智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于多尺度金字塔Transformer的人群计数方法

张少乐, 雷涛, 王营博, 周强, 薛明园, 赵伟强

引用本文:
张少乐,雷涛,王营博,周强,薛明园,赵伟强. 基于多尺度金字塔Transformer的人群计数方法[J]. 智能系统学报, 2024, 19(1): 67–78.
ZHANG Shaole, LEI Tao, WANG Yingbo, et al. A crowd counting network based on multi-scale pyramid Transformer[J]. *CAAI Transactions on Intelligent Systems*, 2024, 19(1): 67–78.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202304044

DOI: 10.11992/tis.202304044
网络出版地址: https://link.cnki.net/urlid/23.1538.TP.20240102.1709.008

基于多尺度金字塔 Transformer 的人群计数方法

张少乐[1], 雷涛[2,3], 王营博[2], 周强[1], 薛明园[2], 赵伟强[4]

(1. 陕西科技大学 电气与控制工程学院, 陕西 西安 710021; 2. 陕西科技大学 电子信息与人工智能学院, 陕西 西安 710021; 3. 陕西科技大学 陕西省人工智能联合实验室, 陕西 西安 710021; 4. 中电科西北集团有限公司西安分公司, 陕西 西安 710065)

摘 要: 针对密集人群场景中背景复杂、目标尺度变化较大导致人群计数精度较低的问题, 本文提出一种基于多尺度金字塔 Transformer 的人群计数方法 (multi-scale pyramid transformer network, MSPT-Net)。在特征提取阶段设计了一种基于深度可分离自注意力的金字塔 Transformer 主干网络结构, 该网络结构能有效捕获图像的局部和全局信息, 从而有效解决人群密度图像背景复杂导致计数精度低的问题; 设计了一种特征金字塔融合模块及多尺度感受野的回归头, 实现了密集人群图像浅层细节特征和深层语义特征的高效融合, 增强了网络对不同尺度目标的捕获能力; 采用深度监督的训练方法在 3 个公开数据集上对提出的方法进行验证。实验结果表明, 本文方法在全监督与弱监督学习策略中, 与目前主流的人群计数方法相比, 实现了更高精度的人群计数, 克服了主流方法对背景复杂、目标尺度变化大的密集人群图像计数精度低的问题, 同时本文方法保持着更小的参数量与计算量。

关键词: 密集人群; 人群计数; 多尺度; 金字塔; Transformer; 自注意力; 密度图; 深度监督
中图分类号: TP391.41 **文献标志码:** A **文章编号:** 1673−4785(2024)01−0067−12

中文引用格式: 张少乐, 雷涛, 王营博, 等. 基于多尺度金字塔 Transformer 的人群计数方法 [J]. 智能系统学报, 2024, 19(1): 67–78.
英文引用格式: ZHANG Shaole, LEI Tao, WANG Yingbo, et al. A crowd counting network based on multi-scale pyramid Transformer[J]. CAAI transactions on intelligent systems, 2024, 19(1): 67–78.

A crowd counting network based on multi-scale pyramid Transformer

ZHANG Shaole[1], LEI Tao[2,3], WANG Yingbo[2], ZHOU Qiang[1], XUE Mingyuan[2], ZHAO Weiqiang[4]

(1. School of Electrical and Control Engineering, Shaanxi University of Science and Technology, Xi'an 710021, China; 2. School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China; 3. Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China; 4. China Electronics Technology Group Corporation Northwest Group Corporation Xi'an Branch, Xi'an 710065, China)

Abstract: A crowd counting network based on multi-scale pyramid Transformer (MSPT-Net) is proposed to address the problem of low accuracy in crowd counting in dense crowd scenes caused by complex backgrounds and large target scale variations. A pyramid transformer backbone network structure based on depth separable self-attention is designed in the feature extraction phase to effectively capture local as well as global information of the image, thereby effectively addressing the problem of low counting accuracy in crowd density images caused by complex backgrounds. A feature pyramid fusion module and a regression head with multi-scale receptive fields are designed to efficiently integrate shallow detail features and deep semantic features in dense crowd scenes, enhancing the network's ability to capture targets of different scales. Lastly, the proposed model is validated using a deep supervision training method on three publicly available datasets. The experimental results show that the proposed MSPT-Net achieves higher crowd counting accuracy in the fully supervised and weakly supervised learning strategies as compared to mainstream crowd counting networks, overcoming the issue of low counting accuracy in dense crowd images with complex backgrounds and significant changes in target scales. At the same time, the method in this paper keeps the parameter number and calculation amount smaller.

Keywords: dense crowd; crowd counting; multi-scale; pyramid; Transformer; self-attention; density map; deep supervision

随着城市化进程的不断推进和人口数量的持续增长, 城市公共场所的人口密度也不断提高, 而这些人口密集场所的管理和安全问题成为城市

收稿日期: 2023−04−30. **网络出版日期:** 2024−01−03.
基金项目: 国家自然科学基金项目 (62271296, 62201334); 陕西省重点研发计划项目 (2021ZDLGY08-07); 陕西省杰出青年科学基金项目 (2021JC-47).
通信作者: 雷涛. E-mail: leitao@sust.edu.cn.

