多尺度注意力人群密度估计模型研究

栾方军，田雨晴*，袁帅

沈阳建筑大学计算机科学与工程学院，辽宁沈阳 110168

摘要： 针对密集人群图像中目标尺度变化大、背景干扰强等问题，本文提出一种基于多尺度选择性密度聚合网络（MSDA-Net）的人群计数模型。该模型引入了三个协同模块：多尺度调制模块（MTM）、选择性注意力模块（SAM）和密度上下文感知模块（DCF），三者共同增强特征表达能力和密度图质量，从而提高拥挤复杂场景下的计数精度。具体而言，MTM 模块通过多分支空洞卷积捕获不同尺度的上下文信息，缓解目标尺度剧烈变化带来的影响；SAM 模块利用空间与通道注意力机制抑制背景噪声，突出前景人群区域；DCF 模块采用卷积长短时记忆网络对密度图的上下文依赖关系进行建模，增强相邻区域间的连续性，以生成更平滑、准确的密度图。实验中采用了联合损失函数，结合了欧氏距离损失与结构相似性（SSIM）损失，同时优化像素级回归和局部结构一致性。在四个公开数据集上进行了充分验证，包括 ShanghaiTech Part_A、ShanghaiTech Part_B、UCF_CC_50 和 UCF_QNRF。所提出的 MSDA-Net 在所有数据集上均取得了优异性能，例如在 Part_A 上 MAE 为 58.1，Part_B 上为 6.9，UCF_CC_50 上为 205.3，UCF_QNRF 上为 84.8，超过了多种最先进的方法。消融实验进一步验证了每个模块的单独贡献。总体而言，MSDA-Net 在真实拥挤场景中表现出强大的泛化能力和实际应用价值。

关键词： 密集人群计数；多尺度调制；注意力机制；上下文建模; MSDA-Net

DOI: 10.57237/j.cst.2026.03.001

The Multi-Scale Attention-Based for Crowd Density Estimation

Luan Fangjun, Tian Yuqing*, Yuan Shuai

School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang 110168, China

Abstract: To address the challenges of large scale variations and strong background interference in dense crowd images, this paper proposes a crowd counting model named Multi-scale Selective Density Aggregation Network (MSDA-Net). The model introduces three collaborative modules: the Multi-scale Tuning Module (MTM), the Selective Attention Module (SAM), and the Density Context-aware Module (DCF), which collectively enhance feature representation and density map quality, thereby improving counting accuracy in complex congested scenes. Specifically, the MTM captures multi-scale contextual information through multi-branch atrous convolutions, alleviating the impact of drastic scale variations. The SAM employs spatial and channel attention mechanisms to suppress background noise and highlight foreground crowd regions. The DCF models the contextual dependencies of density maps using convolutional LSTM architecture, enhancing continuity between adjacent regions to generate smoother and more accurate density maps. A joint loss function combining Euclidean distance loss and structural similarity (SSIM) loss is adopted to optimize pixel-level regression and local

基金项目：国家自然科学基金 (62073227)；辽宁省科技厅项目 (2023JH2/101300212).

*通信作者：田雨晴, tyuqing0619@163.com

收稿日期: 2026-04-14；接受日期: 2026-05-07；在线出版日期: 2026-06-13

structural consistency simultaneously. Extensive experiments are conducted on four public benchmarks, including ShanghaiTech Part_A, ShanghaiTech Part_B, UCF_CC_50, and UCF_QNRF. The proposed MSDA-Net achieves superior performance on all datasets, e.g., MAE of 58.1 on Part_A, 6.9 on Part_B, 205.3 on UCF_CC_50, and 84.8 on UCF_QNRF, outperforming many state-of-the-art methods. Ablation studies further validate the individual contribution of each module. Overall, MSDA-Net demonstrates strong generalization ability and practical application value in real-world crowded scenes.

Keywords: Dense Crowd Counting; Multi-scale Modulation; Attention Mechanism; Context Modeling; MSDA-Net

1 引言

密集人群计数作为计算机视觉领域的一个重要分支任务，旨在估计图像中人数的总量，广泛应用于智慧城市[1]、交通管控[2]、公共安全[3]、应急疏散[4]等实际场景中。该任务具有极大的挑战性，尤其是在密集、遮挡严重或摄像机视角变化大的复杂场景中，模型往往难以准确预测人数。

目前，主流人群计数方法多基于密度图回归技术，将原始图像映射为密度分布图，通过积分计算获得人数。这类方法取得了较好的效果，但仍存在若干关键问题：首先，目标尺度变化大导致模型感受野不足，难以适应从稀疏到极度密集的人群图像。其次，传统多尺度结构（如 MCNN）存在特征冗余问题，不同分支间缺乏有效的信息交互；再次，预测密度图边缘模糊，背景干扰严重，影响整体估计精度。最后，部分数据集图像存在较多噪声和压缩伪影，对模型训练收敛性和泛化能力构成挑战。

为了解决上述问题，本文提出一种基于多尺度注意力密度聚合网络（Multi-scale Selective Density Aggregation Network，MSDA-Net）的人群计数模型。该网络由三大核心模块组成：多尺度调制模块（MTM）、选择性注意力融合模块（SAM）以及密度上下文建模模块（DCF），构成层次化的多尺度密度图预测流程。

具体而言，MTM 模块融合了传统多列结构与特征调制机制，在自顶向下路径中引入多尺度特征信息流，增强尺度感知能力；SAM 模块引入通道与空间注意力机制，引导模型自适应选择最有效分支特征，提升信息融合效果；DCF 模块则通过堆叠卷积结构实现对密度图边缘与上下文的建模，进一步优化预测密度图的边界质量与一致性。

