arXiv:2602.01540v1 [cs.CV] 2 Feb 2026

FSCA-Net：面向稳健多数据集训练的特征分离交叉注意力网络

Yuehai Chen、Jing Yang、Jian Lan、Yi Si、Shaoyi Du、Badong Chen

摘要——人群计数在公共安全、交通管理和智慧城市管理中发挥着至关重要的作用。然而，尽管基于卷积神经网络和Transformer的模型取得了显著进展，但由于严重的域差异，这些模型在不同环境中应用时，性能往往会下降。直观上看，在多个数据集上直接进行联合训练应能提升泛化能力，但实际上却会导致负迁移，因为共享表征与域特定表征会相互纠缠。为应对这一挑战，我们提出特征分离与交叉注意力网络（FSCA-Net），这是一种统一框架，可将特征表征明确解耦为域不变组件和域特定组件。一种新颖的交叉注意力融合模块可自适应地建模这些组件之间的交互，在保留数据集特定判别能力的同时，确保有效的知识迁移。此外，我们引入互信息优化目标，以最大限度地提升域不变特征之间的一致性，并尽量减少领域特定特征之间的冗余，从而促进形成互补的共享—私有表征。在多个群体计数基准上的大量实验表明，FSCA-Net能有效缓解负迁移，并实现先进的跨数据集泛化能力，为真实世界的人群分析提供稳健且可扩展的解决方案。

索引词—人群计数、跨域泛化、特征分离、交叉注意力、互信息正则化。

I. 引言

Rowd群计数旨在估计图像中人群的数量和空间分布，已成为计算机视觉领域的一项基础任务，在智能视频监控、城市规划和公共安全管理等方面有着广泛应用。准确的人群计数不仅能够实现对大型公共集会的实时监控，还能为交通管理、紧急疏散和智慧城市建设提供重要支持 [1], [2]。随着深度学习的快速发展，卷积神经网络（CNN）和视觉Transformer（ViT）在提取用于人群密度估计的判别性特征方面取得了显著成果，大幅提升了单个数据集上的最先进水平 [3], [4]。

Yuehai Chen、JingYang 和 Yi Shi 就职于中国西安 710049 的西安交通大学自动化科学与工程学院。电子邮箱：cyh0518@stu.xjtu.edu.cn、jasmine1976@xjtu.edu.cn 和 shiyi@mail.xjtu.edu.cn。

Jian Lan 任职于中国陕西省西安市 710049 西安交通大学电子与信息工程学院低空管理研究院（电子邮箱：lanjian@mail.xjtu.edu.cn）。

Badong Chen 和 Shaoyi Du 任职于中国陕西省西安市西安交通大学人工智能与机器人研究所，邮编 710049。电子邮箱：chenbd@mail.xjtu.edu.cn、dushaoyi@gmail.com

通讯作者：Jing Yang 和 Jian Lan。电子邮箱：jasmine1976@xjtu.edu.cn 和 lanjian@mail.xjtu.edu.cn。

表 I 使用 BL [6] 基线比较数据集 A 和 B 上单数据集训练与联合训练的性能。指标以平均绝对误差/均方误差表示。

训练设置	A（平均绝对误差/均方误差）	B（平均绝对误差/均方误差）
仅 A	62.8/101.8	15.9/25.8
仅 B	138.1/228.1	7.7/12.7
A+B	65.3/107.9	10.9/16.8

尽管取得了这些成果，真实场景中的人群计数通常仍需应对单个数据集之外的多样化环境 [5]。因此，有必要在多个数据集上进行联合训练，这主要有两个原因。首先，利用规模更大、更多样化的训练样本集合，有望提升泛化能力，使模型更好地适应实际部署，因为人群场景在密度、视角和背景方面差异显著。其次，从头构建规模足够大且质量上乘的数据集成本高昂，因为密度图标注需要投入大量人力，且耗时漫长。因此，整合现有数据集为构建可扩展且稳健的人群计数模型提供了一条务实且经济高效的途径。

然而，当模型在异构数据集上进行联合训练，或部署于未见域时，便会面临一项重大挑战。尽管基于卷积神经网络和视觉Transformer的模型在各自的数据集上表现出色，但由于存在显著的分布差异，它们在跨数据集场景中的泛化能力会严重下降 [7], [8]。如表I所示，在多个数据集上进行联合训练不仅无法带来性能提升，还会导致负迁移，使模型在源域和目标域上的准确率均有所下降。这一现象凸显了数据集之间固有的冲突，并引出了一个根本性问题：我们该如何设计模型，在避免数据集特定偏差干扰的同时，提取普遍可迁移的特征?

进一步调查发现，跨域性能不佳的根本原因之一，在于学习到的特征空间中纠缠着数据集特定特征。在传统的人群计数网络中，低层和高层特征是联合提取的，这不可避免地将域不变的人群表征（例如密度模式、头部形状）与域特定伪影（例如背景纹理、光照条件、相机分辨率）耦合在一起。如图1的 t-SNE 可视化所示，来自不同数据集的特征高度纠缠、相互重叠，这表明数据集不变信息与数据集特定

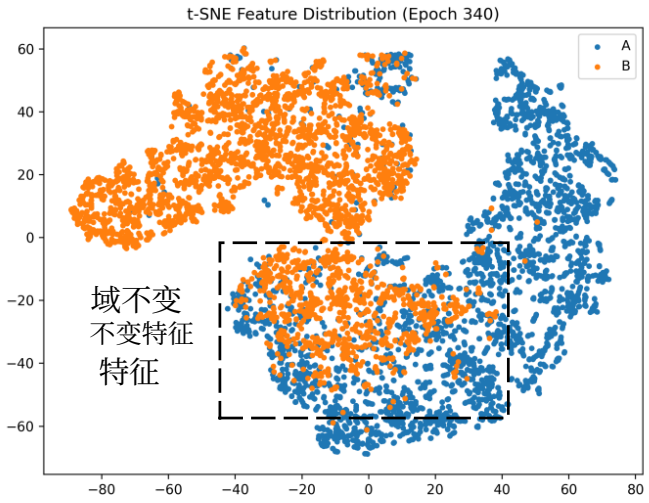


图 1. 使用基线模型从多个数据集中提取的特征的 t-SNE 可视化。来自不同数据集的特征高度纠缠，表明域不变人群表征（例如密度模式、头部形状）与域特定伪影（例如背景纹理、光照、摄像机视角）相互混杂。这种纠缠导致跨数据集泛化能力较差，从而促使我们提出特征分离策略。

信息无法有效解耦。当模型部署到新领域时，这些领域特定线索会变得不可靠或产生误导，最终导致性能显著下降。

为缓解这一问题，研究者已探索了多种策略。对抗式领域对齐方法试图通过学习不同域之间不可区分的特征分布来弥合数据集差异，但此类方法往往忽视共享知识与私有知识之间微妙的平衡，导致过度对齐并丢失领域特定线索。基于元学习的适应增强了对未见域的快速泛化能力，但其对情景式训练的依赖以及有限的适应能力，限制了其在复杂人群场景中的有效性。因此，现有方法仍不足以完全解决域鲁棒性与数据集专门化之间的冲突。

本文提出特征分离交叉注意力人群计数网络（FSCA-Net），这是一种专为多数据集人群计数设计的统一框架。我们将特征空间显式解耦为领域不变组件和领域特定组件，使模型能够捕获可迁移的人群模式，同时隔离数据集特定变化。新颖的交叉注意力融合模块促进这些组件之间的有效交互，从而在不牺牲判别能力的前提下确保适应性。此外，我们采用互信息优化，最大限度地提高领域不变特征之间的一致性，同时尽可能减少领域特定特征之间的冗余，从而强化共享表征与私有表征的互补作用。通过这些设计，FSCA-Net 能够实现卓越的跨数据集泛化，同时保持有竞争力的域内准确率。本文的主要贡献总结如下：

- **解耦领域表征学习**：我们提出一种特征分离策略，将骨干网络表征显式分解为领域不变与领域特定子空间。在正交性和对抗约束的引导下，该设计减少了冗余编码，并隔离数据集特定伪影，从而提升所学特征在不同域中的鲁棒性与可迁移性。

- **基于交叉注意力的自适应融合**：为平衡泛化与专门化，我们引入了一种新颖的交叉注意力融合模块。该模块将域不变特征作为查询，将域特定特征作为键和值，动态捕获共享与私有表征之间的条件依赖关系，从而实现自适应且可解释的跨域交互。

- **面向互补学习的信息论正则化：**我们设计了一种互信息优化目标，在最大化域不变特征间共享信息的同时，最小化领域特定特征之间的冗余。这促使学习到的子空间彼此互补且不重叠，从而确保模型在已见域和未见域中均能保持一致的泛化性能。