管理者和公共安全机构的头号难题[1]。为解决这些问题，人群计数应运而生。

人群计数是计算机视觉领域中一个重要的研究方向，其主要应用于公共场所人群数量的实时监测和分析，从而实现对人流的预测、规划和管理。这项技术可以广泛应用于人口密集的公共场所，如商场、机场、火车站、地铁站等。人群计数可以实现对人员数量、流量和分布的实时监测和分析，帮助城市管理部门进行人流预测和规划，减少拥堵和安全隐患，提高城市管理效率[2]。人群计数的另一个重要应用领域是应急救援。在自然灾害和人为事故发生时，人群计数可以快速计算出现场人员的数量和位置，协助应急救援部门对灾害现场进行快速响应和救援。在冠状病毒（COVID-19）疫情期间，人群计数也被广泛应用于公共场所的人员数量管控，帮助政府和公共场所管理者控制人员流动，减少病毒的传播。

然而，人群计数技术在实际应用中仍然面临很多挑战，其中最为艰巨的挑战为：1) 计数算法对复杂场景的适应性低，在人群密集的公共场所中，人员数量众多、行为复杂，场景复杂多变；2) 场景中的透视畸变、光照变化、复杂的人群形态等问题也会对计数结果产生影响，高鲁棒性的计数算法才能适应复杂多变的场景环境；3) 图像特征利用率低，不同尺度的特征语义信息不同，为了尽可能挖掘当前图像蕴含的细节信息，需要增强融合图像中不同尺度的特征。

为了更好地应对密集人群计数任务中图像场景中背景复杂、目标尺度变化较大导致人群计数精度较低的问题，本文提出一种基于多尺度金字塔 Transformer 的人群计数方法 (multi-scale pyramid Transformer network, MSPT-Net)。针对人群场景背景复杂的问题，设计一个新颖的 Transformer 结构，该结构通过深度可分离自注意力模块同时提取局部与全局特征信息，增强网络特征表达能力。针对目标尺度变化大、遮挡等导致图像特征利用率低，不同尺度的特征语义信息不同的问题，设计特征金字塔模块提取浅层和深层地特征，设计多尺度感受野的回归头模块，通过多尺度的空洞卷积和金字塔平均池化增强特征捕获能力，并能够有效地融合不同尺度的特征进而预测密度图。在训练过程中，使用深度监督，对网络中间的每个阶段添加了额外监督损失，让网络更好的收敛。同时，针对于人群计数任务中人头位置标注繁琐问题，本文使用弱监督学习策略，与目前主流的人群计数方法相比较在计数方面接近最先进的水平。

1　相关工作

早期的人群计数方法通常使用基于检测的方法[3]，如滑动窗口和头部检测等。这些方法虽然简单易行，但往往受到高密度、目标重叠和背景干扰等因素的影响，导致计数结果不准确[4]。为了解决这些问题，一些基于回归的方法被提出，如 Chen 等[5] 提出的累积属性支持向量回归机 (cumulative attribute for support vactor regression, CA-SVR)，该方法的主要思想是学习一种特征到人群数量的映射，通过建立群体数与人群密度图之间的回归模型，可以很容易地获得总体计数结果。虽然基于回归的方法在整体上提升了计数性能，但是它们忽略了图像中的空间信息，只得到一个计数结果，缺乏可靠性和可解释性。

随着计算机视觉技术的发展，基于深度学习的人群计数方法逐渐成为主流。其中，基于密度图的方法是目前最为流行和有效的方法之一，该方法通过将输入图像利用几何自适应高斯核估计图像中每个人头的大小并转换为密度图，然后通过回归密度图中的像素数量估计人群数量[6]，可以在复杂场景下获得更精确的计数结果，如图 1 所示。

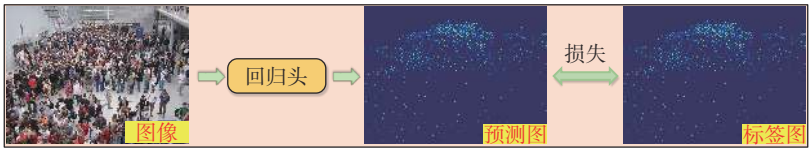


图 1　基于密度图回归的方法

Fig. 1　Methods based on density map regression

伴随着卷积神经网络在计算机视觉领域中迅速发展，使得许多新的人群计数方法[7-13] 被提出。其中，针对多尺度变化问题，Zhang 等[7] 提出多列卷积神经网络 (multi-column convolutional neural network, MCNN)，使用大小不同的卷积核提取人群的多尺度信息。针对多列结构难以训练、结构冗余、输入图片需要根据密度分级等原因，Li 等[8] 提出空洞卷积神经网络模型 (congested scene recognition network, CSRNet)，首次将空洞卷积用于人群计数，通过保持分辨率的同时扩大感受野，