为了提高训练精度与泛化能力，本文设计了联合损失函数，融合了最优传输损失（OTLoss）[5]、结构相似性损失（SSIM Loss）[6]以及 L1 计数损失，共同引导模型在密度图质量与人数计数上的协同优化。

本文主要工作如下：

1) 构建多尺度选择性密度聚合网络整体框架。针对传统人群计数方法在复杂密集场景中泛化能力不足的问题，本文从网络整体结构层面出发，设计了一种融合多尺度特征调制、选择性注意力融合与上下文密度建模的统一计数框架。该网络以卷积神经网络为主干，通过多分支结构提取不同尺度的特征表示，并在统一框架下实现特征调制、融合与密度预测，为后续模块设计提供整体支撑；
2) 设计多尺度特征调制模块（MTM），提升尺度建模能力。针对人群尺度变化剧烈、不同尺度特征语义不一致的问题，本文提出多尺度特征调制模块（Multi-scale Top-down Modulation，MTM）。该模块通过引入自顶向下的特征调制机制，在不同尺度特征之间建立有效的信息交互通道，使高层语义信息能够对低层细节特征进行引导，从而增强网络对不同尺度人群目标的感知能力；
3) 提出选择性注意力融合模块（SAM），抑制冗余特征与背景干扰。针对多尺度特征融合过程中普遍存在的信息冗余和背景噪声干扰问题，本文设计选择性注意力融合模块（Selective Attention Module，SAM）。该模块结合通道注意力与空间注意力机制，对来自不同尺度和不同分支的特征进行自适应加权，引导网络关注更具判别力的人群区域特征，从而提高特征融合的有效性和计数结果的稳定性；
4) 构建上下文密度建模模块（DCF），优化密度图结构一致性。针对基于密度图回归方法中预测结果易出现局部噪声和结构不连续的问题，

本文提出上下文密度建模模块（Density Context Fusion，DCF）。该模块通过对预测密度图进行上下文建模，增强密度图在空间维度上的连续性和一致性，从而在保证整体计数精度的同时，提高密度图的表达质量和可解释性。

2 相关工作

当前，密集人群计数技术已成为计算机视觉中的核心研究方向，其方法主要分为三类：基于检测的方法[7]、基于回归的方法[8]和基于密度估计的方法[9]。其中，基于目标检测的方法通过检测图像中每个个体的边界框或关键点后统计个体数量，但在密集遮挡场景中性能明显下降，且对标注依赖强；基于回归的方法通常学习区域图像特征与人数之间的映射关系，计算效率高但缺乏空间位置信息；相比之下，基于密度图估计的方法通过预测像素级的密度分布，继而对密度图积分获得总人数，已成为当前主流研究范式。

尽管密度图回归方法在多个标准数据集上取得显著进展，如 MCNN [10]、CSRNet [11]、BL [12]等模型在建模密集区域结构上各具优势，但仍面临以下三个关键问题：一是多数方法感受野受限、尺度适应能力弱，难以精准建模从稀疏到极密的人群结构；二是信息融合结构简单，特征表达存在冗余与干扰，导致密度图边缘模糊、背景误识现象严重；三是模型对原始图像质量敏感，尤其在图像存在噪声、伪影或低对比度区域时，预测精度不稳定。

为了缓解上述问题，研究者从多个方向展开改进探索。一方面，多尺度结构被广泛引入以增强感受野和尺度表达能力。例如，MCNN 通过并列多个不同卷积核大小的分支提取不同尺度信息；SANet [13]引入残差模块融合多尺度特征以增强密集区域的鲁棒性；另一方面，注意力机制逐渐被应用于特征选择与信息聚合过程，如 SCA-CNN [14]在通道和空间维度上引导模型关注关键区域，有效降低冗余干扰。同时，上下文建模机制也被引入以优化密度图细节，CSRNet 中引入空洞卷积增强上下文感知能力，Bayesian Loss [15]则从分布角度建模密度图不确定性，提高估计可靠性。

尽管现有方法从多尺度提取、注意力融合、上下文建模、图像预处理等多个方面取得进展，但仍存在以下不足：

1) 多分支结构中信息融合策略仍以简单拼接或加权为主，缺乏有效的分支选择与冗余抑制机制；
2) 不同尺度间语义信息交互不足，难以有效提升密度图的整体结构质量；
3) 部分密度图预测模型过度依赖固定高斯核密度图作为训练监督，忽视了点标注之间的结构分布；
4) 原图质量未加以预处理，背景噪声易对预测结果产生干扰。

为此，本文提出一种融合多尺度特征调制（MTM）、分支注意力融合（SAM）与上下文建模（DCF）的新型密集人群计数网络 MSDA-Net。该方法进一步引入复合损失函数（OT+SSIM+L1）共同优化模型结构，针对多尺度冗余、背景干扰与结构模糊等问题提出系统性解决方案，提升模型在多种密集场景下的计数精度与泛化能力。

3 本文模型

为有效提升密集人群计数任务在多尺度场景中的预测精度与密度图质量，本文提出一种融合多种模块优势的计数网络——MSDA-Net（Multi-scale Selective Density Aggregation Network），旨在解决现有模型在尺度不一致、背景干扰、密度图边界模糊等方面的缺陷。

3.1 网络整体结构

本文提出的密集人群计数网络 MSDA-Net（Multi-scale Selective Density Aggregation Network）面向高密度、尺度变化剧烈的复杂场景，融合多尺度结构调制、注意力选择机制以及上下文建模策略，形成端到端的密度估计流程。

MSDA-Net 主要由以下四个组成部分构成：

1) 主干特征提取网络：基于 VGG16 改进，负责从原始图像中提取多层级、高分辨率的基础视觉特征；
2) MTM 模块（Multi-scale Top-down Modulation）：引导多尺度路径进行语义增强的特征提取，结合上下层结构，提升对不同尺度目标的感知能力；
3) SAM 模块（Selective Attention Module）：将来自不同尺度分支的特征进行融合，同时借助通道与空间注意力机制，实现对有效区域的重点建模；

