_{\text{MLP\&LN}}(CA(\boldsymbol{F}^I, \boldsymbol{F}^C)) \tag{1}$$

式中，$CA(x, y)$ 表示以 x 为查询（query）以及 y 为键（key）与值（value）进行交叉注意力计算。本文通过计算视觉特征与文本特征的交叉注意力权重，将文本语义信息注入视觉特征中。给定图像特征 $\boldsymbol{F}^I$ 和文本特征 $\boldsymbol{F}^C$，计算得到交叉注意力权重 $\boldsymbol{A} = \text{softmax}(\frac{\boldsymbol{F}^I\boldsymbol{W}_Q(\boldsymbol{F}^C\boldsymbol{W}_K)^{\mathrm{T}}}{\sqrt{d}})$ 及输出 $CA(\boldsymbol{F}^I, \boldsymbol{F}^C) = \boldsymbol{A}\cdot(\boldsymbol{F}^C\boldsymbol{W}_V)$，$\boldsymbol{W}_Q$、$\boldsymbol{W}_K$ 和 $\boldsymbol{W}_V \in \mathbf{R}^{d\times d}$ 为可学习的投影矩阵，分别将图像特征映射为查询向量、键向量和值向量；softmax 函数沿图像特征维度归一化注意力权重；$\sqrt{d}$ 为缩放因子，防止点积数值过大。$f_{\text{MLP\&LN}}$ 表示进行层归一化（layer normalization，LN）后通过多层感知机（multilayer perceptron，MLP）进行非线性映射操作。MLP 由两个全连接层构成，中间通过 ReLU（rectified linear unit）激活函数实现非线性映射；LN 用于在层级上将特征进行归一化处理以稳定训练过程。通过上述机制，类别信息被有效地融入到视觉特征中，使得生成的融合特征能够解码出待计数类别的密度信息。

2.3 双分支解码器

通过特征交互获得融合特征 $\boldsymbol{F}^V$ 后，本文设计了双分支解码器（简称双分支解码器）来预测目标类别的密度图 $\boldsymbol{D}^*$。双分支解码器由 CNN 分支、Transformer 分支以及层级化的信息互馈模块组成。其中，CNN 分支擅长捕捉局部细节特征，而 Transformer 分支通过自注意力机制建模全局上下文信息，二者结合显著提升了密度图预测的精度与鲁棒性。具体流程如下：融合特征 $\boldsymbol{F}^V$ 同时进入两个分支，分别生成中间特征 $\boldsymbol{F}^V_{\text{CNN}}$ 和 $\boldsymbol{F}^V_{\text{Trans}}$。随后，通过互馈模块将 Transformer 分支的全局信息适配至 CNN 分支，增强其对全局分布的理解；同时将 CNN 分支的局部信息反馈至 Transformer 分支，强化其对细节特征的感知。两个分支的输出分别通过 1×1 卷积层生成预测密度图 $\boldsymbol{D}^*_{\text{CNN}}$ 和 $\boldsymbol{D}^*_{\text{Trans}}$，通过平均加权得到最终的预测结果 $\boldsymbol{D}^*$。

信息互馈模块通过通道权重标定模块（channel-excitation block，CE）实现了两个分支特征间的交互来传递信息。CE 模块的想法源于 SE（squeeze-and-excitation）（Hu 等，2018）模块，不过 SE 模块是针对单模态特征设计的，本文将之拓展为双模态特征之间的交互。如图 3 所示，以全局信息传递为例，CE 模块通过 MLP 将 Transformer 分支特征 $\boldsymbol{F}^V_{\text{Trans}}$ 投影至 CNN 分支的通道维度，生成通道重标定权重，然后基于通道权重加权原始特征得到反馈特征，其计算为

$$\boldsymbol{F}^{\text{CE}}_{\text{Trans}} = f_{\text{MLP}}\cdot f_{\text{ReLU}}\cdot f_{\text{MLP}}(\boldsymbol{F}^V_{\text{Trans}}) \otimes \boldsymbol{F}^V_{\text{CNN}} \tag{2}$$

式中，$\otimes$ 表示通道级逐元素乘法，f_{ReLU} 为激活函数，f_{MLP} 为多层感知机操作。

通过通道权重标定模块，可生成反馈特征 $\boldsymbol{F}^{\text{CE}}_{\text{Trans}}$，其与 CNN 分支的中间特征逐块相加，得到精炼的中间特征 $\boldsymbol{F}^{\text{Enh}}_{\text{Trans}}$。类似地，通过调换互馈模块的输入特征，可实现 CNN 分支中的局部信息向 Transformer 分支的传递。

两个分支的输出分别通过 1×1 卷积层生成对应的预测密度图 $\boldsymbol{D}^*_{\text{CNN}}$ 和 $\boldsymbol{D}^*_{\text{Trans}}$，然后平均加权得到兼具细粒度感知与空间一致性的预测结果 $\boldsymbol{D}^* = 0.5\boldsymbol{D}^*_{\text{CNN}} + 0.5\boldsymbol{D}^*_{\text{Trans}}$。