保留了更多图像细节信息。然而, 这些方法不能有效处理复杂场景和多尺度问题, 导致人群计数的准确性和稳定性较低。为了解决这些问题, 研究人员又提出许多改进的方法, 如多尺度卷积神经网络[9]、注意力机制[10]、可形变卷积网络[11]、生成对抗网络[12] 等。此外, 还有学者[13-14] 研究了基于密度图的后处理技术, 如弱监督学习[13]、多任务学习[14] 等, 为准确估计人群数量奠定了基础。

随着计算机视觉技术的不断发展, 研究人员发现, 使用简单的卷积神经网络模型难以处理复杂场景下的密集人群计数。因为在这些复杂场景下, 人群的位置相互接近, 互相遮挡, 使得人数估计变得非常困难。因此, 如何处理密集人群计数问题成为研究人员关注的重点。注意力机制[15] 的本质是从关注全部到关注重点, 具有参数少、速度快和效果好等优点。在人群计数方法中引入注意力机制使得人员遮挡、人群分布不均匀等问题得到解决。Liu 等[16] 提出一种融合注意力机制的可变形卷积网络 (attention-injective deformable convolutional network for crowd, ADCrowdNet), 通过引入注意力机制强调了人群区域, 并使用可变形卷积保证了在高度拥挤场景中密度图的准确性。最新研究[17-18] 侧重于设计不同的注意力机制, 以关注全局背景下尺度和密度的变化。同时, 一些研究通过优化新的图像增强和损失函数提高计数的准确性[19-20]。然而, 这些方法通常需要足够的数据和经验, 而且网络结构设计较为复杂, 没有显著的改进效果。

近年来, Transformer[21] 已成为深度学习领域的一个热门模型, 在各种任务中都取得了非常优秀的效果。Transformer 是一种基于自注意力机制的神经网络, 可以处理输入序列中的全局依赖关系。值得关注的是, 视觉自注意力模型 (vision Transformer, ViT)[22] 的出现引起了计算机视觉领域的广泛关注, 并在人群计数任务中显示出良好的结果。一种基于 Transformer 的弱监督人群计数模型 (weakly-supervised crowd counting with Transformer, TransCrowd)[13] 被提出, 这是第 1 个基于 Transformer 的人群计数研究, 它从基于 Transformer 序列计数的角度重新构造了人群计数问题, 通过使用 Transformer 的自注意机制来有效地提取语义人群信息。实验结果表明, 该模型在密集人群计数任务中取得了非常好的性能。虽然 Transformer 在提取上下文信息具有强大的优势, 但是在局部信息获取上仍然不足, 所以如何获取图像的局部信息及全局信息进而提高计数精度仍是人群计数任务的研究重点。综上所述, Transformer 在人群计数领域的应用已经得到了广泛的研究和应用, 并且不断有新的方法和改进被提出, 为人群计数的研究和实践带来了新的思路和可能性。

2　多尺度金字塔 Transformer 网络

本文提出的多尺度金字塔 Transformer 网络结构如图 2 所示, 网络整体包含块嵌入 (patch embedding)、Transformer 模块、特征金字塔模块 (feature pyramid module, FPM)、多尺度感受野的回归头模块 4 个模块。