II. 相关工作

A. 人群计数

人群计数通常通过神经网络生成密度图来实现，对密度图进行积分即可得到图像中的总人数 [3], [4], [9]。为提高预测准确性，大量研究致力于开发复杂的网络架构，以利用多尺度和上下文特征信息。例如，MCNN 使用三个独立分支，从不同视角估算人群数量 [10], ，而 Switching-CNN 则在此基础上引入切换分类器，动态选择最适合进行密度估计的分支 [11]。较新的模型，如 M-SFANet [12] 和 SASNet [13],，提出了用于编码多尺度上下文信息并缓解尺度变化的新机制。MAN[14] 将标准Transformer的全局注意力、可学习局部注意力和实例级注意力整合到计数框架中，从而提升了性能。类似地，FFDBC [15] 重新利用点标注，进一步提高模型准确性。尽管这些方法展示了先进网络设计在提升计数性能方面的有效性，但它们通常假设训练数据和测试数据来自同一分布。在现实应用中，这一假设往往不成立，因为显著的域偏移可能导致性能大幅下降。

与此同时，另一类互补研究试图通过更直接地利用点标注，绕过密度图的显式构建。例如，BL [6] 在真实标注与由估计密度图得到的聚合点预测之间引入了逐点损失。

NoiseCC [16] 将标注噪声明确建模为高斯随机变量，并通过概率密度近似将其纳入损失函数。DM-Count [17]将计数表述为带有一个 L_2 成本的平衡最优传输问题，而包括 GL [18] 和 UOT [19]，在内的扩展方法则利用非平衡最优传输放宽了这一假设。ChfL [20], 则不同，它将密度图转换到频域，并利用特征函数改进表示，从而取得更优结果。尽管这些基于损失函数的方法已经实现了显著的准确率，但它们同样依赖于训练数据与测试数据之间分布一致性的假设。因此，当应用于未见域时，其性能会显著下降。这凸显了跨域评估的重要性，将其作为开发在真实场景中具有鲁棒性和泛化能力的人群计数模型的关键步骤。

B. 域适应

域适应已成为解决人群计数中域偏移问题的一种广泛采用的策略。当训练数据的分布与测试数据的分布存在显著差异时，就会产生域偏移，导致模型性能下降。为缓解这一问题，研究人员开发了多种域适应方法。例如， [21] 采用基于GAN的框架，将在合成数据集上训练的模型迁移到真实世界图像上。然而，为避免负迁移效应，该方法需要预先了解目标域的人群密度分布，以便仔细匹配合成训练样本 [22]。类似地， [23] 引入神经元变换技术，以显式地建模域偏移。

其他方法则采用不同的策略。例如， [24]利用基于动量的更新机制学习跨域映射，在零样本跨域人群计数中取得了最先进的性能。与此同时， [22] 提出利用目标域中的一张或多张未标注图像来适应模型。尽管这些方法展现出良好的效果，但通常都依赖于获取目标域数据（无论是否带有标注）；而在真实世界的人群计数应用中，目标场景高度动态且持续变化，因此获取此类数据往往成本高昂且不切实际。

C. 域泛化

与域适应相比，域泛化旨在开发能够在多个未见域上实现良好泛化的模型，而无需获取目标域数据。这一特性使其尤其适用于人群计数，因为测试数据分布往往会因场景不同而显著变化。现有的域泛化方法大致可分为两类：数据多样化和域对齐。

数据多样化方法着重于提高源域的变化性，以增强模型的鲁棒性。例如， [25] 提出了一种对抗训练策略来扩充训练数据，而 [26]采用条件生成对抗网络来合成来自伪新域的样本。另一方面，域对齐方法旨在尽量减小源域特征之间的差异，从而学习域不变表示。例如， [27]和 [28] 提出对齐不同域的特征分布，后者利用深度跨模态哈希框架生成紧凑的哈希码，从而有效捕捉不同模态之间的内在关系。

尽管这些方法展现出良好的效果，但通常只单独关注域适应或域泛化。相比之下，我们提出了一种用于跨场景人群计数的新型框架，融合了这两种范式的优势。据我们所知，这是首次设计统一方法，同时利用域泛化和域适应来应对跨场景人群计数的挑战，从而提升其在现实应用中的鲁棒性和适应性。

D. 深度学习中的互信息

互信息（MI）是信息论中的一个基本概念，用于量化两个随机变量之间的统计依赖性。 [29] 首次将互信息相关优化引入深度学习，此后大量研究表明，最大化互信息能够有效增强特征表示学习 [30]–[32]。然而，直接在高维特征空间中估计互信息在计算上难以实现。为应对这一挑战，许多研究提出了变分界来近似真实互信息 [33], [34]。

在这项工作中，我们利用基于互信息的正则化来引导特征分离并提升域鲁棒性。具体而言，我们设计了两个互补的互信息目标：一个下界和一个上界。互信息下界用于增强不同域之间域不变特征的相关性，从而提升特征一致性和适应能力。相比之下，互信息上界则作为对域特定特征的惩罚，减少冗余并限制其在不同域之间的相关性，进而通过防止虚假对齐促进更好的泛化。通过结合这两项约束，所提出的框架能够同时增强域不变表征学习并抑制域特定特征的纠缠。此外，我们的设计为域泛化与域适应搭建了一座天然的桥梁：若有未标注的目标域样本可用，只需将互信息上界替换为互信息下界，便可将泛化策略无缝转换为适应方法，从而显式地使特征与目标分布对齐。

III. 所提方法

为应对数据集之间的遮挡、尺度变化和分布偏移等挑战，我们设计了一个特征分离框架，作为所提出方法的基础。具体而言，该框架将表征显式解耦为两个互补的组成部分：域不变（DI）特征和域特定（DS）特征。

DI 特征捕获可在不同数据集之间泛化的语义和结构信息，而 DS 特征则编码数据集特定特征，例如背景纹理、光照条件和透视畸变。通过联合利用这两个组成部分，该框架在数据集之间保留了可迁移知识，同时有选择地保留了独特的域属性。这种设计有效缓解了分布冲突，提高了学习到的表征的适应性，并增强了跨域场景下人群计数的鲁棒性。

A. 总体框架

本文提出的特征分离交叉注意力人群计数网络（FSCA-Net）如图2所示，是一种围绕特征分离、交叉注意力交互和互信息优化设计的多数据集框架。对于来自数据集A和其他数据集的输入，骨干网络首先提取基础特征，并将其明确解耦为域不变（DI）特征和域特定（DS）特征。同时，引入交叉注意力机制，以增强数据集间相关性并促进信息交换。随后，将域不变特征和域特定特征进一步融合为域特定-域不变（DS-DI）特征，并传递给解码器进行细化，最终生成计数密度图。

训练期间，采用互信息最大化来提升域不变特征在不同数据集间的泛化能力，同时通过互信息最小化抑制域特定特征与非目标域信息之间的冗余关联。此外，采用域特定训练策略，将域不变特征与目标域的域特定特征分支融合，并使用目标域样本进行监督。通过这一联合设计，FSCA-Net 既实现了稳健的跨域泛化，也提升了域内性能。

B. 特征分离与融合

如图 2所示，特征分离与融合构成了所提出的特征分离交叉注意力人群计数网络（FSCA-Net）的核心组件，并遵循先分离后融合范式。在这一设计中，特征融合完全通过**交叉注意力机制**实现：该机制首先解耦域不变信息与域特定信息，然后通过交叉注意力交互增强数据集间相关性。该框架为平衡跨域泛化与域内计数准确率奠定了基础。整个过程围绕多源图像特征（数据集A和其他数据集）展开，分为两个连续阶段：特征分离和特征融合。

1）特征分离：解耦域不变特征与域特定特征：特征分离是 FSCA-Net 实现结构化信息表示的关键前提。通过一个**特征分离模块**（记作 $\mathcal{S}$），将骨干网络从多源图像中提取的基础特征分解为两个功能互补且冗余度低的子集。这种解耦减少了共享信息与数据集特定信息之间的干扰，从而使后续融合更加有效。

2）特征分离的数学表述：根据图 2中的框架设计，特征分离过程可形式化表示为：

$$\mathbf{F}_{\text{DI}}, \mathbf{F}_{\text{DS}} = \mathcal{S}(\mathbf{F}_{\text{base}}) \tag{1}$$

其中，$\mathbf{F}_{\text{base}}$ 表示由骨干网络提取的基础特征，涵盖来自数据集A和其他数据集的纹理、空间结构等低级视觉线索。$\mathbf{F}_{\text{DI}}$表示**域不变（DI）特征**，而$\mathbf{F}_{\text{DS}}$表示**域特定（DS）特征**。这两个子集在表征学习中发挥着不同但互补的作用：