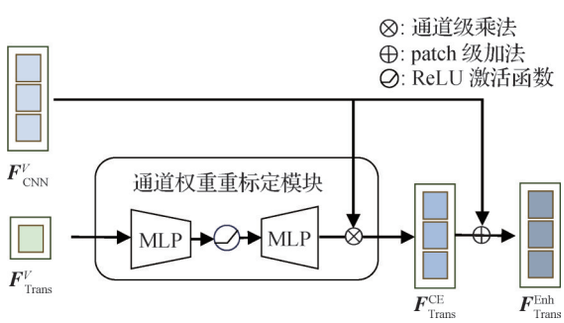


图 3 Transformer 分支信息适配到 CNN 分支

Fig. 3 Adaptation from Transformer branch to CNN branch

2.4 模型优化

本文设计了两种损失函数，分别针对密度图回归与跨模态对齐进行优化，以提升预测性能与鲁棒性。

2.4.1 计数损失

计数损失 L_{count} 采用真实密度图 $\boldsymbol{D}$ 与预测密度图 $\boldsymbol{D}^*$、$\boldsymbol{D}^*_{\text{CNN}}$ 和 $\boldsymbol{D}^*_{\text{Trans}}$ 的像素均方误差（mean square error，MSE）进行约束，其计算为

$$L_{\text{count}} = \left\|\boldsymbol{D}^* - \boldsymbol{D}\right\|_2^2 + \left\|\boldsymbol{D}^*_{\text{Trans}} - \boldsymbol{D}\right\|_2^2 + \left\|\boldsymbol{D}^*_{\text{CNN}} - \boldsymbol{D}\right\|_2^2 \tag{3}$$

该损失通过直接最小化预测密度图与真实密度图

的像素误差，确保模型在密度图估计任务中的准确性。

2.4.2　图像—文本类别对齐损失

为了解决现有方法存在的跨模态类别语义错位的问题，本文提出图像—文本类别对齐损失 L_{align}，以优化模型对目标类别的感知能力。具体而言，将锚点文本提示词 P_0^C 与查询图像组成正样本对；同时从训练集中随机选取 k 个其他类别标签 $\{C_i\}_{i=1}^{k}$ 构造偏移文本提示 $\{P_i^C\}_{i=1}^{k}$ 作为负样本对。样本对之间的相似度 $\{s^i\}_{i=0}^{k}$ 为编码器提取到的图像特征和文本特征之间的余弦相似度，则对齐损失为

$$L_{\text{align}} = -\log\left(\frac{\exp(s^0)}{\sum_{i=0}^{k}\exp(s^i)}\right) \tag{4}$$

对齐损失通过迫使模型区分目标与干扰类别，提高模型对类别的识别能力，缓解了语义错位的问题。

2.4.3　整体损失

训练模型的整体损失为上述两种损失的加权和，即

$$L = L_{\text{count}} + \mu L_{\text{align}} \tag{5}$$

式中，μ 为对齐损失的权重。通过对计数精度、类别对齐的多重约束，模型能够更精确地完成目标计数任务，同时具备更强的泛化性能。

3　实　验

3.1　实验细节

3.1.1　数据集

本文采用 FSC-147（Ranjan 等，2021）数据集作为训练集，该数据集包含 147 个类别的 6 135 幅图像。FSC-147 的训练集包含 89 个类别的 3 659 幅图像，验证集包含 29 个类别的 1 286 幅图像，测试集包含 29 个类别的 1 190 幅图像。为评估方法的泛化性能，除使用 FSC-147 的验证集与测试集外，还引入了 CARPK（Hsieh 等，2017）、PUCPR+（Hsieh 等，2017）和 ShanghaiTech（Zhang 等，2016）数据集进行交叉验证实验。其中，CARPK 的测试集包含 459 幅约 90 000 辆车的无人机航拍图像，PUCPR+的测试集则包含了 25 幅共计 16 456 辆车的航拍图像。ShanghaiTech 分为 A（SHA）和 B（SHB）两部分，其测试集分别包含 182 幅网络收集的人群图像与 316 幅上海公共场景下的人群图像。上述数据集在场景复杂度和目标分布上与 FSC-147 存在显著差异，能够有效验证模型的跨域适应性。

3.1.2　实施细节

本文采用预训练的 DINOv2（self-distillation with no labels）（Oquab 等，2024）作为视觉编码器，BERT（bidirectional encoder representations from Transformers）（Devlin 等，2019）作为文本编码器。训练过程中，本文冻结文本编码器权重并微调图像编码器，遵循 Amini-Naieni 等人（2023）的训练策略。训练时每个样本使用 1 个锚定提示词和 $k = 5$ 个偏移提示词，测试时仅使用 1 个锚定提示词。模型及其内部模块（如 MLP 和 LN 等）训练使用式（5）的损失优化。式（5）中的权重设为 1×10^{-1}，并使用学习率为 1×10^{-4}、衰减系数为 0.33 的 AdamW（Loshchilov 和 Hutter，2019）优化器，在 NVIDIA A40 GPU 上训练 200 个轮次（epoch），批量大小（batch size）设置为 32。

3.1.3　基线网络及评估指标

为验证 CANet 的有效性，本文设置了一个基线网络：其编码器与 CANet 一致，但在交互注意力模块后仅使用单分支的 CNN 解码器预测密度图，且仅采用计数损失 L_{count} 进行优化。