4) DCF 模块（Density Context Fusion）：对融合特征图进行密度建模与上下文细化，以获得结构更清晰、边缘更平滑的预测密度图。

整个网络结构如图 1 所示。

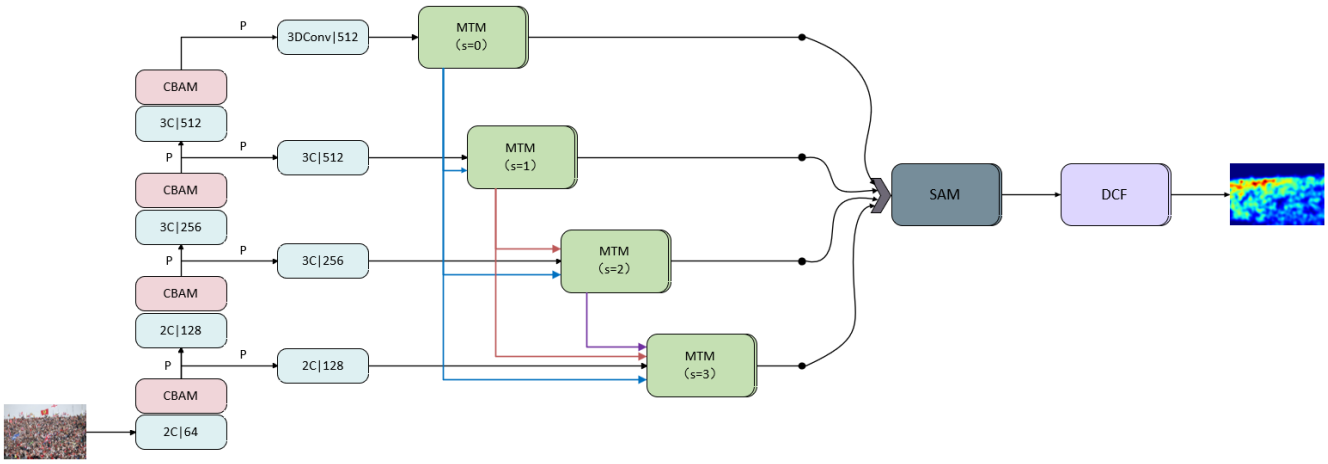

图 1 MSDA-Net 网络整体结构示意图

如图 1 所示，输入图像经过特征提取与多尺度卷积获得多尺度特征表示，经由注意力机制进行选择性融合，最后通过上下文密度建模模块生成高质量的密度图，并对其进行积分得到预测人数。

设输入图像为 I，网络映射函数为 $F_\theta(\bullet)$，预测的密度图为 D，则可表示为：

$$D = F_\theta(I) \tag{1}$$

通过对预测密度图进行像素求和，得到预测人数：

$$\hat{C} = \sum_{x,y} D(x,y) \tag{2}$$

该整体结构设计考虑到密集场景中目标尺度变化大、背景噪声多、结构边界模糊等问题，充分结合多尺度提取与选择性关注机制，在保持高精度预测的同时提升了模型的泛化能力和结构表达能力。

3.2 特征提取模块

特征提取模块以经典的 VGG-16 为基础架构进行深度定制。为避免传统池化操作导致高频空间细节的过早丢失，在主干网络的每一个卷积阶段之后、最大池化下采样层之前均嵌入了卷积块注意力模块 CBAM，以在特征降维前动态滤除背景噪声。网络采用全卷积架构剥离全连接层，并在逐层降采样过程中输出四个不同分辨率与语义层级的特征集合，记为 $\{F_1, F_2, F_3, F_4\}$，其对应的下采样倍率分别为 1/2、1/4、1/8 和 1/16。其中，浅层特征包含丰富的局部结构与边缘细节，深层特征则提供大感受野的抽象语义，共同为后续跨尺度特征调制提供高质量的输入基底。特征提取模块的网络结构图如图 2 所示。

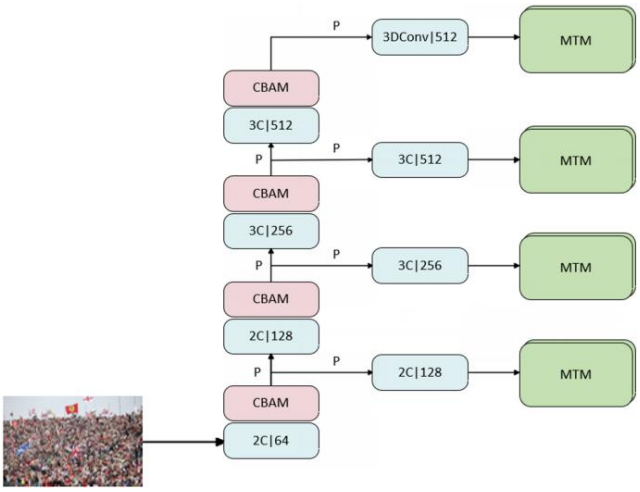

图 2 改进的特征提取模块示意图

如图 2 所示，主干网络采用垂直向上传递路径，网络依次包含了 2C|64、2C|128、3C|256、3C|512 四个主要的卷积阶段。本文特意将 CBAM 模块嵌入在每一个卷积阶段之后、最大池化下采样层 P 之前。这种设计的核心动机在于，在特征图被池化缩小、丢失高频空间细节之前，提前利用注意力机制过滤掉背景噪声，确保进入下一层级的特征是纯净的。