- $\mathbf{F}_{\text{DI}}$：编码跨数据集共享的域无关可迁移知识，例如通用的人体轮廓模式以及密集与稀疏人群区域的空间分布。这些特征为网络的跨域泛化奠定了基础。

- $\mathbf{F}_{\text{DS}}$：编码数据集特定属性，例如依赖场景的背景纹理（如室内场所与室外街道）或尺度分布特征（如小尺度密集人群与大尺度稀疏人群）。这些特征增强了模型适应特定数据集的能力，从而提高域内计数准确率。

3）特征融合。通过交叉注意力机制强化多源特征相关性：特征融合是特征分离之后的关键阶段，旨在通过**交叉注意力交互**强化多源特征相关性。如图2所示，交叉注意力机制是实现特征融合的唯一方法。该机制通过建立 $\mathbf{F}_{\text{DI}}$ 与$\mathbf{F}_{\text{DS}}$ 之间的交互关联，整合来自数据集A和其他数据集的域不变信息与域特定信息，从而生成兼具跨域适应性和域内准确率的融合特征。这些融合表征为解码器提供高质量输入，最终有助于准确估计计数密度图。

基于图2中的框架，交叉注意力机制（记作 $\mathcal{A}_{\text{cross}}$）的运作方式如下：首先，将数据集A的特征指定为**查询（Q）**，而将其他数据集的特征作为**键（K）**和**值（V）**。重要的是，查询和键来自同一种特征类型（即$\mathbf{Q}_{\text{DI}} = \mathbf{F}_{\text{DI-A}}$，$\mathbf{K}_{\text{DI}} = \mathbf{F}_{\text{DI-O}}$，或$\mathbf{Q}_{\text{DS}} = \mathbf{F}_{\text{DS-A}}$，$\mathbf{K}_{\text{DS}} = \mathbf{F}_{\text{DS-O}}$），其中下标“A”表示数据集A，“O”表示其他数据集。这种设计确保了交互能够捕捉同一语义类别特征中的跨域相关性。

交叉注意力输出的计算方式如下：

$$\mathbf{F}_{\text{attn-X}} = \mathcal{A}_{\text{cross}}(\mathbf{Q}_{\text{X}}, \mathbf{K}_{\text{X}}, \mathbf{V}_{\text{X}}) = \text{Softmax}\left(\frac{\mathbf{Q}_{\text{X}}\mathbf{K}_{\text{X}}^{T}}{\sqrt{d_k}}\right)\mathbf{V}_{\text{X}} \tag{2}$$

其中，X 表示特征类型（DI 或 DS），d_k 表示关键特征的维度，用于缩放点积以维持数值稳定性。Softmax

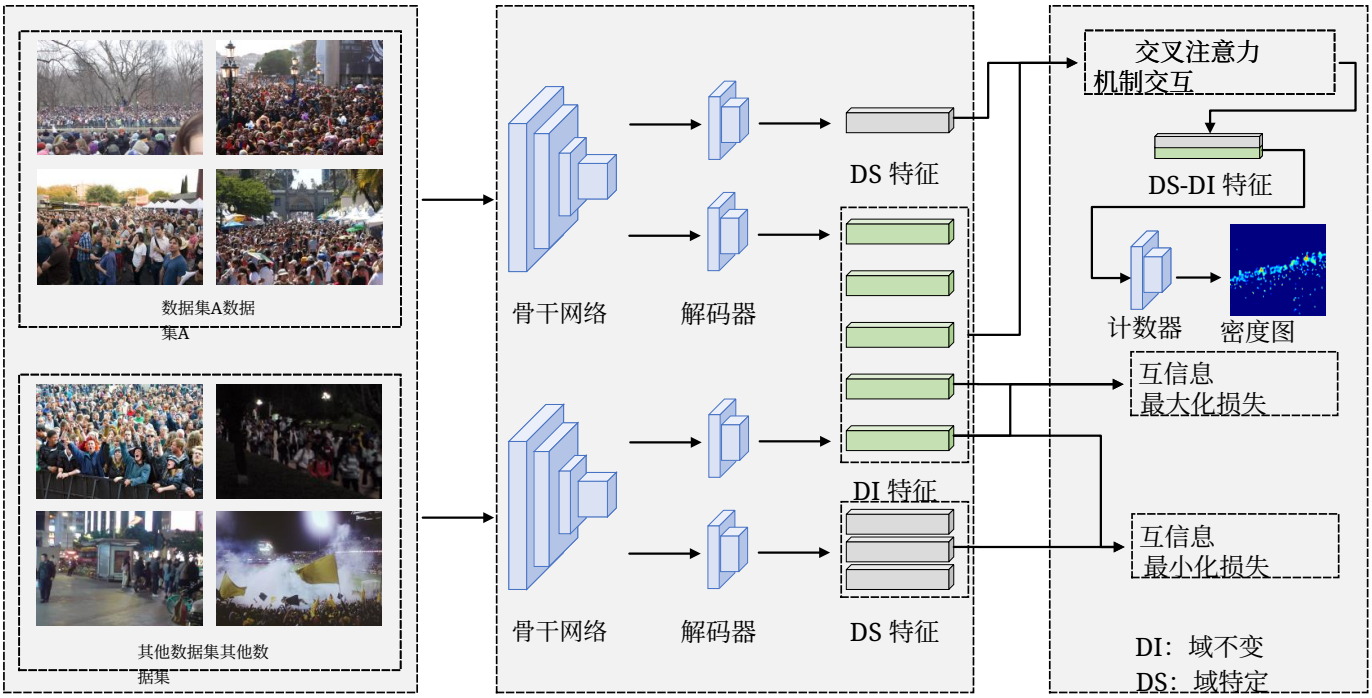


图2. FSCA-Net 概览。骨干网络从多个数据集中提取基础特征，并将其分解为域不变（DI）特征和域特定（DS）特征。交叉注意力机制增强了数据集间交互，随后将 DI 特征和 DS 特征进一步融合为 DS-DI 特征。解码器对所有特征进行细化，以生成 crowds密度图。训练通过互信息优化和域特定策略进行引导，从而兼顾跨域泛化和域内性能。

函数生成注意力权重，突出显示具有较强数据集间相关性的区域（例如，具有相似人群分布模式的区域）。

随后，通过残差连接将交叉注意力输出与同类型的原始特征进行融合，以获得最终的融合特征：

$$\mathbf{F}_{\text{fusion-X}} = \mathbf{F}_{\text{X-A}} + \mathbf{F}_{\text{attn-X}} \tag{3}$$

该过程生成融合的域不变特征 $\mathbf{F}_{\text{fusion-DI}}$ 和融合的域特定特征 $\mathbf{F}_{\text{fusion-DS}}$，从而完成融合过程。

如图 2 所示，交叉注意力机制发挥了两个重要作用：

1）对于 $\mathbf{F}_{\text{DI}}$：增强域不变信息在不同数据集之间的一致性，提高域不变特征的泛化能力。2）对于 $\mathbf{F}_{\text{DS}}$：减轻域偏移引起的变化（例如人群规模分布的差异）所造成的干扰，同时保留数据集特定细节，确保专门化与适应性之间的平衡。

这些融合特征共同为解码器提供了稳健支持，从而实现准确且具有良好泛化能力的人群计数。

C. 互信息最小化与最大化损失

为优化$\mathbf{F}_{\text{DI}}$ （域不变特征）和$\mathbf{F}_{\text{DS}}$ （域特定特征）的表征能力，图2 引入了两个基于互信息（MI）的互补损失函数。具体而言，这些损失引导学习过程，使$\mathbf{F}_{\text{DI}}$ 获得更强的跨域泛化能力，同时$\mathbf{F}_{\text{DS}}$ 保持更高的域特异性。

1）互信息最大化损失（用于$\mathbf{F}_{DI}$ 优化）：该损失专门用于优化$\mathbf{F}_{\text{DI}}$，旨在增强跨域一致性，并在目标域数据可用时促进域适应。互信息（MI）衡量两个变量之间的统计依赖关系，此处为从源域和目标域提取的$\mathbf{F}_{\text{DI}}$ ，通过量化一个变量为另一个变量提供多少信息来实现。