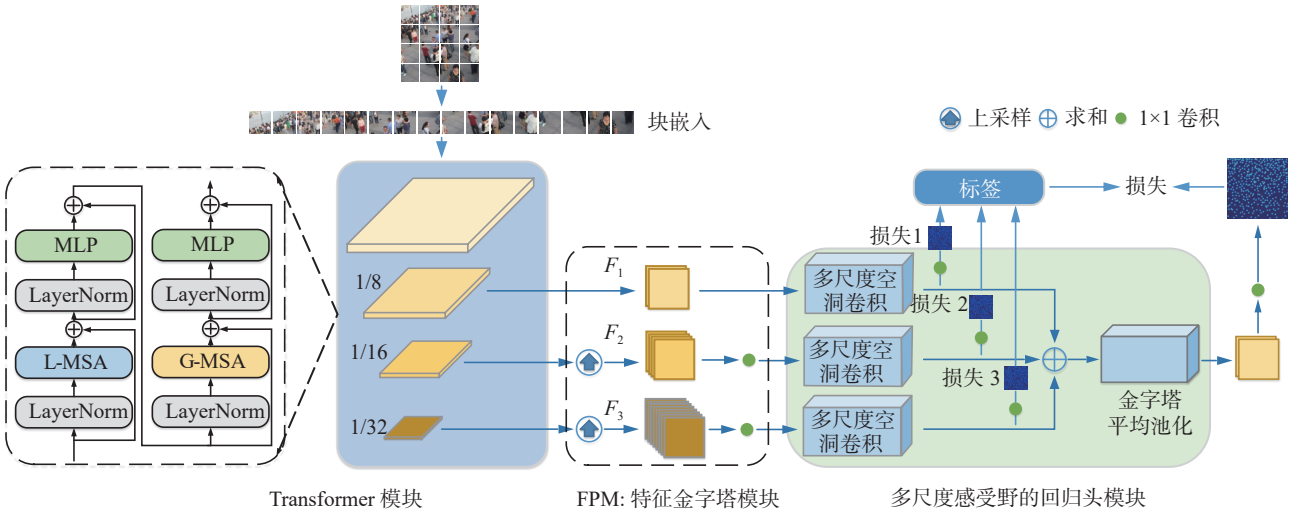


图 2　多尺度金字塔 Transformer 网络结构

Fig. 2　Inter-scale pyramid Transformer network architecture

输入图片经过块嵌入模块分割成固定大小的图像块。输出被展平为一维向量序列输入 Transformer 模块提取全局特征。每个阶段的一维序列都会被重塑为二维的特征图并经过特征金字塔模

块上采样到相同的分辨率，以方便之后的相加操作。三分支的特征图都通过一个多尺度空洞卷积模块 (multi-scale atrous convolution, MAC) 增强特征并回归出中间密度图，为之后的深度监督做准备，这 3 个分支的损失函数分配较小的权重。相加之后的特征图经过一个金字塔平均池化模块 (pyramid average pooling, PAP)，回归出最终的密度图，该分支的损失函数位置分配较大的权重。

2.1　金字塔 Transformer 骨干网络

目前在人群计数任务中，为解决人群场景背景复杂问题，大多数工作还是基于卷积神经网络，但由于卷积神经网络卷积核的感受野固定，其难以捕获全局上下文信息。Transformer 是一种基于自注意力机制的神经网络，可以处理输入序列中的全局依赖关系，但其缺少对局部信息提取的能力。基于此，本文将 Transformer 块堆叠设计成局部注意力和全局注意力交替使用的形式，使其同时拥有局部和全局的感受野进而捕获局部关系和进行全局上下文建模。同时，为了充分利用目标多尺度信息，设计了多层的特征金字塔结构，利用不同阶段的特征图进行浅层和深层特征的信息交互，增强网络的表达能力。

2.1.1　特征图预处理

在人群计数任务中，图像通常是高维数据，具有很大的空间结构，同时也具有丰富的语义信息。为了对图像进行有效地处理，需要将这些高维的数据通过块嵌入模块转换为低维的向量表示，以便能够在网络中进行处理和分析。输入特征图经过块嵌入模块，其将输入图像在进入 Transformer 模块时转换为一维向量。假设输入特征图大小为 $\boldsymbol{x}\in\mathbf{R}^{H\times W\times 3}$，其中 H、W、3 分别代表高度、宽度和通道数。通过块嵌入层将图片裁剪为 HW/K^2 个尺寸相同的块，每个块的大小为 $K\times K\times 3$，本文将这个二维序列展平为一维向量 $\boldsymbol{i}\in\mathbf{R}^{N\times D}$，其中 $N=HW/K^2$，$D=K\times K\times 3$。块嵌入模块可以将输入的图像尺寸统一化，并且在进行特征提取时可以避免过拟合，因为每个图像块的特征表示都是独立的。此外，块嵌入模块在进行特征提取时考虑了图像的局部信息，因此对于网络可以提供更好的性能。

2.1.2　Transformer 模块

由于输入图像通常是高分辨率，这使得 Transformer 全局注意力策略的计算量会随着图像的分辨率成二次方增长，在密集预测任务中会呈现高计算复杂度，如给定 $h\times w$ 的输入，维数为 C 的 Transformer 计算复杂度为 $\Omega\left(4hwC^2+2(hw)^2C\right)$，由于 $h\times w$ 比较大，所以本文旨在将输入的平方降为线性的，从而大幅度降低计算量。因为金字塔视觉自注意力模型 (pyramid vision transformer, PVT)[23] 能够很好地用于密集的像素级预测任务中，所以本文对基于 PVT 的 Transformer 模块的全局注意力策略进行了优化改进，提出深度可分离卷积自注意力 (depthwise separable self-attention, DSSA) 模块，该模块能提取局部与全局特征信息，增强了网络特征表达能力，同时降低了网络的复杂度，如图 3 所示。从自注意力机制的效率和感受野角度出发，将 Transformer 块堆叠设计成局部注意力和全局注意力交替使用的形式，拥有局部和全局的感受野可以捕捉短距离和长距离的关系。

