学习汇报

汇报人：陈珞　　指导老师：郝占军教授　　汇报时间：2025/10/19

PART 01

文献阅读

汇报

西北师范大学 NORTHWEST NORMAL UNIVERSITY

Heterogeneous Dual-Attentional Network for WiFi and Video-Fused Multi-Modal Crowd Counting

期刊： IEEE Transactions on Mobile Computing, 2024

汇报人：陈珞

日期：2025/10/19

研究背景和动机

要点：

1）视频方法：精度高但受遮挡、视角、光照影响。

2）WiFi 方法：覆盖广、隐私友好、低成本，但精度低。

3）核心问题：如何结合两者优势。

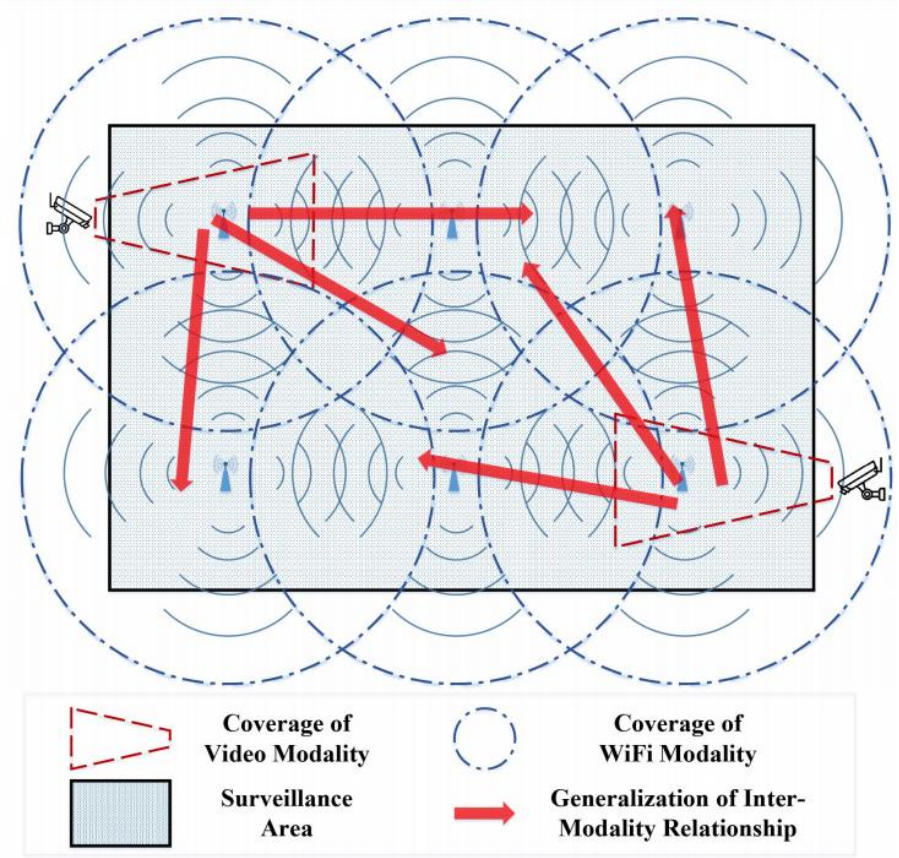
系统设计与数据采集

要点：

1.Hybrid Sensing Network (HSN)： 14 台 Raspberry Pi + 5 台摄像头。

2.时间同步：DS3231 模块 + 网络校时，保证 1s 对齐。

3.每秒采样 1 帧，生成同步的 WiFi 与视频数据对。



a.红色虚线三角形：代表视频模态的覆盖范围。摄像头通过视觉信息监测区域，但存在视野局限（比如被遮挡、远距离模糊等）。

b.蓝色虚线圆形：代表 WiFi 模态的覆盖范围。WiFi 信号通过电磁波反射、衍射等特性感知区域内的物体或人体，可弥补视频的视野缺陷（如非视距场景）。

c.带纹理的矩形：需要监测的目标区域，是视频和 WiFi 模态共同作用的范围。

d.红色箭头：表示跨模态关系的泛化。即通过融合视频和 WiFi 两种模态的信息，挖掘它们之间的关联（比如视觉特征与 WiFi 信号特征的对应关系），让系统能在更复杂场景下稳定工作（如不同光照、遮挡条件）。