设 z 表示来自源域的$\mathbf{F}_{\text{DI}}$ 的潜在表征（例如 $\mathcal{D}_A$，数据集A），z^+ 表示来自目标域的相应潜在表征（例如 $\mathcal{D}_O$的一个子集，其他数据集）。 z 和 z^+ 均通过将$\mathbf{F}_{\text{DI}}$ 传入线性嵌入获得，从而确保特征空间中的一致性。 z与 z^+ 之间的互信息定义为：

$$I(z; z^+) = \mathbb{E}_{p(z,z^+)}\left[\log \frac{p(z^+ \mid z)}{p(z^+)}\right], \tag{4}$$

其中 $p(z, z^+)$是(z, z^+)的联合分布， $p(z^+ \mid z)$是条件分布，$p(z^+)$是 z^+的边缘分布。最大化 $I(z; z^+)$促使$\mathbf{F}_{\text{DI}}$ 在各域中一致地捕获与人群相关的特征，从而支持有效的知识迁移。

由于概率分布未知，直接估计 $I(z; z^+)$ 难以实现，因此我们采用由相似度函数 $f(\cdot, \cdot)$ 引导的下界近似（例如，

余弦相似度）以实现稳定优化。具体而言，使用以下互信息下界 (I_{LB})：

$$I_{\text{LB}}(z;z^+) = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(f(z_i,z_i^+)\right)}{\sum_{j=1}^{N}\exp\left(f(z_i,z_j^+)\right)}\right], \quad (5)$$

其中，N 为采样对的数量，(z_i，z_i^+)表示源域和目标域中的第 i个潜在表示对。通过将 I_{LB} 作为正则项纳入总体目标，$\mathbf{F}_{\text{DI}}$ 的跨域一致性可通过最大化 I_{LB}得到间接增强。

2）互信息最小化损失（用于$\mathbf{F}_{DS}$ 优化）：该损失专门用于$\mathbf{F}_{\text{DS}}$ ，以抑制冗余的跨域相关性，并在无法获得目标域数据的情况下促进域泛化。其主要目标是确保来自目标域的$\mathbf{F}_{\text{DS}}$ （例如，$\mathcal{D}_A$）仅捕获 $\mathcal{D}_A$所特有的属性，同时避免受到非目标域（例如，$\mathcal{D}_O$）的干扰。

由于条件分布未知，直接最小化互信息难以实现。为此，图2 采用对比对数比上界（*CLUB*）来估计目标域的$\mathbf{F}_{\text{DS}}$ （z）与$\mathbf{F}_{\text{DS}}$ 非目标域的（z^+）之间互信息的上界。形式化地，对于潜向量 $z\sim\mathcal{D}_A$ 和 $z^+\sim\mathcal{D}_O$，CLUB上界定义如下：

$$\begin{aligned} I_{\text{CLUB}}(z;z^+) &= \mathbb{E}_{p(z,z^+)}\Big[\log p(z^+\mid z)\Big] \\ &\quad - \mathbb{E}_{p(z)p(z^+)}\Big[\log p(z^+\mid z)\Big], \end{aligned} \quad (6)$$

其中 $I_{\text{CLUB}}(z;z^+)\geq I(z;z^+)$，且两者之间存在非负间隙 $\Delta_1 = I_{\text{CLUB}} - I$。因此，最小化 I_{CLUB} 可间接降低真实互信息。

由于无法获取条件分布 $p(z^+\mid z)$，因此采用变分分布 $q_\theta(z^+\mid z)$（由 θ 参数化，并通过轻量级多层感知机实现）对其进行近似。对于 N 个配对样本 $\{(z_i,\ z_i^+)\}_{i=1}^N$，$I_{\text{CLUB}}$ 的无偏估计量表示如下：

$$\begin{aligned} \hat{I}_{\text{CLUB}}(z;z^+) = \frac{1}{N}\sum_{i=1}^{N}\Bigg[&\log q_\theta\left(z_i^+\mid z_i\right) \\ &- \frac{1}{N}\sum_{j=1}^{N}\log q_\theta\left(z_j^+\mid z_i\right)\Bigg] \end{aligned} \quad (7)$$

为确保 I_{CLUB} 能够可靠地近似真实上界，必须最小化估计误差

$$\begin{aligned} \Delta_2 &= \text{KLD}\big(p(z^+,z)\parallel q_\theta(z^+,z)\big) \\ &= \mathbb{E}_{p(z,z^+)}\big[\log p(z^+\mid z)\big] - \mathbb{E}_{p(z,z^+)}\big[\log q_\theta(z^+\mid z)\big] \end{aligned} \quad (8)$$

由于只有第二项依赖于q_θ，因此可通过最小化负对数似然损失来实现：

$$L_{\Delta_2} = -\frac{1}{N}\sum_{i=1}^{N}\log q_\theta(z_i^+\mid z_i). \quad (9)$$

在训练过程中，采用交替优化策略：(1) L_{Δ_2} 更新 q_θ ，以更好地近似 $p(z^+\mid z)$，

ˆ同时，(2) I_{CLUB} 引导抑制$\mathbf{F}_{\text{DS}}$ 与$\mathbf{F}_{\text{DI}}$之间的互信息。

总而言之，最小化I_{CLUB}可减少域特定特征之间的冗余，从而促进特征多样性，并提升域泛化下的鲁棒性。

D. 计数网络与最终结果生成

特征分离交叉注意力人群计数网络（FSCA-Net）的计数网络主要由一个计数器构成。其核心功能是将优化后的特征（$\mathbf{F}_{\text{DS-DI}}$）转换为准确的人群计数预测，作为图2所示框架中的终端模块。

1）人群计数计算与域特定训-ˆ练：单张输入图像的人群总数 N 通过对预测密度图的所有像素求和得到：

$$\hat{N} = \sum_{i=1}^{H}\sum_{j=1}^{W}\hat{y}_{i,j}. \quad (10)$$

为增强对数据集特定特征的适应能力，我们采用了域特定训练策略，如图2所示。对于目标域（例如 $\mathcal{D}_A$，数据集A），将跨域泛化特征$\mathbf{F}_{\text{DI}}$ 与目标域分支的域特定特征$\mathbf{F}_{\text{DS}}$ 进行融合。随后，使用来自 $\mathcal{D}_A$的带标签样本，对融合特征进行监督学习。该方法确保网络通过$\mathbf{F}_{\text{DI}}$保留可泛化知识，同时通过$\mathbf{F}_{\text{DS}}$捕捉目标域特定模式，从而实现跨域泛化与域内计数准确率之间的平衡。

我们用于优化计数网络的最终损失函数 L_{count} 为：

$$L_{count} = L_{counter} + \beta_1 I(z;z^+) + \beta_2 L_{\Delta_2} \quad (11)$$

其中 $L_{counter}$ 是主流计数方法的计数损失函数 [6]、[14]、[17]、[35]、I（z；z^+）是互信息最小化损失。最小化 L_{Δ_2} 可以确保ˆ使 I_{CLUB} 能够保持互信息上界。我们通过交叉验证评估了模型在 $\beta_1\in\{0.01,0.1,1\}$和 $\beta_2\in\{0.001,0.01,0.1\}$上的性能。结果表明，$\beta_1=0.1$和 $\beta_2=0.01$在最终损失函数的不同损失之间实现了最佳平衡。

如图 2所示，FSCA-Net 构建了一条闭环优化流程：首先，骨干网络和交叉注意力机制对多源图像进行处理，生成增强特征；随后，这些特征被分离为$\mathbf{F}_{\text{DI}}$ 和 $\mathbf{F}_{\text{DS}}$，并通过基于互信息的损失函数分别进行优化；最后，解码器整合优化后的特征，生成密度图及相应的人群数量。通过联合利用特征分离、交叉注意力交互和互信息优化，FSCA-Net 有效应对了多数据集人群计数中的关键挑战，例如域偏移和泛化能力有限，从而在未见场景中获得更准确、更稳健的人群计数结果。

IV. 实验

本节将介绍实验结果，以评估所提出方法的有效性。首先，我们将详细说明实验设置，包括所用数据集、网络架构、训练流程和评估指标。然后，我们将所提出的方法与近期最先进的方法进行比较。最后，我们将开展消融研究，进一步证明所提出方法的有效性。

A. 实验设置

数据集。实验在多个广泛使用的人群计数数据集上进行，包括上海科技大学A（SHA）[10], 上海科技大学B（SHB）[10], UCF-QNRF（QNRF）[36],和JHU-CROWD++（JHU）[37]。SHA包含从互联网收集的482张高密度图像，而SHB由在繁忙街道上通过监控摄像头拍摄的图像组成，因此其人群场景相比SHA较为稀疏。QNRF是一个大规模数据集，包含多样化的人群场景，且以密集人群为主。JHU是一个综合性数据集，包含4,372张图像和151万条标注。与现有数据集相比，我们选取的数据集涵盖了广泛的场景和环境条件，并提供了更丰富的标注，包括点图、近似边界框和模糊等级。