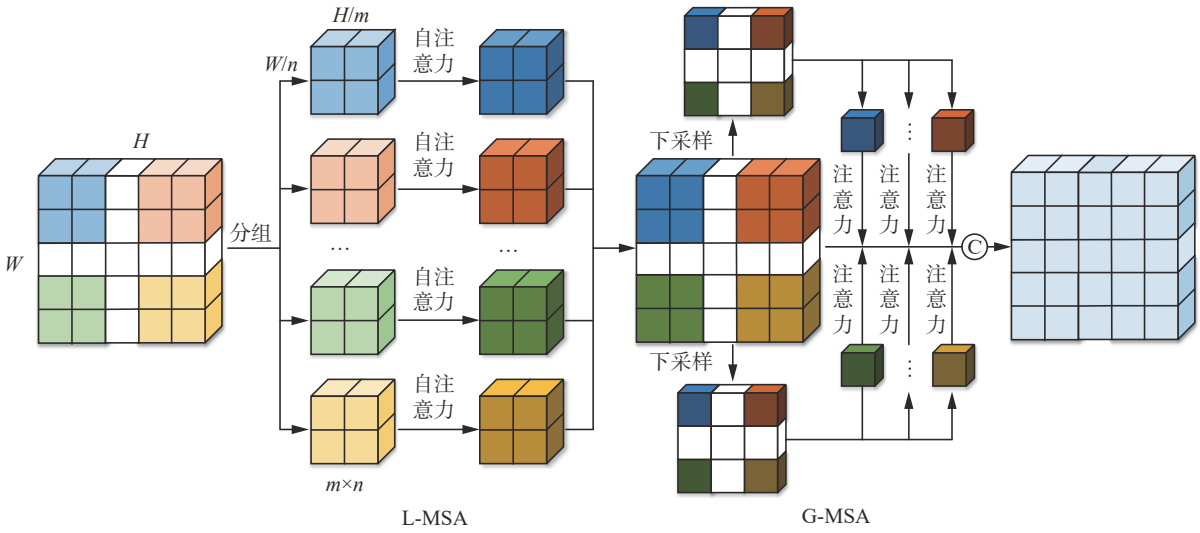


图 3　深度可分离卷积自注意力模块

Fig. 3　Depth-separable convolution self-attention module

受深度可分离卷积[24] 中的分组设计的影响, 首先将第l层Z_{l-1}的输入一维向量进行重塑变成$h\times w$的二维特征图, 其次在局部窗口内对每个特征图进行空间分组, 计算局部分组自注意力 (locally-grouped multi-head self-attention, L-MSA), 最后计算全局下采样自注意力 (global-downsampling multi-head self-attention, G-MSA)。具体而言, 在 L-MSA 阶段, 二维特征图被平均划分为$M\times M$大小的子窗口, 仅在子窗口内部进行自注意力计算, 计算量会大大减少, L-MSA 的计算复杂度和时间复杂度为

$$\Omega(\text{L-MSA}) = 4hwC^2 + 2M^2hwC \tag{1}$$

$$O(\text{L-SMA}) = M^2\times M^2\times \frac{hw}{M^2}C = M^2hwC \tag{2}$$

VIT[22] 中 MSA 模块的计算复杂度和时间复杂度为

$$\Omega(\text{MSA}) = 4hwC^2 + 2(hw)^2C \tag{3}$$

$$O(\text{MSA}) = (hw)^2C \tag{4}$$

式中: $h\times w$为特征图的高度和宽度, C为特征图的深度, M为每个窗口的大小。

相比于 VIT 中的自注意力, 计算复杂度和时间复杂度从输入的平方降为线性的, 并且每个子窗口之间没有交互通信, 以此获得局部特征信息。

在 G-MSA 阶段, 从每个子窗口提取一个低维度特征作为各个窗口的表征, 然后基于这个表征再与其他原始窗口的每个像素点进行注意力操作, 相当于自注意力中键的作用, 一直重复此过程直到所有窗口间都进行了注意力计算。在这个过程中通过下采样每个子窗口使其包含整个窗口的所有信息, 同时用这个特征与每个窗口进行交互, 能够实现全局交互。并且 G-MSA 过程中的K、V是在缩小特征的基础上计算的, 但Q是全局的, 因此注意力仍然可以恢复到全局, 这种做法显著减少了计算量。计算复杂度和时间复杂度为

$$\Omega(\text{G-MSA}) = 2hwC^2 + 2\frac{hw}{M^2}C^2 + 2hwC \tag{5}$$

$$O(\text{G-MSA}) = 1\times 1\times hw\times \frac{hw}{M^2}\times C = \left(\frac{hw}{M}\right)^2 C \tag{6}$$

同时网络结构中也交叉了必要的多层感知器 (multi-layer perceptron, MLP), 归一化层 (layer normalization, LN) 和残差连接等。如图 1 Transforme 模块所示, Z_{l-1}、Z_l、Z_l'、Z_l''、Z_l'''分别表示每层的输出。