WIFI数据处理与WDM构造

定位：WiFi probe → MAC 地址定位 (KNN)。

生成脉冲图：$(I(p) = \sum_i \delta(p - p_i))$

高斯平滑：$(WDM(p) = I(p) * G_\sigma(p))$，σ=2，ks=15。

输出尺寸：140 × 80，滑窗 ΔT=60s，步长 1s，共 2280 样本。

视频数据预处理

要点：

1.预处理方法： ROI+CROP。

ROI，Region of Interest，感兴趣区域是指在图像中划定需要重点关注的区域，排除无关背景干扰；CROP裁剪则是将划定的 ROI 区域从原始图像中截取出来。

2.分辨率： 80 × 160。

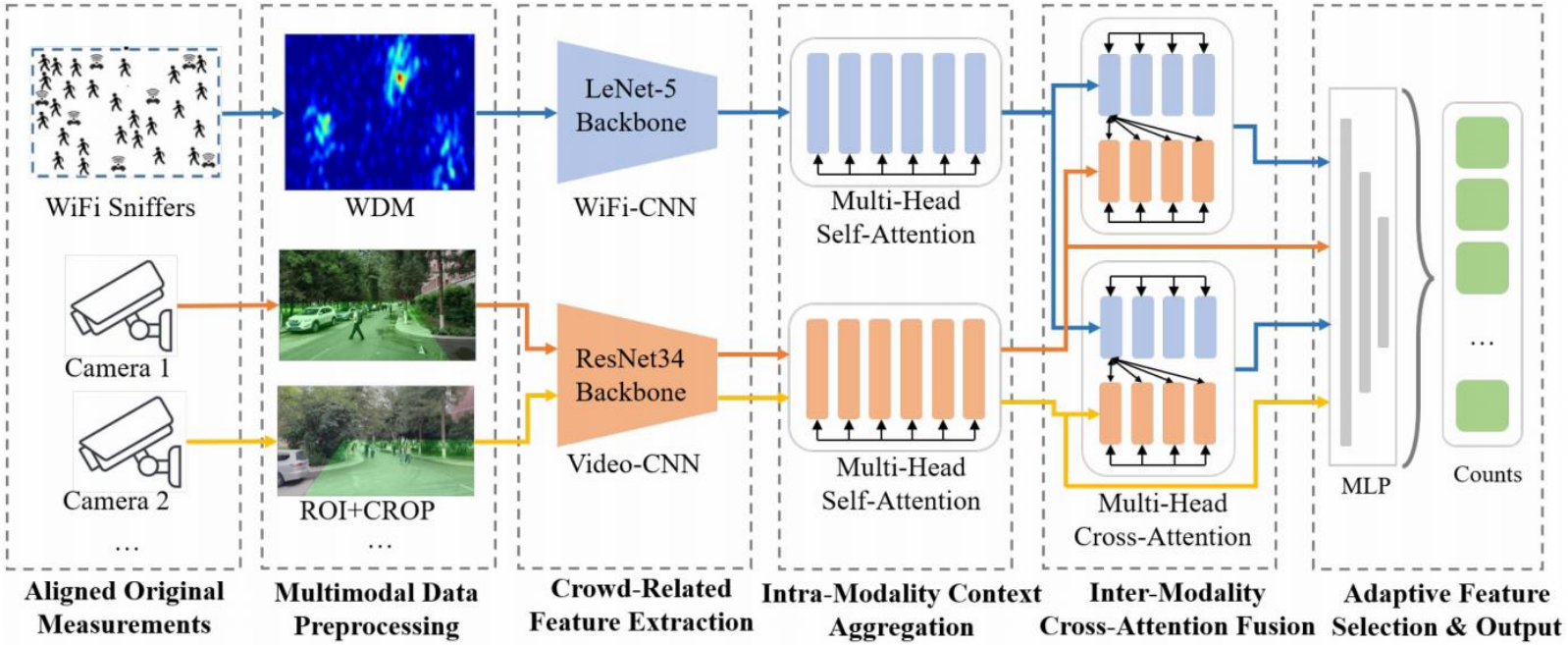