网络架构与训练细节。我们采用 BL [6], CSRNet [35], DM-Count [17] 和 MAN [14] 作为所提方法的基线，并保持训练设置与 BL [6]中报告的一致，以确保公平比较。我们采用三层卷积回归器作为解码器，

$Conv_{512\times256\times3\times3}$ + $Conv_{256\times256\times1\times1}$ +$Conv_{256\times512\times3\times3}$（输入通道数 × 输出通道数 ×卷积核高度 × 卷积核宽度）。具体而言，我们采用三层卷积回归器作为计数器，

$Conv_{512\times256\times3\times3}$+ $Conv_{256\times128\times3\times3}$ + $Conv_{128\times1\times3\times3}$（输入通道数 ×输出通道数 × 卷积核高度 ×卷积核宽度）。

评估指标。我们使用广泛采用的平均绝对误差（MAE）和均方根误差（MSE）指标评估所提方法的性能，其定义如下：

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{C}_i - C_i|, \quad MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{C}_i - C_i)^2}, \tag{12}$$

ˆ其中，N 表示测试图像的数量，C_i 表示根据预测密度图得到的估计人群数量，C_i 表示相应的真实人数。

B. 实验比较

与基线方法的比较。

我们将所提出的 FSCA-Net 与四种具有代表性的先进人群计数方法结合，以评估其性能：BL [6], CSRNet [35], DM-Count [17], 和 MAN [14]。如表 II所示，对于所有基线模型，同时在多个数据集上训练（记为 (M)）的效果都不如分别在每个数据集上训练（记为 (S)），尽管可用训练样本有所增加。性能下降的原因在于，单个计数模型无法有效适应不同数据集之间的大幅尺度变化和分布差异，因而导致欠拟合和泛化能力较差。

相比之下，采用我们的 FSCA-Net 后，这些基线模型在多数据集训练下均取得了显著提升。例如，FSCA-Net+CSRNET(M) 和 FSCA-Net+BL(M) 不仅优于各自对应的多数据集基线，还超过了单数据集基线。这些结果表明，通过 FSCA-Net 显式地解耦并重新组织特征，能够促进跨数据集的知识迁移，同时增强数据集特定适应能力。

与当前最先进方法的比较。

为验证所提出的特征分离交叉注意力人群计数网络（FSCA-Net）的有效性，我们针对四个具有代表性的人群计数基准——UCF-QNRF、Shanghai A、JHU-Crowd++和 UCF CC 50——开展了大量实验，并与最先进的方法进行了比较。该网络通过显式分离域不变（DI）特征和域特定（DS）特征，并结合交叉注意力交互与互信息优化，解决遮挡、尺度变化和跨数据集分布偏移问题。表III中的定量结果表明，FSCA-Net 在跨域泛化和域内性能方面均表现出优势。值得注意的是，FSCA-Net 在多个数据集上实现了领先的计数准确率：FSCA-Net+CSRNET(M) 在 Shanghai A 上取得了最低的平均绝对误差（46.0）和具有竞争力的均方误差（87.6），同时在 UCF CC 50 上取得了最佳的平均绝对误差（156.9）和均方误差（231.4）；FSCA-Net+MAN(M) 则在 JHU-Crowd++ 上创下新的最佳纪录，取得最低的平均绝对误差（46.0）和均方误差（192.4），此外还在 UCF-QNRF 上取得最佳均方误差（124.5），并在 Shanghai A 上取得最低均方误差（78.2）。

值得注意的是，所有这些性能提升都是通过在多个数据集上联合训练的通用模型实现的，无需针对特定数据集定制模型——这对现实场景部署而言是一项关键优势。同样令人瞩目的是，FSCA-Net 能够作为一种有效的即插即用增强模块，应用于现有基线方法：将其与 CSRNet、BL 和 DM-Count 等传统模型集成后，得到的 FSCA-Net 变体（FSCA-Net+CSRNET(M)、FSCA-Net+BL(M)、FSCA-Net+DM(M)）均能在各数据集上持续降低平均绝对误差和均方误差。例如，FSCA-Net+CSRNET(M) 将原始 CSRNet 在 Shanghai A 上的平均绝对误差从 68.2 降至 46.0 ，并将其在 UCF CC 50 上的平均绝对误差从 266.1 降至156.9；而 FSCA-Net+DM(M) 则将 DM-Count 在 JHU-Crowd++上的平均绝对误差从 66.0 降至 54.7。这些结果共同证实，FSCA-Net 的设计通过利用解耦的域不变/域特定特征来保留可迁移知识并维持域特定属性，有效缓解了跨数据集分布冲突，增强了表征适应性，从而优于现有方法，并验证了其作为人群计数任务通用增强方案的价值。

表II 与基线方法的计数性能比较。(S)表示模型在每个数据集上分别训练。(M)表示模型在多个数据集上联合训练。

方法	UCF-QNRF		Shanghai A		JHU-Crowd++		UCF CC 50	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CSRNET(S) [35]	110.6	190.1	68.6	115.0	85.9	309.2	266.1	397.5
CSRNET(M)	121.5	168.2	70.4	123.1	94.8	282.7	332.6	394.5
FSCA-Net+CSRNET(M)	**87.6**(-27.9%)	**134.4**(-20.1%)	**46.0**(-34.7%)	**87.6**(-28.8%)	**56.2**(-40.7%)	**231.6**(-18.1%)	**156.9**(-52.9%)	**231.4**(-41.3%)
BL(S) [6]	88.7	154.8	68.4	97.8	67.1	268.9	229.3	308.1
BL(M)	92.2	158.6	70.6	99.7	69.4	276.2	235.0	317.3
FSCA-Net+BL(M)	**83.8**(-9.1%)	**144.5**(-8.9%)	**46.1**(-34.7%)	**78.1**(-21.7%)	**58.4**(-15.8%)	**237.1**(-14.2%)	**230.2**(-2.0%)	**274.7**(-13.4%)
DM(S) [17]	85.6	148.3	59.7	95.7	66.0	261.4	211.0	291.5
DM(M)	105.7	167.3	65.2	110.1	62.8	234.2	271.2	403.4
FSCA-Net+DM(M)	**75.1**(-29.0%)	**135.0**(-19.3%)	**49.9**(-23.5%)	**85.3**(-22.6%)	**54.7**(-12.9%)	**229.0**(-2.2%)	**181.7**(-33.0%)	**242.9**(-39.8%)
MAN(S) [14]	77.3	131.5	56.8	90.3	53.4	209.9	-	-
MAN(M)	83.2	142.0	61.8	97.5	57.6	226.6	-	-
FSCA-Net+MAN(M)	**73.5**(-11.7%)	**124.5**(-12.3%)	**49.3**(-20.3%)	**78.2**(-19.8%)	**46.0**(-20.1%)	**192.4**(-15.1%)	-	-

表 III 计数性能与最先进-水平-方法-的比较。 红色表示最佳性能，蓝色表示次佳。更低的值表示更好的性能。

方法	UCF-QNRF		Shanghai A		JHU-Crowd++		UCF CC 50	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
ASNPE [38]	89.1	159.7	57.8	97.1	-	-	174.8	251.6
RPNNet [38]	-	-	61.2	96.9	-	-	-	-
MNA [39]	85.8	150.6	61.9	99.6	67.7	258.5	-	-
ADSCNET [40]	71.3	132.5	55.4	97.7	-	-	198.4	267.3
CSRNet [35]	110.6	190.1	68.2	115.0	85.9	309.2	266.1	397.5
BL [6]	88.7	154.8	68.4	101.8	67.1	268.9	229.3	308.1
M-SFANET [12]	85.6	151.2	59.7	95.7	65.5	265.4	212.3	291.5
DM [17]	85.6	148.3	59.7	95.7	66.0	261.7	161.0	276.8
GL [18]	84.3	147.5	61.3	95.4	-	-	-	-
ChfL [20]	80.3	137.6	57.5	94.3	-	-	-	-
GauNet [41]	81.6	153.7	54.8	89.1	-	-	186.3	256.5
PET [42]	79.5	144.3	47.4	75.0	58.5	238.0	-	-
CrowdHat [43]	75.1	126.7	51.2	81.9	52.3	211.8	-	-
STEERER [44]	74.3	128.3	54.5	86.9	54.3	238.3	-	-
APGCC [45]	80.0	136.6	48.8	76.7	54.3	255.9	-	-
CrowDiff [46]	68.9	125.6	47.4	75.0	47.3	198.9	-	-
P2R [47]	83.3	138.1	51.0	79.7	58.8	253.1	-	-
FSCA-Net+CSRNET(M)	87.6	134.4	46.0	87.6	56.2	231.6	156.9	231.4
FSCA-Net+BL(M)	83.8	144.5	46.1	78.1	58.4	237.1	230.2	274.7
FSCA-Net+DM(M)	75.1	135.0	49.9	85.3	54.7	229.0	181.7	242.9
FSCA-Net+MAN(M)	73.5	124.5	49.3	78.2	46.0	192.4	-	-