评估指标采用平均绝对误差（mean absolute error，MAE）和均方根误差（root mean square error，RMSE），与现有研究（Jiang 等，2023；Kang 等，2024；Amini-Naieni 等，2023）保持一致。

3.2　与 SOTA 方法的实验比较结果

表 1 展示了 CANet 在 FSC-147 数据集上的实验结果。如表 1 所示，CANet 与 5 种基于示例的目标计数方法及 4 种基于文本提示的目标计数方法进行了对比。实验表明，基于示例的方法整体优于基于文本提示的方法，这是因为文本提示仅提供宽泛的类别描述，难以根据不同的样本捕捉类内差异，而从样本中裁剪的示例可以捕捉到更丰富的类别细节特征和类内差异。尽管如此，CANet 与基于示例方法的性能差距较小，仅略逊于 CounTR（Liu 等，2023b）和 UPBC（unified prompt-based counting）（Lin 和 Chan，2024）等先进模型。其次，在 FSC-147 测试集上，CANet 在基于文本提示的方法中取得了最优性能：相较于最先进方法（state-of-the-art，SOTA），如 CounTX（Amini-Naieni 等，2023），MAE 与 RMSE 分别降低 1.22 与 8.45；相较于基线模型，MAE 与 RMSE 分别减少 5.05 与 11.97。

表1 在FSC-147数据集上的定量对比

Table 1 Quantitative comparisons on FSC-147 dataset

方法	提示	验证集		测试集	
		MAE	RMSE	MAE	RMSE
GMN(Lu等,2019)	示例	29.66	89.91	26.52	124.57
FamNet(Ranjan等,2021)	示例	23.75	79.44	24.90	112.68
BMNet+(Shi等,2022)	示例	15.74	58.53	14.62	91.83
CounTR(Liu等,2023b)	示例	13.13	49.83	11.95	91.23
UPBC(Lin和Chan,2024)	示例	12.80	48.65	11.86	89.40
ZSOC(Xu等,2023)	文本	26.93	88.63	22.09	115.17
CounTX(Amini-Naieni等,2023)	文本	17.70	63.61	15.73	106.88
CLIP-Count(Jiang等,2023)	文本	18.79	61.18	17.78	106.62
VLCounter(Kang等,2024)	文本	18.06	65.13	17.05	106.16
Baseline(基线)	文本	20.00	66.62	19.56	110.40
CANet(本文)	文本	**15.29**	**57.72**	**14.51**	**98.43**

注:加粗字体表示文本提示方法中各列的最优结果。

基于文本的提示方法中,CounTX使用类别名作为提示词,而CLIP-Count(Jiang等,2023)和VLCounter(Kang等,2024)使用模板{a photo of [class]}作为提示词。本文使用了模板{a photo of [class]}作为锚点提示词,同时修改[class]生成偏移提示词来强化模型的识别能力。在图4中给出了可视化结果,计数值显示在图像及预测密度图的右上角。这进一步验证了CANet的优越性。无论输入图像中目标类别的分布如何,其密度图均能精准聚焦于目标区域,极少出现在背景或无关类别区域,表明模型能够有效利用文本提示引导计数任务,并显著缓解现有方法的类别语义错位问题。

为进一步评估CANet的泛化能力,本文进行了跨数据集验证实验。按照现有研究(Jiang等,2023;Kang等,2024;Shi等,2022)的设置,在FSC-147数据集上训练CANet,并在CARPK、PUCPR+和ShanghaiTech数据集上直接评估,无需微调。如表2所示,CANet在CARPK与PUCPR+数据集上的MAE性能优于VLCounter(Kang等,2024),展现了强大的跨域适应性。表3展示了ShanghaiTech数据集上的泛化

表2 在CARPK和PUCPR+数据集上的交叉评估结果

Table 2 Cross-dataset evaluation on CARPK and PUCPR+ datasets

方法	提示	CARPK		PUCPR+	
		MAE	RMSE	MAE	RMSE
FamNet(Ranjan等,2021)	示例	28.84	44.47	87.54	117.68
BMNet+(Shi等,2022)	示例	10.44	13.77	62.42	81.74
CounTR(Liu等,2023b)	示例	5.75	7.45	–	–
CounTX(Amini-Naieni等,2023)	文本	11.96	16.61	–	–
CLIP-Count(Jiang等,2023)	文本	8.13	10.07	–	–
VLCounter(Kang等,2024)	文本	6.46	8.68	48.94	**69.08**
Baseline(基线)	文本	9.15	11.66	61.23	87.35
CANet(本文)	文本	**6.38**	**8.26**	**45.36**	71.64

注:加粗字体表示文本提示方法中各列的最优结果。“–”表示该方法在对应数据集上的结果未在原论文中报告。

结果。相较于 CLIP-Count(Jiang 等,2023),CANet 在 SHA 上将 RMSE 从 308.4 降低至 255.8,在 SHB 上将 MAE 从 45.7 降低至 38.9。这些结果表明,CANet 能够有效学习与类别相关的密度概念,在未知场景下具备优异的泛化性能。

图 5 展示了在 ShanghaiTech 人群数据集上的可视化结果。计数值显示在图像及预测密度图的右上角。自顶向下依次为:查询图像、CNN 分支预测、Transformer 分支预测及双分支解码器融合预测结果。图 5 的可视化结果进一步验证了 CANet 在人群计数任务中的良好表现和泛化性能,可以看到模型预测的密度峰值都在人群上。