将预处理后的图像或视频帧调整为 80（高度） × 160（宽度）的分辨率：

3.标注： 1 fps 人工标注，总人数 145,617。

以每秒 1 帧的频率对视频进行人工标注。视频通常是 30 fps 或更高帧率，选择 1 fps 标注，在保证标注精度（捕捉人群关键变化时刻）的前提下，大幅减少标注工作量，降低人力成本。

比较： ROI+CROP 最优，平衡精度与计算量。

HDANet 总体结构

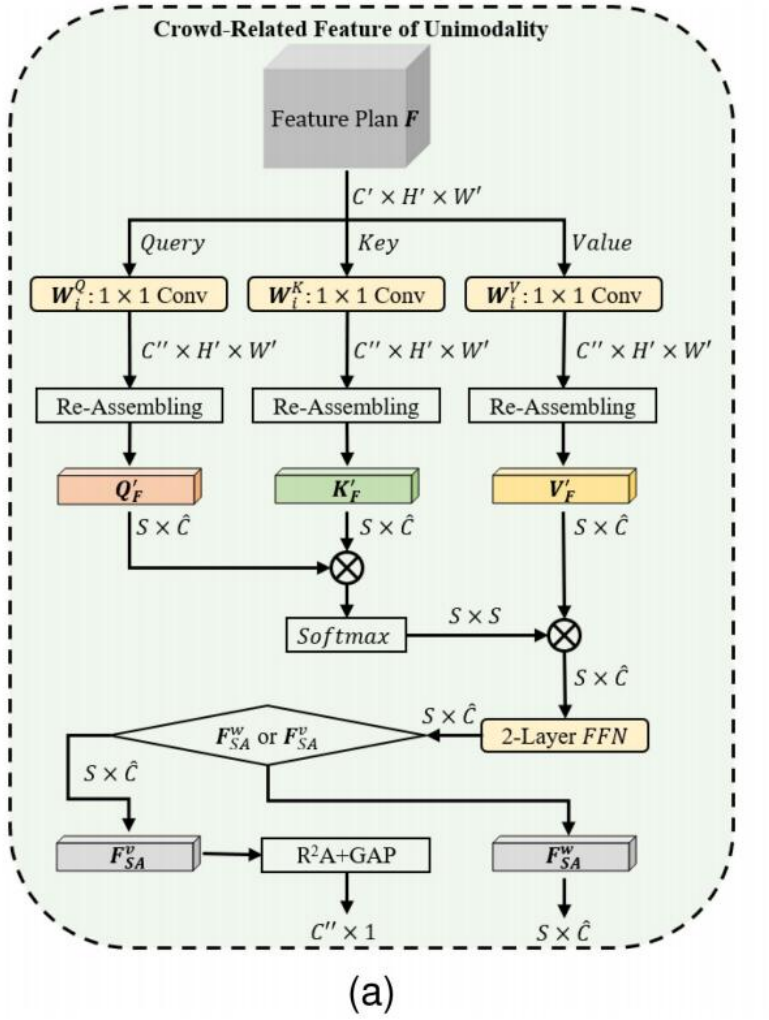


1. 对齐的原始数据
2. 多模态数据预处理
3. 人群相关特征提取
4. 模态内上下文聚合
5. 跨模态交叉注意力融合
6. 自适应特征选择与输出

整体流程:

多模态数据采集→预处理→单模态特征提取→模态内上下文聚合→跨模态融合→自适应输出

模态内 Self-Attention



(a)

要点：

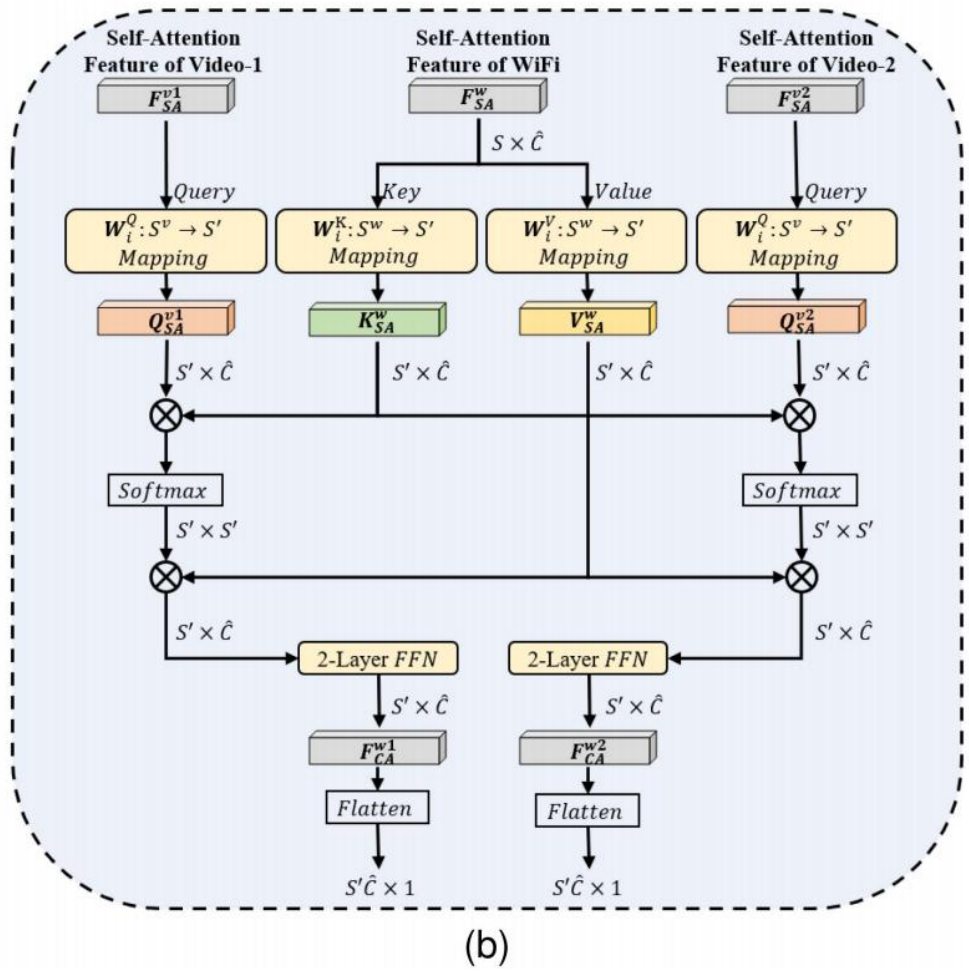
输入特征 → Q/K/V → 多头注意力。

捕获空间与通道依赖关系。

输出 (F_{SA})，增强模态内部一致性。

这个模块的核心是自注意力机制：让单模态特征“自己关注自己”，挖掘特征内部（不同空间位置）的关联。比如，视频模态中“人群密集区域”的特征会与“周围相关区域”的特征建立强关联，从而让特征更精准地反映人群分布

模态间 Cross-Attention（关键创新）



(b)

1. 输入:多模态自注意力特征
2. Query/Key/Value 映射（跨模态）
3. 跨模态注意力计算（点积 + Softmax）
4. 加权求和与前馈网络（2-Layer FFN）
5. 输出：扁平化