C. 消融研究

为评估 FSCA-Net 核心组件的独立贡献及协同效应——包括特征分离模块（FSM）、交叉注意力特征融合模块（CAFFM）、互信息最小化损失（MIML）和互信息最大化损失（MAXL）——我们以 FSCA-Net+BL(M) 变体为基线，在 UCF-QNRF 数据集上开展了系统的消融研究。各项消融设置均遵循“移除组件”的原则，以分离各个模块的影响：移除 FSM 意味着放弃对域不变（DI）和域特定（DS）特征的显式解耦，转而采用原始 BL(M) 模型的特征提取流程；移除 CAFFM 则是以简单拼接取代基于交叉注意力的域不变和域特定特征融合，tion（concat）操作仅聚合特征通道，而不捕获特征间相关性；移除 MIML 或 MAXL 即省略相应的互信息约束，从而分别无法优化特征判别性（针对 MIML）或跨域泛化（针对 MAXL）。

表 IV展示了消融研究的定量结果，评估指标为平均绝对误差和均方误差（值越低，性能越好）。完整的 FSCA-Net+BL(M)模型（启用所有组件，即 FSM✓、CAFFM ✓、MIML ✓、MAXL ✓）取得了最佳性能，平均绝对误差为 83.8，均方误差为 144.5。这证实了所有核心组件的协同效应对于最大限度地提高模型的计数准确率至关重要。

移除 FSM 后（表IV中的第二行），模型性能显著下降：平均绝对误差

增加了 4.9（从 83.8 增至 88.7），均方误差增加了 10.3（从 144.5 增至 154.8）。这种性能下降表明，FSM 通过显式解耦域不变特征（用于捕获跨域可泛化语义信息）和域特定特征（用于编码背景纹理和视角等数据集特定特征），有效缓解了不同域特征之间的分布冲突，而原始模型（不含 FSM）不具备这一能力。

去除 CAFFM（第三行）同样会导致明显的性能下降：与完整模型相比，平均绝对误差增加 1.2（达到 85.0），均方误差增加 3.7（达到 148.2）。这一结果凸显了简单特征拼接的局限性：与利用交叉注意力来建模域不变特征与域特定特征之间的细粒度交互并强调判别性信息的 CAFFM 不同，CONCAT 仅执行通道级聚合，而未优化特征相关性——最终降低了用于密度图回归的融合特征质量。

去除 MIML（第四行）或 MAXL（第五行）同样会损害模型性能。在没有 MIML 的情况下，平均绝对误差升至 86.8，均方误差升至 150.6；在没有 MAXL 的情况下，平均绝对误差达到 87.2，均方误差仍为 148.8。MIML 旨在最小化域特定特征与非目标域信息之间的冗余相关性，确保域特定特征仅编码数据集特定属性；将其移除会导致“噪声”域特定特征，从而干扰跨域适应。相比之下，MAXL 最大化域不变特征在各数据集上的泛化能力，省略它会削弱模型在域之间迁移知识的能力——二者都会导致计数准确率下降。

总体而言，这些消融结果验证了FSCA-Net中每个核心组件的必要性：FSM为有效的特征解耦奠定基础，CAFFM优化了解耦特征的融合，而MIML/MAXL共同提升特征判别性与泛化能力。它们的协同整合是FSCA-Net实现优异性能的关键，任何单一组件都不可被替换或省略，否则会导致性能下降。

表IV 使用FSCA-NET+BL(M)在UCF-QNRF数据集上的消融研究。移除FSM意味着使用原始模型。移除CAFFM意味着在域不变（DI）特征与域特定（DS）特征之间使用简单拼接（CONCAT）操作。

FSM	CAFFM	MIML	MAXL	结果（平均绝对误差/均方误差）
✓	✓	✓	✓	83.8 / 144.5
	✓	✓	✓	88.7 / 154.8
✓		✓	✓	85.0 / 148.2
✓	✓		✓	86.8 / 150.6
✓	✓	✓		87.2 / 148.8

D. 对未见数据集的泛化

为进一步说明 FSCA-Net 如何让计数模型仅利用**域不变（DI）特征**进行跨数据集适配，从而更好地泛化到未见场景，我们进行了跨数据集评估。在这项实验中，FSCA-Net（侧重于 DI 特征提取）及其集成基线模型仅使用 UCF-QNRF 数据集进行训练，未使用额外训练数据，也未在目标数据集上微调参数；随后直接在三个未见过的基准数据集（Shanghai A、JHU-Crowd++和 UCF CC 50）上进行评估。评估的定量结果见表V。

如表所示，FSCA-Net 始终能够提升所有基线模型的跨数据集泛化性能。例如，与 CSRNet 集成时，FSCA-Net 可使 CSRNet 在 Shanghai A、JHU-Crowd++ 和 UCF CC 50 上的 MAE 分别降低 16.9%、18.5% 和 37.0%；即便对于相对稳健的 MAN 基线，FSCA-Net 仍可使其在 Shanghai A 上的 MAE 降低 22.4%，在 UCF CC 50 上降低 24.7%。这些改进证实，FSCA-Net 提取的 DI 特征能够捕捉可跨域泛化的语义和结构信息，同时排除背景纹理或光照变化等数据集特定噪声，从而有效缓解源数据集（UCF-QNRF）与未见目标数据集之间的分布偏移。

值得注意的是，FSCA-Net 仅依赖 DI 特征进行泛化，因此无需在训练过程中进行域特定特征适应或引入目标域数据，使其成为一种轻量且实用的跨数据集人群计数解决方案。与采用混合特征表示（结合域特定信息和不变信息）的基线模型相比，FSCA-Net 对 DI 特征的关注确保了其在未见场景中的性能更加稳定且具有更强的可迁移性，最终在所有对比方法中取得了最佳的整体跨数据集评估结果。

表 V 未见数据集上的泛化能力。模型在 UCF-QNRF 上训练，并在其他数据集上进行测试。改进幅度计算为相对于基线方法的百分比降低。

UCF-QNRF →	Shanghai A		JHU-Crowd++		UCF CC 50	
方法	MAE	MSE	MAE	MSE	MAE	MSE
L2SM [48]	73.4	119.4	-	-	-	-
S-DCNET [49]	61.8	102.8	-	-	-	-
CSRNET [35]	75.3	138.7	91.4	317.0	389.8	659.6
FSCA-Net+CSRNET	62.6	100.5	74.5	268.7	245.6	398.2
提升（%）	16.9	27.5	18.5	15.2	37.0	39.6
BL [6]	69.8	123.8	81.2	303.8	309.6	537.1
FSCA-Net+BL	56.0	91.4	70.8	264.0	226.7	328.5
提升（%）	19.8	26.2	12.8	13.1	26.8	38.8
DM [17]	69.3	120.6	85.2	303.4	317.8	550.2
FSCA-Net+DM	55.0	92.5	73.9	266.2	243.5	360.1
提升（%）	20.6	23.3	13.3	12.3	23.4	34.6
MAN [14]	67.1	122.1	79.0	282.1	295.3	512.7
FSCA-Net+MAN	52.1	90.6	71.0	252.9	222.4	337.2
提升 (%)	22.4	25.8	10.1	10.3	24.7	34.2

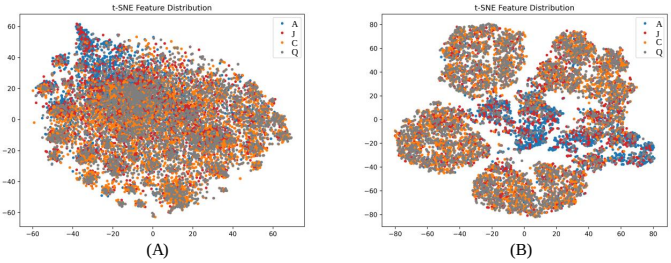
E. 复杂度分析

为评估我们方法的复杂度，我们在表VI中展示了在 ShanghaiTech PartA 数据集上的参数量、浮点运算次数、推理速度和计数结果对比。所有实验均在相同环境下进行：



(A) 输入图像 (B) 真实值 (C) BL(M) (D) FSCA-Net+BL

图3. 人群计数和密度预测结果的可视化对比。(A) 输入图像；(B) 真实值；(C) BL (M)；(D) FSCA-Net+BL。



(A) (B)