表 3　在 ShanghaiTech 数据集上的交叉评估结果

Table 3　Cross-dataset evaluation on ShanghaiTech dataset

方法	SHA		SHB	
	MAE	RMSE	MAE	RMSE
RCC(Hobley 和 Prisacariu,2022)	240.1	366.9	66.6	104.8
CLIP-Count(Jiang 等,2023)	192.6	308.4	45.7	77.4
Baseline(基线)	204.1	322.8	48.7	78.1
CANet(本文)	**145.6**	**255.8**	**38.9**	**63.9**

注:加粗字体表示各列最优结果。

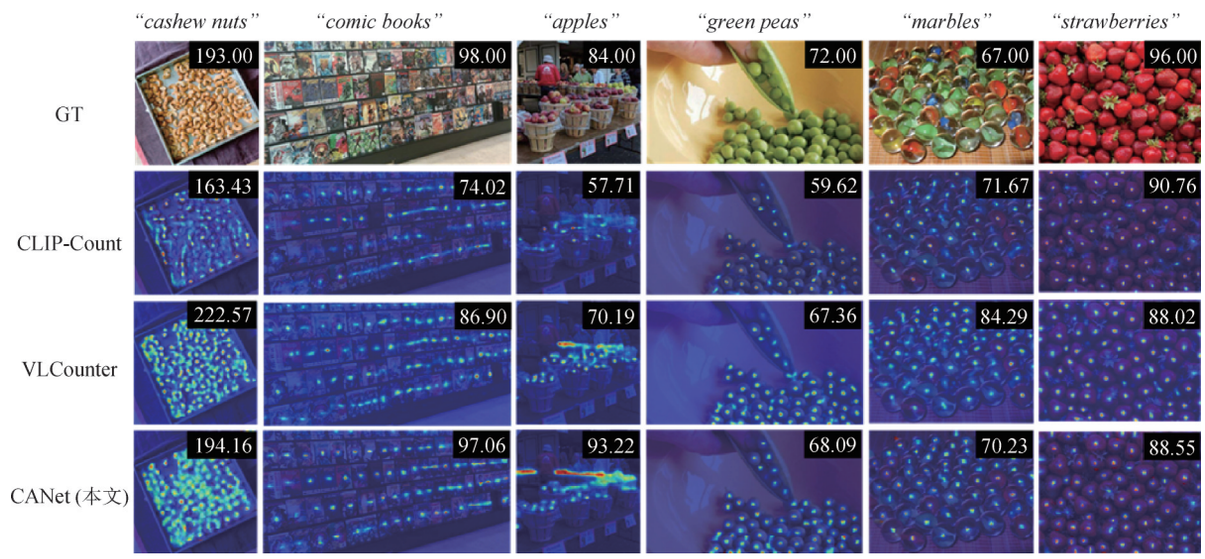


图 4　基于文本提示的目标计数方法在多种目标类别上的可视化结果

Fig. 4　Visualization comparison between different text-promptable object counting methods across a diverse range of object categories