3.3 多尺度特征调制模块 MTM

为充分发挥不同层级特征的互补优势，缓解剧烈尺度变化带来的特征不匹配问题，设计了自顶向下的多尺度特征调制模块（MTM）。该模块摒弃了直接的特征拼接，转而构建逐级的跨尺度密集连接机制 。对于特定的尺度层级，MTM 同时接收来自上层尺度分支的高级语义特征与本层局部分支的特征 。具体操作中，高层特征首先通过 1×1 卷积进行通道与尺度对齐，并与本层特征拼接后经 CBAM 进行联合重标定，形成语义引导信号；与此同时，本层特征通过并联的 3×3、5×5、7×7 卷积核提取多尺度动态响应以应对特征相似性。最后，将语义引导信号与多尺度特征融合，通过卷积与 ReLU 激活后输出调制特征，从而实现高层全局语义对低层细节的精准降维引导。MTM 模块结构示意图如图 3 所示。

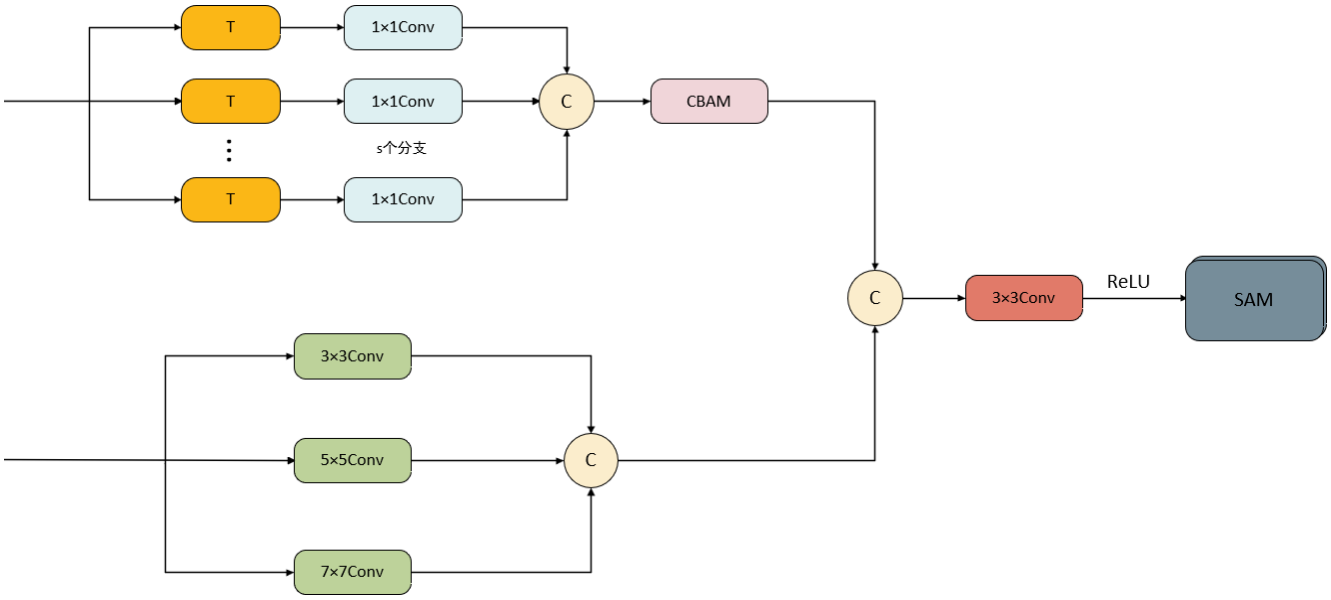

图 3 MTM 模块结构示意图

如图 3 所示，每个 MTM 模块的输入由两部分组成。上面端口接收来自不同尺度的特征，由 s 个并行尺度分支（s=0，1，2，3）构成，每个分支对应不同的下采样倍率（如 $1/2^0$、$1/2^1$、$1/2^2$、$1/2^3$），以不同感受野提取目标特征。考虑到不同尺度特征在空间分辨率与通道维度上的差异，调制过程首先对高层特征进行尺度对齐与通道匹配，使其在特征维度上与低层特征保持一致。该对齐过程仅用于保证特征在交互时的维度可比性，并不引入跨尺度的并行融合操作。在完成对齐后，经 1×1 卷积压缩后与本层特征拼接，构建跨层语义增强机制。拼接后的多源特征进一步通过 CBAM（Convolutional Block Attention Module）进行通道注意力与空间注意力调制，引导网络更聚焦于人头密集区域与关键结构位置。

下面端口输入取自于特征提取网络对应的尺度特征分支，通过三列并行的 3×3 分支、5×5 分支、7×7 分支，对应的卷积滤波器分别为 3、5、7，可以捕获更大的尺度范围，有效解决特征相似性。输出特征拼接后与上面端口处理后的结果进行拼接，调制后的特征再通过一组卷积与 ReLU 激活形成最终输出。