该模块的核心是跨模态特征交互：让“视频模态的局部视觉细节”与“WiFi 模态的全局信号分布”建立关联，实现优势互补（视频精准但易遮挡，WiFi 覆盖广但精度略低）。最终输出的跨模态融合特征，既包含视频的视觉精准性，又包含 WiFi 的全局感知能力，为后续多模态任务提供更好的输入。

训练策略与超参

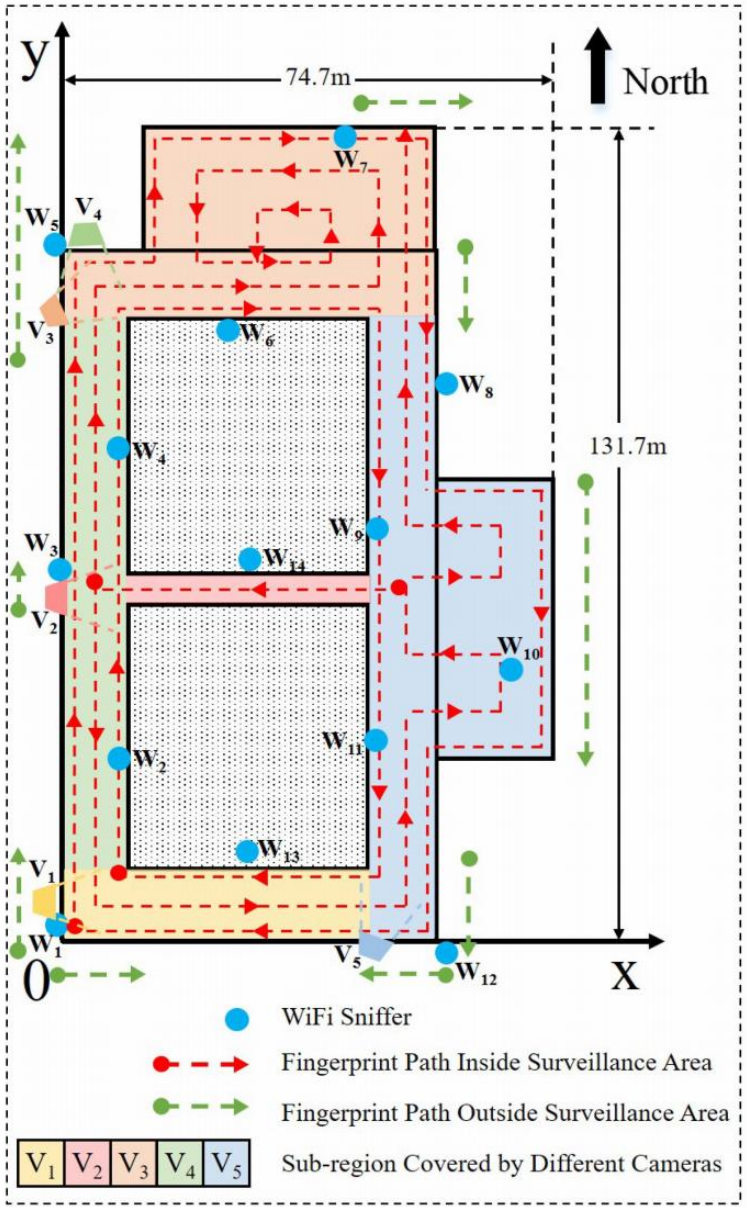
要点：
分步预训练：WiFi / Video 分支先独立训练，再冻结。

Fine-tune Cross-Attention + MLP。

Adam 优化器，MSE 损失，lr(WiFi)=5e-4，lr(Video)=1e-3。

训练 500 epoch，A100 GPU。

多模态数据采集的硬件部署方案



这张图展示的是真实世界数据集测试平台的布局，用于多模态（WiFi + 视频）感知任务（如人群计数、目标定位等），各元素解释如下：

1. 坐标系与区域尺寸图中建立了二维坐标系（x、y），标注了区域的长度（74.7m）和宽度（131.7m），明确了监测场景的物理范围。箭头“North”指示北方，为场景提供方位参考。