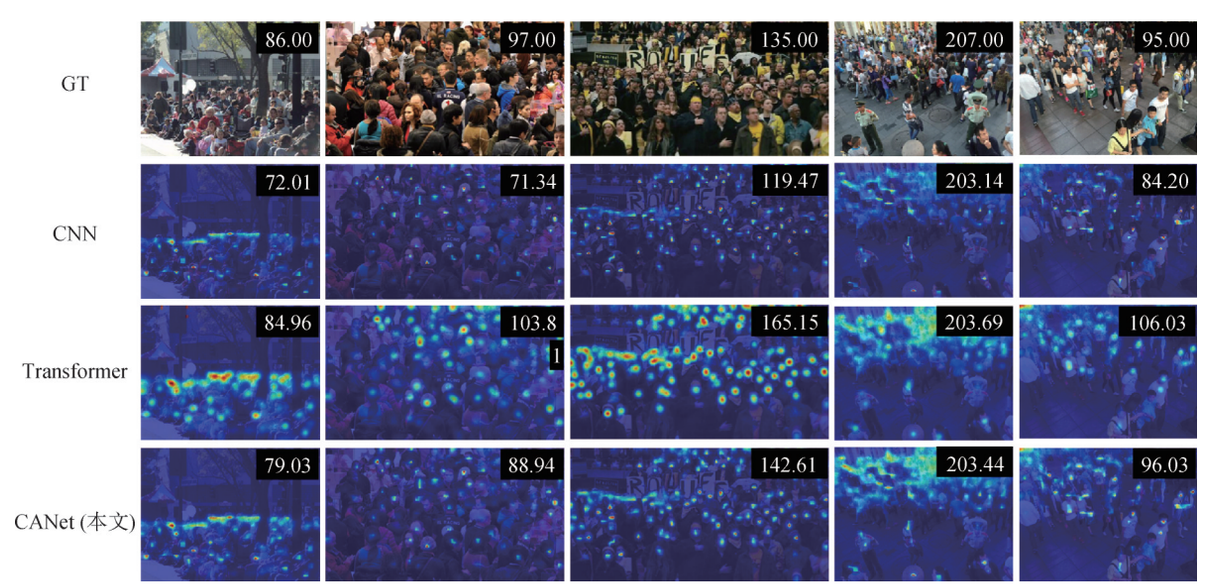


图 5　在 ShanghaiTech 人群数据集上的可视化结果

Fig. 5　Visualization results of ShanghaiTech crowd dataset

3.3 消融实验

3.3.1 关于图像文本类别对齐损失的消融实验

为增强模型对目标类别的感知能力，本文提出图像—文本类别对齐损失，通过引入偏移文本提示构建正负样本对。为了验证对齐损失的有效性，本文构建了未引入该损失的变体CANet w/o L_{align}。如表4所示，移除对齐损失后，CANet在验证集和测试集上的MAE分别上升1.28和0.92。这一结果表明，该损失通过对比学习机制强化了模型对目标语义的跨模态对齐能力，从而提高计数性能。

图6展示了参数变化对模型性能的影响，纵坐标表示FSC-147数据集的MAE指标。如图6(a)所示，当对齐损失的权重$\mu = 1 \times 10^{-1}$时，模型在验证集和测试集上取得最佳性能。过大的μ会导致对齐损失主导了优化过程，使得针对计数的优化过程受到抑制；而过小的μ会削弱对齐效果，导致语义错位问题无法得到解决。如图6(b)所示，当偏移文本数量k为5时，CANet在FSC-147验证集与测试集上取得最优性能。当k过小时，对比样本过少导致模型难以学习不同样本的差异，进而难以区分与目标类别语义相似的物体；当k过大时，对比学习过多强调了类别之间差异，使得计数的学习优化受到抑制。当$k = 5$时，CANet在识别能力和计数能力之间达到平衡。这些实验验证了偏移文本提示在类别对齐中通过提供干扰信息迫使模型学习区分不同类别的关键作用，能够缓解现有方法存在的类别语义错位问题。

表4 关于图像文本对齐损失的消融实验

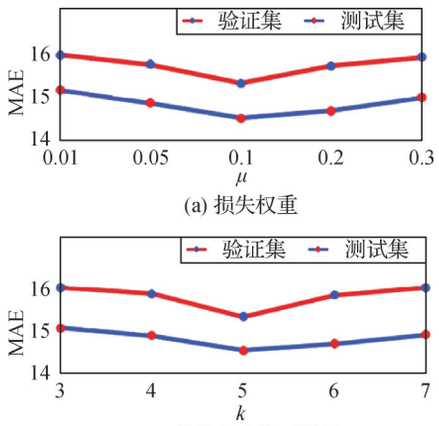
Table 4 Ablation study on the vision-text alignment loss

方法	验证集		测试集	
	MAE	RMSE	MAE	RMSE
CANet w/o L_{align}	16.57	60.31	15.43	102.24
CANet	**15.29**	**57.72**	**14.51**	**98.43**

注：加粗字体表示各列最优结果。

3.3.2 关于双分支解码器的消融实验

首先在FSC-147验证集上对采用单分支解码器的CANet进行多个等级的GAME(grid average mean absolute error)性能指标(Li等，2018)评估，测试了基于CNN分支和Transformer分支的变体，分别记为CANet(CNN)和CANet(Trans)。如表5所示，实验结果表明，Transformer分支在全局层面的计数能力更强(G0和G1指标更低)，而CNN分支则在局部区域的计数估计上表现更优(G2和G3指标更低)。这一观察促使本文设计了信息互馈模块实现两种解码架构优势的有效传递和融合，从而使得CANet在各项GAME指标上均取得最低值，进一步验证了设计的有效性。



(a) 损失权重

(b) 偏移文本提示数量

图6 参数变化对模型性能的影响

Fig. 6 The impact of parameter variations on model performance ((a) the loss weight; (b) the number of shift text prompts)

表5 FSC-147验证集上的GAME指标实验结果

Table 5 GAME results on the validation set of FSC-147

方法	验证集			
	G0	G1	G2	G3
CANet(CNN)	17.83	20.12	22.74	27.43
CANet(Trans)	17.25	19.33	23.32	28.95
CANet	**15.29**	**17.63**	**20.54**	**25.62**

注：加粗字体表示各列最优结果。

表6中提供的单分支解码器的消融实验结果显示，两种变体均明显劣于本文提出的双分支解码器，进一步说明同时捕捉全局与局部信息对提升计数性能至关重要。具体而言，CNN分支侧重于捕捉细节信息，这对于准确识别目标对象至关重要，而Transformer分支则更擅长提取全局上下文信息，有助于整体目标计数的准确性。本文设计的双分支解码器，通过信息互馈模块进一步强化了两种架构的预测能力，使得CANet能够得到兼具细粒度感知和空间一致性的结果。进一步，为了评估互馈模块在双分支解码器中的作用，本文在表6中进行了相关消

融实验。通过去除互馈模块(记为CANet w/o IMF)，发现验证集和测试集上的MAE分别上升了0.86和1.25，这表明互馈模块在Transformer分支和CNN分支之间的信息双向传递中发挥了关键作用。