3.4 选择性注意力融合模块 SAM

由于 MTM 模块输出的多尺度特征在维度与空间位置上对最终密度预测的贡献度存在显著差异，为进一步聚焦人群核心区域，设计了选择性注意力融合模块 SAM。

SAM 模块结构示意图如图 4 所示。

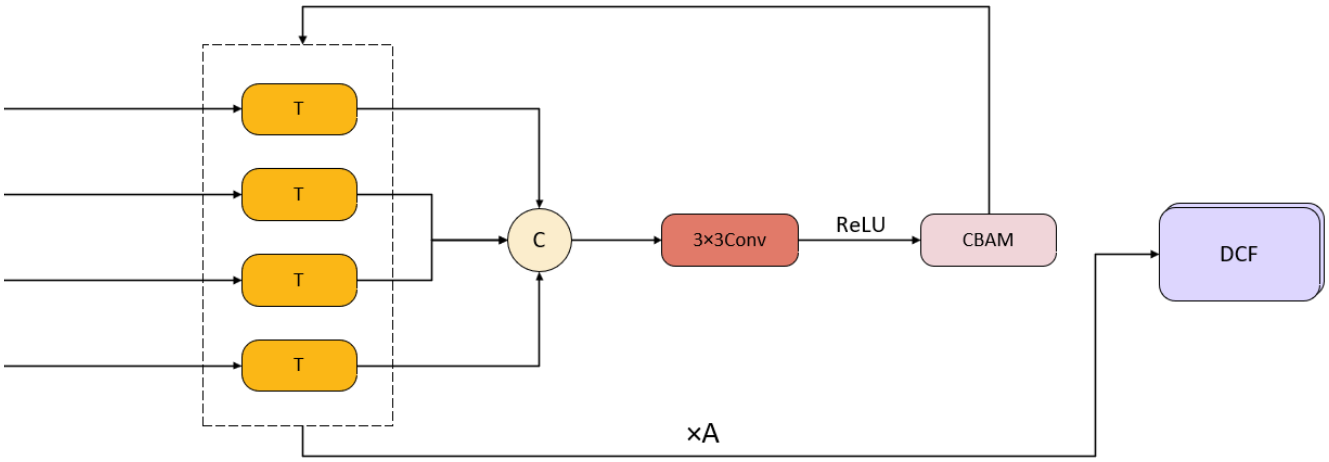

图 4 SAM 模块结构示意图

如图 4 所示，首先将多尺度特征 $\{T_0,T_1,T_2,T_3\}$ 上采样至相同分辨率并沿通道维拼接，通过 3×3 卷积融合生成初步特征 F_{fused}：

$$F_{fused}=\mathrm{Re}LU\left(Conv_{3\times3}\left(Concat\left(T_0,T_1,T_2,T_3\right)\right)\right) \tag{3}$$

随后，执行通道注意力机制。利用全局平均池化和最大池化对特征进行空间压缩，经共享多层感知机 MLP 交互后生成通道权重 M_c，进而得到通道重标定特征 F_c'：

$$M_c=\sigma\left(MLP\left(AvgPool\left(F_{fused}\right)\right)+MLP\left(MaxPool\left(F_{fused}\right)\right)\right) \tag{4}$$

$$F_{out}=M_s\otimes F_{fused} \tag{5}$$

在此基础上执行空间注意力机制。将 F_c' 沿通道维分别池化并拼接，通过 7×7 卷积提取二维空间依赖生成空间权重 M_s，输出最终的选择性增强特征 F_{out}：

$$M_s=\sigma\left(Conv_{7\times7}\left(Concat\left(MaxPool\left(F_c'\right),AvgPool\left(F_c'\right)\right)\right)\right) \tag{6}$$

$$F_{out}=M_s\otimes F_c' \tag{7}$$

3.5 DCF 模块

为克服网络末端感受野扩大导致的局部细节丢失、密度图边缘模糊等问题，引入上下文密度建模模块 DCF 以重构密度分布的空间拓扑结构。DCF 模块结构示意图如图 5 所示。

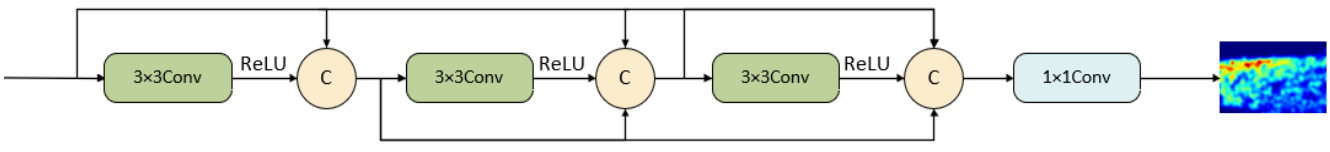

图 5 DCF 模块结构示意图

如图 5 所示，DCF 模块采用密集前馈机制（Dense Connection），内部包含三个连续的 3×3 卷积块。设输入的高质量选择性特征为 F_{SAM}，其密集信息流传递过程如下：

$$H_1 = \mathrm{Re}LU\left(Conv_{3\times3}\left(F_{SAM}\right)\right) \tag{8}$$

$$F_{SAM1} = Concat\left(F_{SAM}, H_1\right) \tag{9}$$

$$H_2 = \mathrm{Re}LU\left(Conv_{3\times3}\left(F_{SAM1}\right)\right) \tag{10}$$

$$F_{SAM2} = Concat\left(F_{SAM}, H_1, H_2\right) \tag{11}$$

$$H_3 = \mathrm{Re}LU\left(Conv_{3\times3}\left(F_{SAM2}\right)\right) \tag{12}$$

$$F_{SAM3} = Concat\left(F_{SAM}, H_1, H_2, H_3\right) \tag{13}$$

通过极致的跨层特征复用，F_{SAM3} 充分聚合了不同深度的局部细节与全局上下文依赖 。最后，利用 1×1 回归头对高维特征张量进行通道压缩，输出最终预测的二维人群密度图 $\hat{Y}$：

$$\hat{Y} = Conv_{1\times1}\left(F_{SAM3}\right) \tag{14}$$

3.6 损失函数

在人群计数任务中，损失函数的设计对模型的训练效果起着至关重要的作用。传统方法大多采用 L2 损失或L1损失来约束预测密度图与真值密度图之间的差异，但这类基于逐像素差异的度量在密集场景下往往不足以捕获全局与局部结构特征，从而导致预测结果与真实分布存在偏差。针对这一问题，本文在 MSDA-Net 中设计了由最优传输损失（OT Loss）、结构相似性损失（SSIM Loss）和 L1 损失共同组成的复合损失函数，以平衡全局分布、局部结构和总数误差三方面的优化目标。

3.6.1 最优传输损失

OT Loss 的目标是通过 Sinkhorn 算法计算预测密度分布与点注释分布之间的最优传输成本，从而衡量预测结果与真实标注在整体分布上的差异：

$$L_{OT}\left(P,G\right) = \langle \alpha, \mu \rangle + \langle \beta, \nu \rangle \tag{15}$$