2. 关键组件与标识

WiFi Sniffer（WiFi 嗅探器）：蓝色圆点标注，用于采集环境中的 WiFi 信号（如设备的 MAC 地址、信号强度等），是 WiFi 模态数据的来源。

Fingerprint Path（指纹路径）：

红色虚线箭头：监测区域内的指纹路径，用于在监测区域内部采集“位置 – WiFi 信号”的对应关系（构建指纹库，支持定位等任务）。

绿色虚线箭头：监测区域外的指纹路径，用于采集区域外的 WiFi 信号特征，辅助区分“区域内 / 外”的目标。

Sub–region Covered by Different Cameras（不同摄像头覆盖的子区域）：下方彩色方块（V1到V5）对应图中不同颜色的区域，代表每个摄像头的监控覆盖范围，是视频模态数据的采集来源。

结果

Modality	Method	MAE↓ Reduction	MSE↓	MAPE↓	ACC↑
WiFi Unimodality	G-POLY [28]	13.11	296.96	23.86%	76.14%
	DNN-SCC [29]	8.45	119.50	15.79%	84.21%
	WDM+CNN (ours)	**7.98**	**108.70**	**15.16%**	**84.84%**
WiFi+1Video Multi-modality	EF	7.15 ↓ 10.42%	89.31	14.01%	85.99%
	MF	6.82 ↓ 14.54%	75.81	**13.59%**	**86.41%**
	LF	7.15 ↓ 10.42%	88.47	13.80%	86.20%
	HDANet (ours)	**6.18 ↓ 22.61%**	**61.62**	13.73%	86.27%
WiFi+2Videos Multi-modality	EF	6.75 ↓ 15.41%	80.50	12.66%	87.34%
	MF	6.60 ↓ 17.24%	78.97	12.51%	87.49%
	LF	6.91 ↓ 13.36%	89.77	12.74%	87.26%
	HDANet (ours)	**5.89 ↓ 26.21%**	**56.06**	**12.03%**	**87.97%**

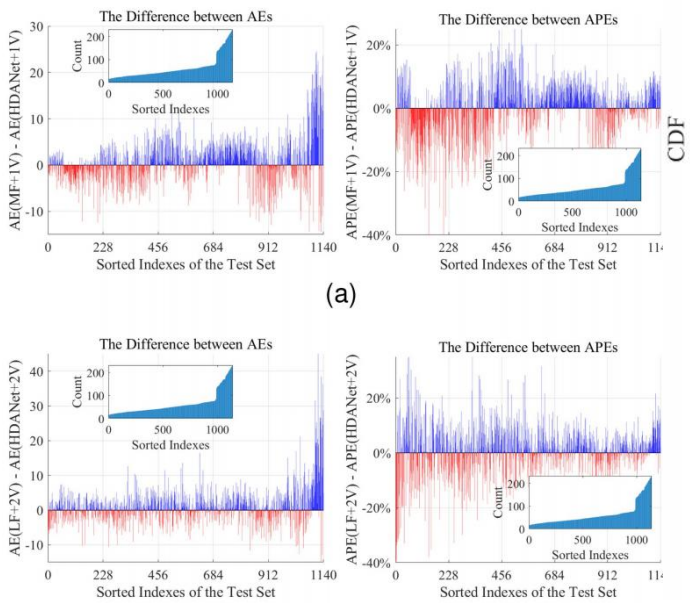
结果解读

1.**单模态 WiFi**：作者提出的WDM+CNN 方法在 MAE、MSE、MAPE、ACC 指标上均优于其他单模态方法（G-POLY、DNN-SCC），说明该方法在单模态 WiFi 计数上更精准。