整个 Transformer 块可以表示为

$$\begin{cases} Z_l' = \text{L-MSA}(\text{LN}(Z_{l-1})) + Z_{l-1} \\ Z_l'' = \text{MLP}(\text{LN}(Z_l')) + Z_l' \\ Z_l''' = \text{G-MSA}(\text{LN}(Z_l'')) + Z_l'' \\ Z_l = \text{MLP}(\text{LN}(Z_l''')) + Z_l''' \end{cases} \tag{7}$$

该注意力机制是对图像特征的空间维度进行分组, 分别计算各局部空间的自注意力, 再利用全局自注意力机制对分组注意力结果进行融合, 这种机制计算效率更好, 性能更优, 同时易于部署, 便于实际应用。

2.1.3　特征金字塔模块

尽管 Transformer 能够提取全局特征, 但是提取出的深层特征图在通过上采样后, 仍会缺少很多细节信息。此外, 深层次的特征虽然包含了丰富的语义信息, 但是由于分辨率低通常较模糊, 难以区分不同对象的边界, 这使得人群计数网络很难准确获取人群的位置信息。相比之下, 浅层的特征虽然语义信息较少, 但由于分辨率高, 可以准确地包含物体的位置信息[25]。为了充分利用深层的语义信息和浅层的细节信息, 设计了一个特征金字塔模块, 如图 2 中的特征金字塔模块 (feature pyramid module, FPM) 所示。它能够有效地融合深层和浅层的特征, 增强多尺度信息交互, 从而提高小目标检测的精度。具体而言, 将所有阶段的特征图上采样到输入图像的 1/8 大小, 方便与其他方法进行比较[8]。

本文设计的基于深度可分离自注意力的金字塔 Transformer 主干网络结构, 不但能有效捕获图像的局部和全局信息, 而且在不显著损失性能的情况下能够降低计算量。设计的特征金字塔结构实现了密集人群图像浅层细节特征和深层语义特征的高效融合, 能够补充深层特征图在通过上采样后缺失的细节信息。

2.2　多尺度感受野的回归头

因为本文的 Transformer 骨干和特征金字塔模块已经捕获了足够的局部和全局信息, 所以, 只需要使用一个简单的回归头来回归精确的密度图。但针对人群计数场景目标尺度变化的问题, 引入空洞卷积[26] 不仅可以减少参数, 而且还可以通过调整卷积核的空洞大小增加其感受野来实现不同尺度的特征提取, 从而提高模型准确性。考虑到不同阶段语义信息的丢失, 本文采用多尺度空洞卷积模块和金字塔平均池化模块进行特征增强, 以进一步提高模型的性能。

2.2.1　多尺度空洞卷积模块

由于人群计数场景复杂, 实际图像存在人群密度明显变化及图像严重遮挡等问题。虽然改进的 Transformer 模块可以同时提取全局和局部信息, 但是不同阶段的语义信息内容丰富程度也不同, 通过特征金字塔模块上采样后对于人群密度和不同场景信息可能会丢失, 为了加强网络对于

不同尺度目标捕获的特征并且精确回归密度图，本文设计了一个多尺度空洞卷积模块，该模块通过并行堆叠具有不同膨胀率的空洞卷积层，增大了感受野，更适用于人群密度估计任务，如图 4 所示，它能够用于挖掘图像细节信息，加强对特征的提取。

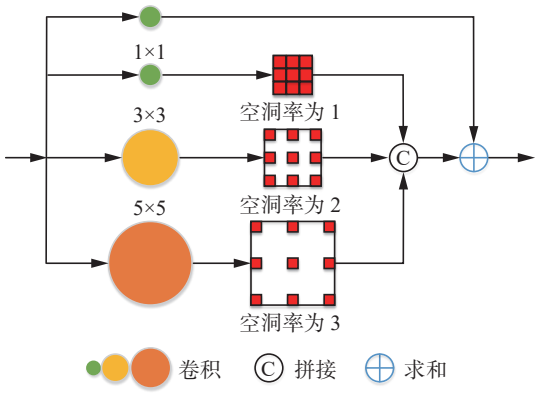


图 4　多尺度空洞卷积模块

Fig. 4　Multi-scale dilated convolutional module

具体而言，多尺度空洞卷积模块包含 3 列路径和 1×1 路径，每列由单个卷积层和空洞卷积层组成。本文将相应的卷积核和空洞率设置得尽可能小，目的是适应小规模对象的人群计数场景。每个卷积层之后是归一化 (batch normalization, BN) 层和 ReLU 激活函数。每列输出的特征图和快捷路径相加以利用它们的多尺度特征，使得多尺度空洞卷积模块更轻量化的同时获得更好的性能。