表7展示了CANet与其他基于文本提示的目标计数模型在参数量和推理速度上的比较结果。结果显示，与CLIP-Count相比，CANet参数量增加了0.3倍，但推理速度仅略微下降(0.9倍)，表明CANet的性能提升并不仅仅依赖于参数量和计算量的增加。具体而言，CANet中解码器的参数量为29.68 MB(3.51%)，而两个编码器的参数量合计为750.93 MB(88.68%)，证明了对解码器结构的优化是合理且有效的。

表6　关于双分支解码器的消融实验

Table 6　Ablation study on the dual-branch decoder

方法	验证集		测试集	
	MAE	RMSE	MAE	RMSE
CANet(CNN)	17.83	61.57	16.92	112.88
CANet(Trans)	17.25	63.71	16.54	105.17
CANet(DC)	16.79	59.42	16.30	107.98
CANet(DT)	16.36	61.33	16.45	103.47
CANet w/o IMF	16.15	60.48	15.76	101.36
CANet	**15.29**	**57.72**	**14.51**	**98.43**

注：加粗字体表示各列最优结果。

表7　参数量和推理速度比较

Table 7　Comparison on parameters and inference speed

方法	参数量/MB	推理速度/(帧/s)
CLIP-Count(Jiang等，2023)	633.56	**23.89**
VLCounter(Kang等，2024)	**577.37**	17.76
CANet(本文)	846.77	21.33

注：加粗字体表示各列最优结果。

此外，本文还对由两个CNN或两个Transformer构成的双分支变体进行了消融实验，分别记为CANet(DC)和CANet(DT)，结果显示这些变体虽较对应的单分支模型有所提升，但效果不及本文提出的双分支解码器，进一步证明了双分支解码器在融合全局与局部特征方面的有效性。

图5展示了双分支解码器内部两个分支预测结果的可视化，结果显示，Transformer分支输出的密度图更侧重于整体区域的连续预测，而CNN分支则更聚焦于目标局部区域的细节，两者的输出特性与理论分析及模型设计初衷高度一致。

3.3.3　关于编码器选取的消融实验

本文默认采用DINOv2与BERT作为图像与文本编码器。为验证模型对编码器选择的适应性，表8对比了CLIP与BLIP作为图像编码器的性能。结果显示，CANet(CLIP)仍优于使用CLIP作为编码器的方法VLCounter(Kang等，2024)，表明模型对编码器选择具备较强的适应性。

3.3.4　关于训练方式的消融实验

本文采用冻结文本编码器并微调图像编码器的训练策略，而CLIP-Count(Jiang等，2023)使用了VPT(visual prompt tuning)这种高效微调策略。为公平对比，表9提供了采用VPT高效微调的实验结果，其参数设置与CLIP-Count一样。结果显示，CANet在更少参数训练下性能未显著下降，验证了本文训练策略的有效性与方法优越性。

表8　关于编码器选取的消融实验

Table 8　Ablation study on the backbone

方法	验证集		测试集	
	MAE	RMSE	MAE	RMSE
CANet(CLIP)	17.75	62.4	16.69	104.51
CANet(BLIP)	17.66	61.51	16.52	103.26
CANet	**15.29**	**57.72**	**14.51**	**98.43**

注：加粗字体表示各列最优结果。

表9　关于训练方式的消融实验

Table 9　Ablation study on the training setting

方法	验证集		测试集	
	MAE	RMSE	MAE	RMSE
CANet-VPT	15.93	59.47	15.26	100.87
CANet	**15.29**	**57.72**	**14.51**	**98.43**

注：加粗字体表示各列最优结果。

3.3.5　鲁棒性验证

本文将鲁棒性定义为模型在高斯噪声干扰下的计数误差波动率，计算式为 $\Delta MAE = \left| MAE_{\text{clean}} - MAE_{\text{noisy}} \right| / MAE_{\text{clean}}$，式中，$MAE_{\text{clean}}$ 表示模型在测试图像不加干扰的MAE性能，MAE_{noisy} 表示模型在测试图像加干扰的MAE性能。本文通过向测试图像添加均

值为0、标准差为0.1的高斯噪声验证CANet的鲁棒性，记为CANet(Noise)。实验结果如表10所示，CANet在测试图像在高斯噪声干扰下的Δ*MAE*为6.80%，依旧保持了良好的计数性能，验证了CANet对输入噪声干扰的鲁棒性。

表10　CANet的鲁棒性验证实验

Table 10　Robustness evaluation about the CANet

方法	验证集		测试集	
	MAE	RMSE	MAE	RMSE
CANet(Noise)	16.33	61.31	15.73	101.36
CANet	**15.29**	**57.72**	**14.51**	**98.43**

注：加粗字体表示各列最优结果。

4　结　论

对于现有的基于文本提示的目标计数方法存在的类别语义错位和解码器架构局限两个挑战，本文提出一种全新的解决方法，通过跨模型对齐和双分支解码器的引入显著提升了模型的预测鲁棒性和精度。针对给定的查询图像，CANet首先生成类别提示，以明确待计数的对象。本文将类别提示与查询图像组成的图像—文本对定义为正样本对，而由其他类别生成的偏移文本提示组成的样本对则视为负样本对。随后，利用多模态视觉语言模型提取类别提示与查询图像的特征，并引入图像—文本类别对齐损失以优化编码特征，解决现有方法存在的类别语义错位的问题，使两种模态的特征进一步对齐，从而提升模型的类别感知能力。为了预测密度图，本文提出双分支解码器，其中CNN分支与Transformer分支分别用于处理互补的局部特征与全局特征。通过多层级的信息互馈模块在双分支之间进行双向信息传递，使两个分支都能够融合全局与局部信息，从而进一步增强单分支解码性能，并最终通过融合双分支的结果得到兼具细粒度感知和空间一致性的预测。实验结果表明，在4个公开数据集上的测试结果验证了本文方法的有效性、优越性和泛化性。未来我们将进一步探索图像—文本类别对齐损失的通用性，通过将数目信息也嵌入到提示词中，探究数目信息在编码阶段对模型计数性能的影响。