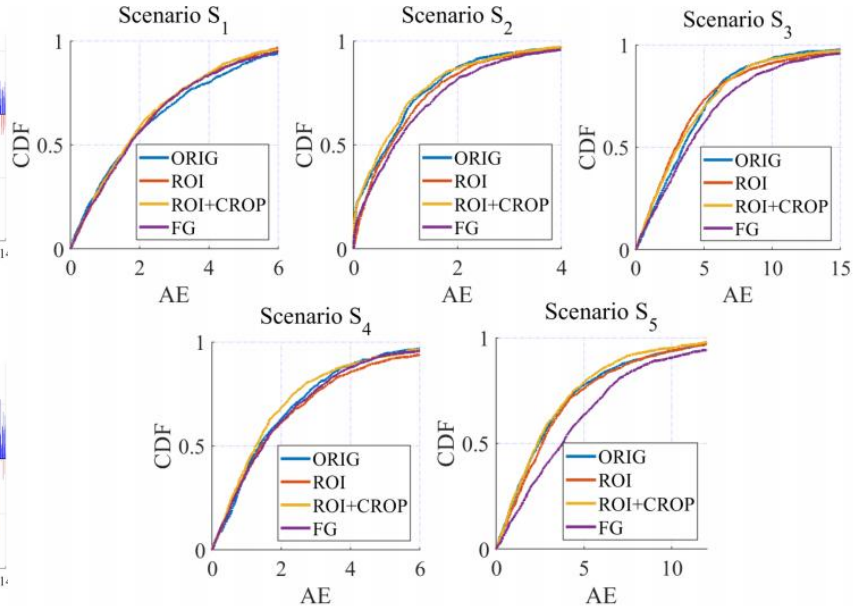
2.**多模态 WiFi+1 个视频（WiFi+1Video Multi-modality）**：作者提出的“HDANet (ours)”在 MAE（下降 22.61%）、MSE、ACC 指标上最优，MAPE 也接近最优，体现出“WiFi + 单视频”多模态融合后，计数性能比单模态 WiFi 有明显提升。

3.**多模态 WiFi+2 个视频（WiFi+2Videos Multi-modality）**：“HDANet (ours)”在所有指标上依旧最优，且 MAE 下降幅度（26.21%）、MSE、MAPE、ACC 的表现都比“WiFi+1 个视频”时更优，说明增加视频数量（从 1 个到 2 个），多模态融合的计数性能进一步提升。