2.2.2　金字塔平均池化模块

在通过多尺度空洞卷积模块后，已经获得了充分的结构和空间信息。但在人群计数任务中，同一个场景中往往存在不同尺度的人头大小，此时需要应对人头尺度变化并减少解码部分的信息损失。为防止丢掉太多高维语义信息，为此设计了金字塔平均池化模块，用于融合不同尺度的特征图。该模块通过不同层级输出不同尺度的特征图，以保持全局特征的权重，并通过 1×1 卷积核处理每个金字塔层级后的输出。通过双线性插值直接对低维特征图进行上采样，使其与原始特征图尺度相同。将不同层级的特征图拼接为最终的金字塔，平均池化全局特征。通过一层 1×1 卷积生成最终的密度图。该模块能够应对人头尺度变化，帮助提高人群计数的准确性，如图 5 所示，它可以满足不同尺度的输入图像，在不丢失信息的情况下可以提取每个空间位置的特征。

本文采用多尺度空洞卷积模块和金字塔平均池化模块进行特征增强并作为回归头，是因为其简单且高效。在特征图相同情况下，空洞卷积可以得到更大的感受野，更大的感受野可以提高在密集任务中小物体识别分割的效果，从而获得更加密集的数据。金字塔平均池化可以提取不同尺寸的空间特征信息，提升网络对于空间布局和人头尺度变化的鲁棒性。

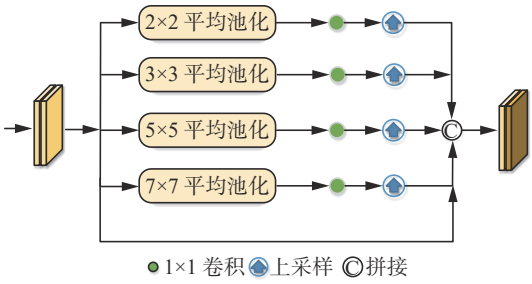


图 5　金字塔平均池化模块

Fig. 5　Pyramid average pool module

2.3　深度监督

2.3.1　全监督损失函数

本文设计一种基于基于分布匹配 (distribution matching, DM-Count) 的损失函数[20]。在人群计数过程中，使用高斯平滑每一个注释点会损害泛化性能，因此采用分布匹配方法进行人群计数。DM-Count 使用计数损失、最优传输损失和总变量损失 3 个项来制定损失函数。

人群计数的目标是使$\|\hat{Z}\|_1$尽可能接近$\|Z\|_1$，计数损失被定义为它们之间的绝对差：

$$L_1(Z,\hat{Z}) = \left|\|Z\|_1 - \|\hat{Z}\|_1\right| \tag{8}$$

式中：Z和$\hat{Z}$分别代表真值和预测密度图，$\|\cdot\|_1$代表L_1范数。

最优传输是指将一种概率分布转化为另一种概率分布的最优成本。本文使用最优传输来度量预测密度图和地面真实密度图之间的相似性，因为最优传输损失可使模型具有强大的拟合能力。最优传输成本也可以通过双重公式计算为[20]

$$\begin{aligned} &W(\mu,\nu) = \max_{\alpha,\beta\in\mathbf{R}^n} \langle\alpha,\mu\rangle + \langle\beta,\nu\rangle \\ &\text{s.t.}\alpha_i + \beta_j \leqslant c(x_i,y_j), \forall i,j \end{aligned} \tag{9}$$

式中：μ和ν分别为定义在向量空间上的 2 组点x和y上的 2 个概率度量。

最优传输损失L_{OT}定义为

$$L_{\text{OT}}(Z,\hat{Z}) = W\left(\frac{Z}{\|Z\|_1},\frac{\hat{Z}}{\|\hat{Z}\|_1}\right) = \left\langle\alpha^*,\frac{Z}{\|Z\|_1}\right\rangle + \left\langle\beta^*,\frac{\hat{Z}}{\|\hat{Z}\|_1}\right\rangle \tag{10}$$

式中：W为式 (9) 代表的最优传输成本，α^*、β^*为式 (9) 的解。

DM-Count 中的总变量损失使用了原始的地面真相的头部注释，不够平滑，无法建立一个强有力的人物形象。特别是在一些稀疏的场景中，

人群的规模更大, 用像素表示一个人是不合理的。为了解决这个问题, 采用均方误差损失(即 L_2 损失)正则化预测与真值之间的差距。L_2 损失为

$$L_2(Z,\hat{Z})=\frac{1}{N}\sum_{i=1}^{N}(Z-\hat{Z})^2 \tag{11}$$

式中 N 为样本数。

因此, 通过计数损失L_1、最优传输损失L_{OT}及均方误差损失L_2的加权和制定的损失函数L_M为

$$L_M(Z,\hat{Z})=L_1(Z,\hat{Z})+\lambda_1 L_{\mathrm{OT}}(Z,\hat{Z})+\lambda_2 L_2(Z,\hat{Z}) \tag{12}$$