图4. 特征分布的 t-SNE 可视化：(A) 联合训练的原始方法提取的特征（无法区分域不变特征和特定领域特征）；(B) 所提出的 FSCA-Net 提取的特征。

F. 可视化

为证明所提出的 FSCA-Net 能够有效提取域不变特征，我们分析了联合训练的原始方法（图4 (A)）和 FSCA-Net（图4 (B)）的 t-SNE 特征分布可视化结果。

在图 4 （A）中，原始方法未能将域不变特征与域特定相关特征分离，不同颜色的点（对应不同域或特征类型）高度混杂并聚集在一起，尤其在中心区域形成密集的聚集。这表明，普通方法学习到的特征相互纠缠，以无结构的方式同时包含域不变信息和域特定信息，使得难以单独区分和利用域不变特征。

参数量和浮点运算次数是在单张 NVIDIA 3090 GPU 上以 512×512 为输入尺寸计算的，推理时间则是测试 1024×768 张图像样本进行 100 次运行的平均值，这与 ShanghaiTech A 基准 [10]一致。如表VI所示，尽管我们的 FSCA-Net+CSRNet 模型参数量略有增加（20.60M，相比 CSRNet 的 16.26 M），推理时间也略长（0.042 s，相比 CSRNet 的 0.038 s），但其在 PartA 人群计数任务中取得了显著更优的 87.6 均方误差。值得注意的是，我们的模型推理速度约为每秒 23.8 张图像（1/0.042 = 23.8），仍能满足实际人群计数场景中的实时应用要求。

相比之下，图 4 （B）中 FSCA-Net 的特征点分布更加有序且分散。不同颜色的特征分离得更好，不像图 4 （A）那样过度集中在中心。可以观察到，域不变特征（本应在不同域之间共享，因此呈现出更一致的分布模式）形成了更清晰、彼此区分更明显的簇，与域特定特征的交织程度也更低。特征点分离度和分散度的提升表明，FSCA-Net 能够有效地将域不变特征与领域特定特征解耦，以更有结构性和判别性的方式提取域不变特征，

表 VI 比较参数量（M）、浮点运算次数（G）、推理速度（秒/ 100 张图像）及PartA 数据集上的结果 [10]。注意：参数量和浮点运算次数以输入尺寸 512×512 计算，运行于单张 NVIDIA 3090 GPU。推理时间为测试 1024×768 样本时运行 100 次的平均时间。

方法	骨干网络	参数量（M）	浮点运算次数（G）	推理速度（秒）	PartA 中的均方误差
PCC-Net [50]	FS	0.51	43.87	0.013	124.0
CAN [51]	VGG16	18.10	114.83	0.047	100.0
SCAR [52]	VGG16	16.29	108.44	0.047	110.0
M-SFANet [12]	VGG16	22.88	115.14	0.058	94.5
SFCN [53]	ResNet101	38.60	162.03	0.096	104.5
CSRNet [35]	VGG16	16.26	108.34	0.038	115.0
FSCA-Net+CSRNet	VGG16	20.60	124.64	0.042	87.6

这对需要跨域泛化的任务至关重要。

为展示所提出的 FSCA-Net 的有效性，我们分析了图中人群计数和密度预测的可视化结果3。在计数准确率方面，通过比较预测计数（PRE）与真实值（GT），FSCA-Net+BL 相较于基线 BL(M) 有显著提升。例如，当 GT 为 1111 时，BL(M) 预测为 1743（严重高估），而 FSCA-Net+BL 预测为 1254，更接近真实值；当 GT 为 = 297 时，BL(M) 预测为 466（明显高估），而 FSCA-Net+BL 预测为 308（几乎与 GT 相符）；当 GT 为 = 1210 时，BL(M) 预测为 1633（严重高估），而 FSCA-Net+BL 预测为 1324，大幅降低了误差。

在密度预测精度方面，BL(M) 的密度热力图（C 列）存在不准确之处，例如出现人为的高密度区域，或未能捕捉真实的密集区域。相比之下，FSCA-Net+BL 的热力图（D 列）与真实值（B 列）高度吻合，能够准确捕捉密集区域的空间分布和强度（黑框标出）。总之，FSCA-Net 不仅通过减少预测误差提高了计数准确率，还通过更准确地捕捉人群空间分布提升了密度图精度，验证了其在人群计数任务中的有效性。

V. 结论

在本研究中，我们解决了跨域人群计数这一关键挑战。传统的基于卷积神经网络和视觉Transformer的模型，由于域不变特征与域特定特征相互纠缠，容易产生负迁移。为克服这一局限，我们提出了特征分离与交叉注意力网络（FSCA-Net），该网络能够显式解耦特征空间，通过交叉注意力增强域间交互，并利用互信息约束抑制无关变化。在多个基准数据集上进行的全面实验表明，FSCA-Net显著提升了泛化能力，同时缓解了负迁移，并在跨域场景下取得了最先进的性能。未来的工作将把这一框架扩展至实时部署，并探索其在其他密集场景理解任务中的适用性，例如车辆计数和人群行为预测。

参考文献

[1] J. Gao、Z. Huang、Y. Lei、H. Shan、J. Z. Wang、F.-Y. Wang和J. Zhang，“用于增强人群计数的深度秩一致金字塔模型”，IEEE 神经网络与学习系统汇刊，第36卷，第 1， 299 312, 2025 1 页–。J. Gao、T. Han、Y. Yuan和Q. Wang，“通过高质量图像转换与密度重建实现域自适应人群计数”，IEEE 神经网络与学习系统汇刊，第1–13页，2021年1–。[3] R. Wang、Y. Hao、L. Hu、X. Li、M. Chen、Y. Miao和I. Humar，“通过双重知识蒸馏实现高效人群计数”，IEEE图像处理汇刊，第 33， 569 583, 2024 1, 2 页–。Y. Chen、Q. Wang、J. Yang、B. Chen、H. Xiong和S. Du，“学习用于人群计数的判别性特征”，IEEE图像处理汇刊，第 33， 3749 3764, 2024 1, 2 页–。Q. Wang、T. Han、J. Gao和Y. Yuan，“神经元线性变换：为人群计数建模域偏移”，IEEE 神经网络与学习系统汇刊，第 1 13, 2021 1 –页。Z. Ma、X. Wei、X. Hong和Y. Gong，“基于点监督的贝叶斯损失用于人群数量估计”，见IEEE/CVF国际计算机视觉会议论文集，2019年，第6142–6151页。1, 2, 6, 7, 8, 9 [7] Y. Luo、J. L¨u、X. Jiang和B. Zhang，“从架构冗余中学习：深度多路径编码器-解码器网络中的增强深度监督”，IEEE 神经网络与学习系统汇刊，第 1 14, 2021 1 –页。X. Chen、X. Chen、Y. Zhang、X. Fu和Z.-J. Zha，“用于复杂场景密集连续值回归的拉普拉斯金字塔神经网络”，IEEE 神经网络与学习系统汇刊，第 32， 11， 5034 5046, 2021 1 卷，第 页–。Y.-K. Hsieh、J.-W. Hsieh、X. Li、Y.-M. Zhang、Y.-C. Tseng和M.-C. Chang，“带标注误差建模的尺度感知人群计数网络”，IEEE图像处理汇刊，第34卷，第2750– 2764, 2025 2 页。Y. Zhang、D. Zhou、S. Chen、S. Gao和Y. Ma，“基于多列卷积神经网络的单图像人群计数”，见IEEE 计算机视觉与模式识别会议（CVPR）论文集，6月 2016 2, 7, 10, 11 。D. Babu Sam、S. Surya和R. Venkatesh Babu，“用于人群计数的切换卷积神经网络”，见IEEE 计算机视觉与模式识别会议（CVPR）论文集，2017年7月。2 [12] P. Thanasutives、K.-i. Fukui、M. Numao和B. Kijsirikul，“基于编码器-解码器并带有多尺度感知模块的卷积神经网络用于人群计数”，见2020年第25届国际模式识别会议（ICPR）。IEEE， 2021， 2382 2389 2, 8, 11 第–页。Q. Song、C. Wang、Y. Wang、Y. Tai、C. Wang、J. Li、J. Wu和J. Ma，“选择还是融合？人群计数中的尺度选择”，见AAAI 人工智能会议论文集，第35卷，第3期，2021年，第 2576 2583 2 –页。H. Lin、Z. Ma、R. Ji、Y. Wang和X. Hong，“通过多面向注意力提升人群计数”，见IEEE/CVF 计算机视觉与模式识别会议论文集，2022年，第19 628–19 637页。[15] 2, 6, 7, 8, 9 Z. Shi、P. Mettes和C. G. Snoek，“基于密度的计数中的免费聚焦”，国际计算机视觉杂志，第 1 18, 2024 2 –页。J. Wan和A. Chan，“为人群计数建模噪声标注”，神经信息处理系统进展，第33卷，2020年。3