其中，P 表示预测密度图像素点集合，G 表示点注释图像素点集合，μ,ν 分别为两者的概率分布，α,β 为 Sinkhorn 算法学习得到的参数。OT Loss 能够直接利用点标注进行优化，从而避免高斯核生成真值密度图带来的误差，并提升全局一致性。

3.6.2 结构相似性损失

OT Loss 注重全局分布，而 SSIM Loss 关注局部结构特征。其核心思想是通过亮度、对比度和结构三方面来度量预测密度图与真值密度图的相似性：

$$SSIM\left(P,G\right) = \frac{\left(2\mu_p\mu_g + C_1\right)\left(2\sigma_{pg} + C_2\right)}{\left(\mu_p^2 + \mu_g^2 + C_1\right)\left(\sigma_p^2 + \sigma_g^2 + C_2\right)} \tag{16}$$

其中，μ_p,μ_g 为预测与真实密度图的均值，σ_p,σ_g 为方差，σ_{pg} 为协方差，C_1,C_2 为稳定常数。

在训练时，我们采用如下形式定义 SSIM 损失：

$$L_{SSIM} = 1 - \frac{1}{N}\sum_{x} SSIM\left(x\right) \tag{17}$$

该损失能够有效约束模型保留局部结构信息，使预测密度图在空间分布上更贴近真实情况。

3.6.3 L1 损失

为了降低预测人数与真实人数之间的误差，本文引入 L1 损失：

$$L_{L1} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{C}_i - C_i\right| \tag{18}$$

其中，$\hat{C}_i$ 为预测人数，C_i 为真实人数。该损失直接约束计数值，提升模型在整体人数估计上的精度。

综上所述，MSDA-Net 的最终优化目标由三种损失加权组成：

$$L_{Total} = \alpha \cdot L_{OT} + \beta \cdot L_{SSIM} + \gamma \cdot L_{L1} \tag{19}$$

其中，权重系数设定为 α =1.0，β =0.5，γ =0.3，以确保全局分布约束占主导，同时兼顾局部结构信息和总数精度。该复合损失函数在平衡全局与局部特征的同时，能够有效提升预测密度图的质量和计数的准确性。

4 实验结果与分析

4.1 实验设置

为了验证所提出的 MSDA-Net 网络在人群计数任务中的有效性，本文在四个具有代表性的数据集上进行评估，包括 ShanghaiTech Part_A（SHA）、Part_B（SHB）、UCF-QNRF 和 UCF-CC-50 数据集。所有图像在训练前均进行了统一预处理。首先对每个训练图像用随机缩放、智能裁剪以及水平翻转的方式进行数

据增强图像尺寸分别归一化为 512、384、512、512。

在训练阶段，使用学习率为 1e-4 的 Adam 算法进行参数优化，使用余弦学习率调度器，采用复合损失函数（OT Loss + SSIM Loss + L1 Loss）进行监督训练，批次大小设置为 6，训练 200-300 轮以达到最优性能。模型使用的主干网络为 VGG16 的前 10 层，并替换最后卷

积为扩张卷积以扩大感受野。所有实验均在相同环境下完成，模型使用 PyTorch 框架在单张 NVIDIA GPU 上训练。

4.2 评价指标

本文采用人群计数中最常用的两个指标对模型性能进行评估，分别是：

平均绝对误差（Mean Absolute Error, MAE）：表示预测值与真实值之间的平均绝对差值，衡量计数精度。

均方根误差（Root Mean Square Error, RMSE）：表示误差的标准差，衡量模型在极端样本下的稳定性。

定义如下：

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|p_i - g_i| \tag{20}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|p_i - g_i|^2} \tag{21}$$

其中，p_i 表示第 i 张图像预测人数，g_i 为真实人数，N 为测试图像总数。

4.3 数据集与实验结果

为提高模型在不同场景下的泛化能力，研究者构建了若干用于人群计数的标准数据集。其中包括 ShanghaiTech Part_A 、 ShanghaiTech Part_B 、 UCF_CC_50、以及 UCF_QNRF 等。各数据集上的实验结果汇总如表 1 和表 2 所示。

表 1 在 SHA 和 SHB 数据集上的实验结果

方法	SHA		SHB	
	MAE	RMSE	MAE	RMSE
MCNN	110.2	173.2	25.4	41.3
Switch	90.4	135.0	21.6	33.4
CP-CNN	73.6	106.4	20.1	30.1
Ic-CNN	68.5	116.2	10.7	16.0
CSRNet	68.2	115.0	10.6	16.0
SDANet	63.6	101.8	7.8	10.2
SCAR	66.3	114.1	9.5	15.2
Bayesian	62.8	101.8	7.7	12.7
HA-CNN	62.9	94.9	8.1	13.4
DENet	65.5	101.2	9.6	15.4
LSC-CNN	66.4	117.0	8.1	12.7
MSCANet	66.5	102.1	—	—
MFPNet	65.5	112.5	8.7	13.8
MSDA-Net	58.1	98.7	6.9	11.7

表 2 在 UCF_CC_50 和 UCF_QNRF 数据集上的实验结果

方法	UCF_CC_50		UCF_QNRF	
	MAE	RMSE	MAE	RMSE
MCNN	377.6	509.1	277.0	426.0
Switch	318.1	439.2	228.0	445.0
CP-CNN	295.8	320.9	—	—
Ic-CNN	260.9	365.5	—	—
CSRNet	266.1	397.5	—	—
SDANet	227.6	316.4	—	—
SCAR	259.0	374.0	—	—
Bayesian	229.3	308.2	88.7	154.8
HA-CNN	256.2	348.4	118.1	180.4
DENet	241.9	345.4	—	—
LSC-CNN	225.6	302.7	120.5	—
MSCANet	242.8	329.8	104.1	183.8
MFPNet	240.8	384.4	112.0	190.7
MSDA-Net	205.3	285.4	84.8	145.2