参考文献(References)

Amini-Naieni N, Amini-Naieni K, Han T D and Zisserman A. 2023. Open-world text-specified object counting [EB/OL]. [2025-04-03]. https://arxiv.org/pdf/2306.01851.pdf

Cao F, Zhang X W, Li L and Shi M J. 2024. Revisiting multi-scale neural network for crowd counting. System Simulation Technology, 20(2): 180-187 (曹锋, 张孝文, 李莉, 史淼晶. 2024. 重塑多尺度神经网络用于人群计数研究. 系统仿真技术, 20(2): 180-187) [DOI: 10.16812/j.cnki.cn31-1945.2024.02.009]

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A and Zagoruyko S. 2020. End-to-end object detection with transformers//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 213-229 [DOI: 10.1007/978-3-030-58452-8_13]

Cherti M, Beaumont R, Wightman R, Wortsman M, Ilharco G, Gordon C, et al. 2023. Reproducible scaling laws for contrastive language-image learning//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 2818-2829 [DOI: 10.1109/CVPR52729.2023.00276]

Devlin J, Chang M W, Lee K and Toutanova K. 2019. BERT: pre-training of deep bidirectional transformers for language understanding//Proceedings of 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics: 4171-4186 [DOI: 10.18653/v1/N19-1423]

Đukić N, Lukežič A, Zavrtanik V and Kristan M. 2023. A low-shot object counting network with iterative prototype adaptation//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 18826-18835 [DOI: 10.1109/ICCV51070.2023.01730]

Guo M Y, Yuan L, Yan Z Y, Chen B H, Wang Y W and Ye Q X. 2024. Regressor-Segmenter mutual prompt learning for crowd counting//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 28380-28389 [DOI: 10.1109/CVPR52733.2024.02681]

Hobley M and Prisacariu V. 2022. Learning to count anything: reference-less class-agnostic counting with weak supervision [EB/OL]. [2025-04-03]. https://arxiv.org/pdf2205.10203.pdf

Hsieh M R, Lin Y L and Hsu W H. 2017. Drone-based object counting by spatially regularized regional proposal network//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 4165-4173 [DOI: 10.1109/iccv.2017.446]

Hu J, Shen L and Sun G. 2018. Squeeze-and-excitation networks//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 7132-7141

[DOI: 10.1109/CVPR.2018.00745]

Jiang R X, Liu L B and Chen C W. 2023. CLIP-count: towards text-guided zero-shot object counting//Proceedings of the 31st ACM International Conference on Multimedia. Ottawa, Canada: ACM: 4535-4545 [DOI: 10.1145/3581783.3611789]

Kang S, Moon W J, Kim E and Heo J P. 2024. VLCounter: text-aware visual representation for zero-shot object counting//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 2714-2722 [DOI: 10.1609/aaai.v38i3.28050]

Li J N, Li D X, Xiong C M and Hoi S. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation [EB/OL]. [2025-04-03]. https://arxiv.org/pdf/2201.12086.pdf

Li Y H, Zhang X F and Chen D M. 2018. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1091-1100 [DOI: 10.1109/CVPR.2018.00120]

Liang D, Xu W and Bai X. 2022. An end-to-end transformer model for crowd localization//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 38-54 [DOI: 10.1007/978-3-031-19769-7_3]

Lin W and Chan A B. 2024. A fixed-point approach to unified prompt-based counting//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI: 3468-3476 [DOI: 10.1609/aaai.v38i4.28134]

Liu C, Zhong Y J, Zisserman A and Xie W D. 2023b. CounTR: transformer-based generalised visual counting [EB/OL]. [2025-04-03]. https://arxiv.org/pdf/2208.13721.pdf

Liu C X, Lu H, Cao Z G and Liu T L. 2023a. Point-query quadtree for crowd counting, localization, and more//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 1676-1685 [DOI: 10.1109/ICCV51070.2023.00161]

Liu Y T, Shi M J, Zhao Q J and Wang X F. 2019. Point in, box out: beyond counting persons in crowds//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6462-6471 [DOI: 10.1109/CVPR.2019.00663]

Loshchilov I and Hutter F. 2019. Decoupled weight decay regularization [EB/OL]. [2025-04-03]. https://arxiv.org/pdf/1711.05101.pdf

Lu E, Xie W D and Zisserman A. 2019. Class-agnostic counting//Proceedings of the 14th Asian Conference on Computer Vision. Perth, Australia: Springer: 669-684 [DOI: 10.1007/978-3-030-20893-6_42]

Mundhenk T N, Konjevod G, Sakla W A and Boakye K. 2016. A large contextual dataset for classification, detection and counting of cars with deep learning//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 785-800 [DOI: 10.1007/978-3-319-46487-9_48]