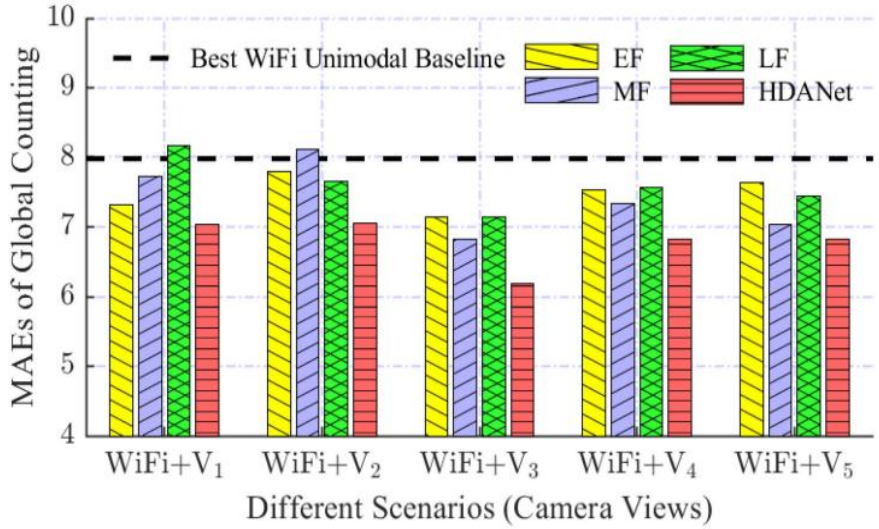
评估



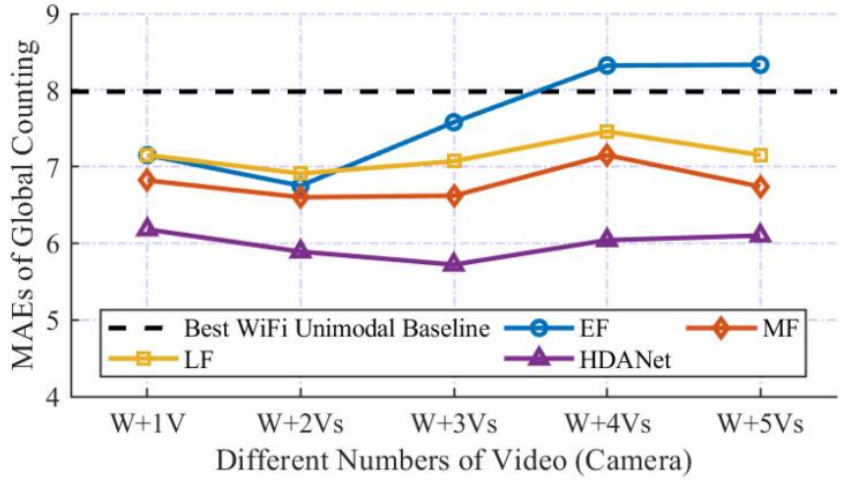
误差差异对比图



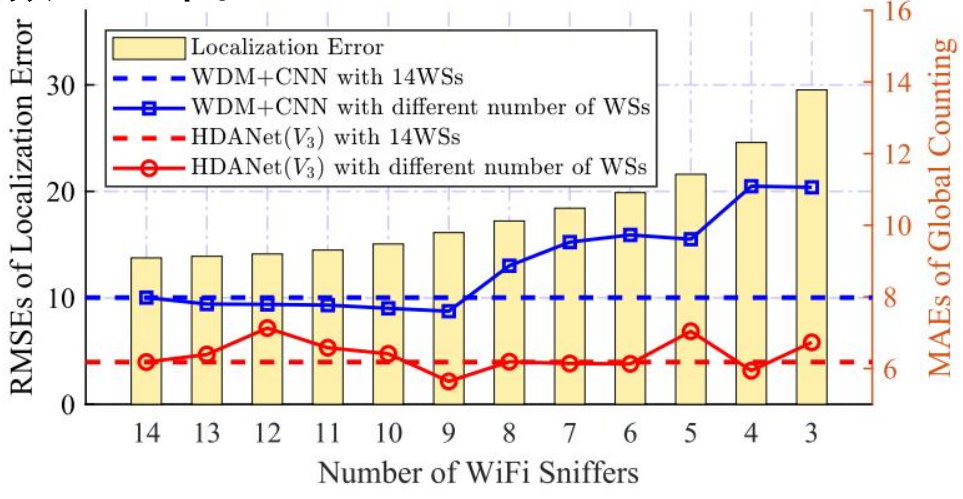
不同预处理的累计分布函数 CDF 图



不同场景下的全局计数 MAE 柱状图



不同视频数量下的全局计数 MAE 折线图



WiFi 嗅探器数量对定位误差与计数误差的影响图

局限与改进方向

要点：

1.时序特征未被充分利用。可引入 RNN、LSTM 等时序模型，学习连续时序关联，提升动态场景计数精度。

2.多视频融合的权衡问题：融合更多视频虽能扩大覆盖，但会增加计算负担与噪声干扰。可设计动态视频选择与加权机制，筛选高质量、互补性强的视频并分配自适应权重。

3.WiFi 定位误差仍有残余影响：WiFi 定位误差会传导至计数任务。可结合多源定位技术修正，并在跨模态融合时用视频精准视觉位置校准 WiFi 信号位置，削弱误差影响。

THANKS