4.3.1 ShanghaiTech Part_A（SHA）

SHA 数据集是一个密集人群计数数据集，包含高密度大场景图像，共 482 张训练图像和 316 张测试图像，平均每图人数超过 500 人。表 1 中显示，MSDA-Net 在该数据集上取得了优异的性能，说明模型在高密度人群场景中具有良好的尺度感知和上下文建模能力。

4.3.2 ShanghaiTech Part_A（SHB）

SHB 数据集相比 Part_A 更加稀疏，主要采集于上海街头，具有更加规律的人群分布和较低的人群密度。在此数据集上，MSDA-Net 依然取得了良好的 MAE 和 RMSE 表现，表明模型具有较强的泛化能力。

4.3.3 UCF_CC_50

UCF_CC_50 是一个包含 50 张图像的小样本数据集，其中人群密度极高，每张图像的头部标注数量从几十到几千不等。由于数据稀少，训练时容易过拟合，MSDA-Net 通过模块间的信息融合机制，稳定学习不同密度场景下的统计规律，利用五折交叉验证实现了对

小样本高密度场景的有效建模。

4.3.4 UCF-QNRF

UCF-QNRF 是当前最具挑战性的大规模人群计数数据集之一，包含 1535 张高分辨率图像，标注了超过 125 万个头部点。由于其复杂的拍摄角度、密度和背景干扰，该数据集对于模型的鲁棒性要求极高。MSDA-Net 在该数据集上的表现表明其设计的多尺度注意力机制和上下文建模模块能够有效捕捉复杂场景下的语义信息，实现精准计数。

4.4 模块消融实验

表 3 给出了不同模型结构在 SHB 数据集上的性能表现。可以看到，基础网络性能最差，说明主干网络本身虽能进行人群计数，但缺乏多尺度特征提取和上下文信息建模能力。单独加入 MTM、SAM 或 DCF 模块后，网络性能有所提升，其中 MTM 模块对多尺度特征提取起到明显作用，SAM 模块增强了注意力选择能力，DCF 模块改善了密度预测的上下文一致性。

最终，完整的 MSDA-Net 将三者融合，获得了最优的 MAE 和 RMSE，充分说明 MTM、SAM 和 DCF 模块的有效性及互补性。

表 3 模块组合消融实验结果

MTM	SAM	DCF	MAE	RMSE
×	×	×	8.2	13.0
√	×	×	7.6	12.4
×	√	×	7.7	12.7
×	×	√	8.0	12.8
√	√	√	6.9	11.7

4.5 可视化分析

为进一步验证所提出方法的可解释性和密度图质量，本文展示了不同模型在测试图像上的密度图可视化结果。如图 6 所示，MSDA-Net 生成的密度图在头部聚集区域具有更清晰的边界，且背景区域噪声更低。与 MCNN 和 CSR_Net 等方法相比，本文方法的预测结果与真值密度图更为接近。

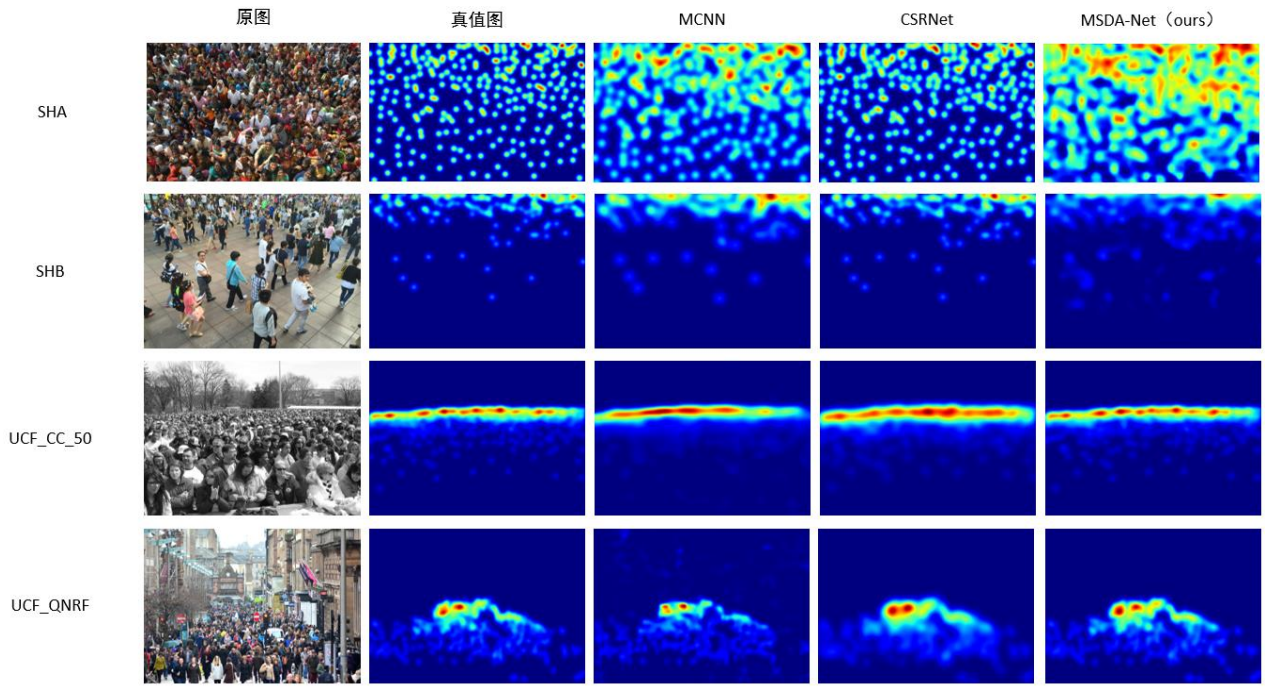

图 6 MSDA-Netket 可视化结果对比图

5 结论

本文旨在解决高密度人群计数任务中，由于尺度变化剧烈、背景复杂等因素导致的计数误差问题。为此，设计了一种基于多尺度注意力机制的密集人群计数网络 MSDA-Net（Multi-scale Selective Density Aggregation Network）。该网络结合了多列感受野增强、注意力信息融合与上下文密度建模三类策略，具有强大的尺度感知能力和密度图精度。