Oquab M, Darcet T, Moutakanni T, Vo H, Szafraniec M, Khalidov V, et al. 2024. DINOv2: learning robust visual features without supervision [EB/OL]. [2025-04-03]. https://arxiv.org/pdf/2304.07193.pdf

Paiss R, Ephrat A, Tov O, Zada S, Mosseri I, Irani M, et al. 2023. Teaching CLIP to count to ten//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 3147-3157 [DOI: 10.1109/ICCV51070.2023.00294]

Pelhan J, Lukežič A, Zavrtanik V and Kristan M. 2024. DAVE-A Detect-and-verify paradigm for low-shot counting//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 23293-23302 [DOI: 10.1109/CVPR52733.2024.02198]

Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. 2021. Learning transferable visual models from natural language supervision [EB/OL]. [2025-04-03]. https://arxiv.org/pdf/2103.00020.pdf

Ranasinghe Y, Nair N G, Bandara W G C and Patel V M. 2024. CrowdDiff: multi-hypothesis crowd density estimation using diffusion models//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 12809-12819 [DOI: 10.1109/CVPR52733.2024.01217]

Ranjan V, Sharma U, Nguyen T and Hoai M. 2021. Learning to count everything//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 3393-3402 [DOI: 10.1109/CVPR46437.2021.00340]

Ranjan V and Nguyen M H. 2023. Exemplar free class agnostic counting//Proceedings of the 16th Asian Conference on Computer Vision. Macao, China: Springer: 71-87 [DOI: 10.1007/978-3-031-26316-3_5]

Schuhmann C, Vencu R, Beaumont R, Kaczmarczyk R, Mullis C, Katta A, et al. 2021. LAION-400M: open dataset of CLIP-filtered 400 million image-text pairs [EB/OL]. [2025-02-02]. https://arxiv.org/pdf/2111.02114.pdf

Shi M, Lu H, Feng C, Liu C X and Cao Z G. 2022. Represent, compare, and learn: a similarity-aware framework for class-agnostic counting//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 9519-9528 [DOI: 10.1109/CVPR52688.2022.00931]

Wang S Y, Feng C B, Liu C X and Jin Y S. 2025. Multivariate and soft blending sample-driven image-text alignment for deepfake detection. Journal of Image and Graphics, 30(5): 1334-1345 (王诗雨, 冯才博, 刘春晓, 金逸胜. 2025. 多元软混合样本驱动的图文对齐人脸伪造检测. 中国图象图形学报, 30(5): 1334-1345) [DOI: 10.11834/jig.240252]

Xiao G, Fang J W, Zhang H, Liu Y, Zhou X F and Xu J. 2025. Celadon cross-modal knowledge graph construction via a visual-language model. Journal of Image and Graphics, 30(5): 1318-1333 (肖刚, 方静雯, 张豪, 刘莹, 周晓峰, 徐俊. 2025. 视觉语

言模型引导的青瓷跨模态知识图谱构建. 中国图象图形学报, 30(5): 1318-1333)[DOI: 10.11834/jig.240234]

Xu J Y, Le H, Nguyen V, Ranjan V and Samaras D. 2023. Zero-shot object counting//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 15548-15557[DOI: 10.1109/CVPR52729.2023.01492]

Zhang X Y, Gao Z G and Feng J W. 2025. Lightweight pedestrian detection algorithm for dense crowd scenes based on YOLOv7-tiny. Software Engineering, 28(1): 46-51(张芯源, 高志刚, 冯建文. 2025. 基于YOLOv7-tiny的轻量级密集人群场景行人检测算法. 软件工程, 28(1): 46-51)[DOI: 10.19644/j.cnki.issn2096-1472.2025.001.009]

Zhang Y Y, Zhou D S, Chen S Q, Gao S H and Ma Y. 2016. Single-image crowd counting via multi-column convolutional neural network//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 589-597[DOI: 10.1109/CVPR.2016.70]

Zou M, Huang H, Du W and Huang J F. 2023. Crowd counting and density estimation based on feature fusion codec. Computer Engineering and Design, 44(7): 2110-2117(邹敏, 黄虹, 杜浸, 黄继风. 2023. 基于特征融合编解码的人群计数和密度估计. 计算机工程与设计, 44(7): 2110-2117)[DOI: 10.16208/j.issn1000-7024.2023.07.025]

作者简介

曹锋,男,正高级工程师,主要研究方向为神经网络与人工智能、模式识别与智能系统、计算机软件与理论、计算机应用技术。E-mail: caofeng@rtom.com.cn

史淼晶,通信作者,男,教授,主要研究方向为计算机视觉、机器学习、多媒体计算、医学与遥感图像处理。E-mail:mshi@tongji.edu.cn

张孝文,男,硕士研究生,主要研究方向为计算机视觉、自然图像处理和目标计数。E-mail:2331860@tongji.edu.cn

岳子杰,男,助理教授,主要研究方向为计算机视觉和医学图像处理。E-mail:23310203@tongji.edu.cn

李莉,女,教授,主要研究方向为智能生产、复杂制造系统计划与调度、计算智能和面向工业4.0的大数据应用。E-mail:lili@tongji.edu.cn