[17] B. Wang、H. Liu、D. Samaras和M. H. Nguyen，“用于人群计数的分布匹配”，神经信息处理系统进展，第 33，1595–1607，2020 3, 6, 7, 8, 9 页–。[18] J. Wan、Z. Liu和A. B. Chan，“用于人群计数与定位的广义损失函数”，见2021年 IEEE/CVF 计算机视觉与模式识别会议（CVPR），2021，1974–1983 3, 8 第–页。[19] Z. Ma、X. Wei、X. Hong、H. Lin、Y. Qiu和Y. Gong，“通过非平衡最优传输学习计数”，见AAAI 人工智能会议论文集，第35卷，第3期，2021年，第2319–2327页。3[20] W. Shu、J. Wan、K. C. Tan、S. Kwong和A. B. Chan，“频域中的人群计数”，见IEEE/CVF 计算机视觉与模式识别会议论文集，2022年，第19618–19 627页。[21] 3, 8 Q. Wang、J. Gao、W. Lin和Y. Yuan，“从合成数据中学习以进行真实场景人群计数”，见IEEE/CVF 计算机视觉与模式识别会议论文集，2019，8198–8207 3 第–页。[22] M. K. K. Reddy、M. Rochan、Y. Lu和Y. Wang，“Adacrowd：用于人群计数的无标签场景适应”，IEEE 多媒体汇刊，第 24，1008–1019，2022 3 页–。[23] Q. Wang、T. Han、J. Gao和Y. Yuan，“神经元线性变换：为人群计数建模域偏移”，IEEE 神经网络与学习系统汇刊，第33卷，3250，2021年8，3238–3 第 页–。[24] Q. Wu、J. Wan和A. B. Chan，“用于零样本跨域人群计数的动态动量适应”，见第29届 ACM 多媒体国际会议论文集，2021，658–666 3 第–页。[25] I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville和Y. Bengio，“生成对抗网络”，ACM 通讯，第 63，11，139–144，2020 3 卷，第 页–。[26] J. Gauthier等，“用于卷积人脸生成的条件生成对抗网络”，斯坦福大学 CS231N课程项目：用于视觉识别的卷积神经网络，冬季学期，第 2014，5，2，2014 3 卷，第 页。[27] S. Motiian、M. Piccirilli、D. A. Adjeroh和G. Doretto，“统一的深度监督域适应与泛化”，见IEEE 国际计算机视觉会议论文集，2017年，第5715–5725页。3[28] E. Yang、C. Deng、W. Liu、X. Liu、D. Tao和X. Gao，“成对关系引导的深度哈希用于跨模态检索”，见 AAAI 人工智能会议论文集，第31卷，第1期，2017年。3[29] A. A. Alemi、I. Fischer、J. V. Dillon和K. Murphy，“深度变分信息瓶颈”，见国际学习表征会议，2017 3 。[30] P. Bachman、R. D. Hjelm和W. Buchwalter，“通过跨视图最大化互信息学习表征”，神经信息处理系统进展，第 32，2019 3 卷。[31] M. Tschannen、J. Djolonga、P. Rubenstein、S. Gelly和M. Lucic，“论用于表征学习的互信息最大化”，见第八届国际学习表征会议。OpenReview.net，2020 3 。[32] K. Zhou、H. Wang、W. X. Zhao、Y. Zhu、S. Wang、F. Zhang、Z. Wang和J.-R. Wen，“S3-rec：基于互信息最大化的序列推荐自监督学习”，见第29届 ACM 信息与知识管理国际会议论文集，2020，1893–1902 3 第–页。[33] R. Fritschek、R. F. Schaefer和G. Wunder，“通过神经互信息估计实现信道编码的深度学习”，见2019年 IEEE 第20届无线通信信号处理进展国际研讨会（SPAWC）。IEEE，2019，1–5 3 第–页。[34] P. Cheng、W. Hao、S. Dai、J. Liu、Z. Gan和L. Carin，“CLUB：互信息的对比对数比上界”，见国际机器学习会议。PMLR，2020，1779–1788 3 第–页。[35] Y. Li、X. Zhang和D. Chen，“CSRNet：用于理解高度拥挤场景的空洞卷积神经网络”，见IEEE 计算机视觉与模式识别会议论文集，2018，1091–1100 6, 7, 8, 9, 11 第–页。[36] H. Idrees、M. Tayyab、K. Athrey、D. Zhang、S. Al-Maadeed、N. Rajpoot和M. Shah，“用于密集人群中计数、密度图估计与定位的组合损失”，见欧洲计算机视觉会议（ECCV）论文集，9月 2018 7 。[37] V. A. Sindagi、R. Yasarla和V. M. Patel，“JHU-Crowd：大规模人群计数数据集与基准方法”，技术报告，2020年。7[38] Y. Yang、G. Li、Z. Wu、L. Su、Q. Huang和N. Sebe，“用于透视感知目标计数的反向透视网络”，见IEEE/CVF 计算机视觉与模式识别会议，2020，4374–4383 8 页码 – 。[39] J. Wan 和 A. Chan，《面向人群计数的噪声标注建模》，神经信息处理系统进展，第 33，2020 8 卷。[40] S. Bai、Z. He、Y. Qiao、H. Hu、W. Wu 和 J. Yan，《用于计数的自校正监督自适应空洞网络》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2020，4594–4603 8 页码 – 。[41] Z.-Q. Cheng、Q. Dai、H. Li、J. Song、X. Wu 和 A. G. Hauptmann，《重新审视卷积网络在目标计数中的空间不变性》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2022，19 638–19 648 8 页码 – 。[42] C. Liu、H. Lu、Z. Cao 和 T. Liu，《用于人群计数、定位等任务的点查询四叉树》，载于 IEEE/CVF国际计算机视觉会议论文集，2023，1676–1685 8 页码 – 。[43] S. Wu 和 F. Yang，《利用未充分使用的输出特征提升人群分析中的检测性能》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2023，15 609–15 618 8 页码 – 。[44] T. Han、L. Bai、L. Liu 和 W. Ouyang，《STEERER：通过选择性继承学习解决计数与定位中的尺度变化》，载于 IEEE/CVF国际计算机视觉会议论文集，2023，21 848–21 859 8 页码 – 。[45] I.-H. Chen、W.-T. Chen、Y.-W. Liu、M.-H. Yang 和 S.-Y. Kuo，《基于辅助点引导改进基于点的人群计数与定位》，载于欧洲计算机视觉会议。施普林格，2024，428–444 8 页码 – 。[46] Y. Ranasinghe、N. G. Nair、W. G. C. Bandara 和 V. M. Patel，《Crowddiff：利用扩散模型进行多假设人群密度估计》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2024，12 809–12 819 8 页码 – 。[47] W. Lin、C. Zhao 和 A. B. Chan，《用于半监督基于点的人群计数的点到区域损失》，载于计算机视觉与模式识别会议论文集，2025，29 363–29 373 8 页码 – 。[48] C. Xu、K. Qiu、J. Fu、S. Bai、Y. Xu 和 X. Bai，《学会缩放：生成人群计数用的多极归一化密度图》，载于 IEEE/CVF国际计算机视觉会议论文集，2019，8382–8390 9 页码 – 。[49] H. Xiong、H. Lu、C. Liu、L. Liu、Z. Cao 和 C. Shen，《从开集到闭集：通过空间分而治之进行目标计数》，载于 IEEE/CVF国际计算机视觉会议论文集，2019，8362–8371 9 页码 – 。[50] J. Gao、Q. Wang 和 X. Li，《PCC Net：通过空间卷积网络进行透视人群计数》，IEEE视频技术电路与系统汇刊，第 30，10，3486–3498，2020 11 卷第期，页码 – 。[51] W. Liu、M. Salzmann 和 P. Fua，《上下文感知人群计数》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2019，5099–5108 11 页码 – 。[52] J. Gao、Q. Wang 和 Y. Yuan，《SCAR：用于人群计数的空间/通道注意力回归网络》，神经计算，第363卷，页码 1–8，2019 11 – 。[53] Q. Wang、J. Gao、W. Lin 和 Y. Yuan，《利用合成数据进行真实场景人群计数学习》，载于 IEEE/CVF 计算机视觉与模式识别会议论文集，2019年，页码8198–8207。11