其中，MTM 模块融合了传统 TFM 的自顶向下调制能力和 MCM 多列结构的尺度捕获能力，通过多个尺度的特征调制与融合，使网络在复杂场景中仍能精准定位头部区域，提升了模型的结构解析能力。SAM 模块通过空间-通道联合注意力机制对多尺度特征进行选择性融合，提升模型对关键区域的响应能力，避免了信息冗余与误导。DCF 模块则通过堆叠多个 3×3 卷积核进行上下文建模，使得模型能够更好地理解密集场景下目标之间的关联性，并有效生成高质量密度图。

实验结果表明，本文提出的 MSDA-Net 在四个主流人群计数数据集（SHA、SHB、UCF-QNRF、UCF-CC-50）上均取得了具有竞争力的结果，在密集场景下的表现尤为突出，验证了多模块协同设计的有效性和实用性。相较于传统的单尺度或简单注意力机制网络，MSDA-Net 更适应于密度变化剧烈、目标拥挤复杂的计数任务。

然而，本研究也存在一定局限性。一方面，当前模型仍然需要较为复杂的模块组合与较高的计算资源，部署在边缘设备上存在一定挑战；另一方面，虽然模型在密集区域表现良好，但在极稀疏区域的人群识别精度仍有待提升，后续可考虑引入 Transformer 或更轻量化的全局建模机制以增强鲁棒性与泛化性。

综上所述，MSDA-Net 为高密度人群计数任务提供了一种高效可靠的解决方案，并为后续相关研究提供了理论支撑和方法基础，未来将继续优化模型结构，提升模型部署效率，并结合半监督与小样本方法进一步扩展模型的适应性和实用范围。

参考文献

[1] ZANNELLA C, BOVE S. Smart city context and the role of crowd analysis [J]. Sustainable Cities and Society, 2021, 68: 102761. https://www.sciencedirect.com/science/article/pii/S2210670721000226

[2] ZHANG J, ZHENG Y, QI D. Deep spatio-temporal residual networks for citywide crowd flows prediction [C] // Proceedings of the AAAI Conference on Artificial Intelligence. 2017: 1655-1661. https://doi.org/10.1609/aaai.v31i1.10735

[3] SINDAGI V A, PATEL V M. A survey of recent advances in CNN-based single image crowd counting and density estimation [J]. Pattern Recognition Letters, 2018, 107: 3-16. https://www.sciencedirect.com/science/article/pii/S0167865517302111

[4] HELBING D, FARKAS I, VICSEK T. Simulating dynamical features of escape panic [J]. Nature, 2000, 407(6803): 487-490. https://doi.org/10.1038/35035023

[5] WANG B, LIU H, SAMARAS D, et al. Distribution matching for crowd counting [C] // Advances in Neural Information Processing Systems (NeurIPS). 2020: 159-171. https://doi.org/10.48550/arXiv.2009.13077

[6] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity [J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612. https://doi.org/10.1109/TIP.2003.819861

[7] DALAL N, TRIGGS B. Histograms of oriented gradients for human detection [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2005: 886-893. https://doi.org/10.1109/CVPR.2005.177

[8] IDREES H, SALEEMI I, SEIBERT C, et al. Multi-source multi-scale counting in extremely dense crowd images [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2013: 2547-2554. https://ieeexplore.ieee.org/document/6615976

[9] LEMPITSKY V, ZISSERMAN A. Learning to count objects in images [C] // Advances in Neural Information Processing Systems (NeurIPS). 2010: 1324-1332. https://papers.nips.cc/paper/2010/hash/819f46e52c25763a55cc642422644317-Abstract.html

[10] ZHANG Y, ZHOU D, CHEN S, et al. Single-image crowd counting via multi-column convolutional neural network [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016: 589-597.

[11] Li Y, Zhang X, Chen D. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes [C] // Proceedings of the IEEE conference on computer vision and pattern recognition. Washington: IEEE Computer Society, 2018: 1091-1100. https://doi.org/10.1109/CVPR.2018.00120

[12] CHENG Z Q, LI J X, DAI Z, et al. Learning spatial awareness to improve crowd counting [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 6152-6161. https://openaccess.thecvf.com/content_ICCV_2019/html/Cheng_Learning_Spatial_Awareness_to_Improve_Crowd_Counting_ICCV_2019_paper.html

[13] CAO X, WANG Z, ZHAO Y, et al. Scale aggregation network for accurate and efficient crowd counting [C] // Proceedings of the European Conference on Computer Vision (ECCV). 2018: 734-750. https://doi.org/10.1007/978-3-030-01228-1_45

[14] CHEN L, ZHANG H, XIAO J, et al. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017: 5659-5667. https://openaccess.thecvf.com/content_cvpr_2017/html/Chen_SCA-CNN_Spatial_and_CVPR_2017_paper.html

[15] MA Z, WEI X, HONG X, et al. Bayesian loss for crowd count estimation with point supervision [C] // Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). 2019: 6142-6151. https://openaccess.thecvf.com/content_ICCV_2019/html/Ma_Bayesian_Loss_for_Crowd_Count_Estimation_With_Point_Supervision_ICCV_2019_paper.html