式中: λ_1和λ_2为损失系数, 在 DM-Count[20] 中分别设置为 0.01 和 0.1。

为了解决训练过程中的梯度下降和梯度消失的问题, 本文设计了一个辅助损失函数增强训练[27]。由于不同深度的卷积层对最终损失的贡献不同, 一般来说, 网络越深, 感知范围越广, 提取到的特征表达能力越强, 因此较深层的特征图输出在某种程度上比浅层的输出更重要。根据这个原则及相关实验的支持, 将辅助监督损失均设置为均方误差损失 (L_2 损失), 每个阶段的辅助损失如图 2 所示, 并将不同深度的辅助损失函数的权重设置为可学习的$n_i, i\in(1,2,3)$, 所以最终的损失函数为

$$L_{\mathrm{oss}}=L_M+\sum_{i=1}^{3}n_i L_2 \tag{13}$$

2.3.2 弱监督损失函数

为了训练方便, 本文对 3 个多尺度分支和多尺度融合网络设计了 4 个$s_{\mathrm{moothL_1}}$损失, 因为不同图像中人群的数量变化很大, 并且 L_1、L_2 损失单独对离散值很敏感, 存在梯度爆炸问题。因为网络的不同分支在检测不同尺寸的头部时将具有不同的精度, 所以损失函数中使用任务特定的可学习权重$\omega_i, i\in(1,2,3,4)$, 本文的弱监督损失函数定义为

$$L_w=\sum_{i=1}^{4}\omega_i s_{\mathrm{moothL_1}}(Z,\hat{Z}) \tag{14}$$

3 实验结果与分析

为了验证本文提出方法的有效性, 设计了一系列实验进行探讨分析, 验证所提出方法的有效性和优越性。在本节中, 首先说明实验细节、数据集和评估指标, 其次展示所提出的 MSPT-Net 在 3 个数据集上的实验结果以及可视化结果, 最后在 ShanghaiTech 数据集[7] 进行消融研究。

3.1 实验细节

本实验在训练的过程中主要在 NVIDIA GeForce RTX 3 090 24 GB、Python3.7 和 PyTorch 1.7 的服务器上实现, 骨干采用的是基于 PVT 改进模型。ST_PartB 和 UCF-QNRF 的裁剪大小均为 512, ST_PartA 和 NWPU 裁剪大小均为 256。本文使用了 AdamW 优化算法并且训练批次大小为 8, 初始学习率设为 1×10^{-5}。

3.2 数据集

ShanghaiTech 数据集包含 1 198 张图片, 共标记了 330 165 个人头数量。数据集分为 Part A 和 Part B 2 部分。Part A 包含 482 张图像, 图像目标较为密集, 其中训练集 300 张图像, 测试集 182 张图像, 图像平均分辨率为 589×868。Part B 共包含 716 张图像, 图像目标较为稀疏, 其中训练集 400 张图像, 测试集 316 张图像, 图像平均分辨率为 768×1 024。

UCF-QNRF 数据集[28] 共包括 1 535 张人群图像, 其中训练集 1 201 张图像, 测试集 334 张图像。就注释数量而言, UCF-QNRF 是迄今为止最大的数据集, 可用于训练和评估大规模人群密集计数模型。该数据集的图像是从不同网站收集的, 这些图像在场景和图像大小方面都具有很大的多样性。其中训练集 1 201 张图像, 测试集 334 张图像, 图像平均分辨率为 2 013×2 902。

NWPU 数据集[29] 是目前最大的人群计数数据集, 由 5 109 张高分辨率图像和 213 万多条注释组成。图像中的人数范围从 0~20 033。此外, 在这个数据集中引入了负样本, 它指的是没有人或具有与人群场景相似纹理的图像。其中训练集 3 109 张图像, 验证集 500 张图像, 测试集 1 500 张图像, 图像平均分辨率为 2 311×3 383。

3.3 评价指标

采用平均绝对值误差 (mean absolute error, MAE) 和均方根误差 (root mean square error, MSE) 用于评价模型性能, 即

$$M_{\mathrm{AE}}=\frac{1}{N}\sum_{i=1}^{N}\left|Y_i-\hat{Y}_i\right| \tag{15}$$

$$M_{\mathrm{SE}}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|Y_i-\hat{Y}_i\right|^2} \tag{16}$$

式中: N 为来自测试集的样本数, Y_i 和 $\hat{Y}_i$ 分别为第 i 个测试图像中的预测数量和真实数量。通过对模型输出的人群密度图求和获得预测计数。对于 MAE 和 MSE, 如果值越小, 则测试样本越接近真实人数。

3.4 实验结果对比

为了评估所提出模型的有效性, 在 3 个公共的数据集上进行了实验验证, 包括 Shanghai-

Tech[7]、UCF-QNRF[28] 和 NWPU[29] 数据集。此外，在全监督方法上也与 CSRNet[8]、DM-Count[20]、SUA-Fully[30]、MFP-Net[18]、TransCrowd[13]、DLMP-Net[31]、SC2Net[32]、FIDTM[33] 等先进的网络模型进行比较。同时，本文在弱监督方法上也与 CCSLL[34]、MATT[35]、TransCrowd[13]、MPS[36] 进行比较。具体结果如表 1 所示，其中√表示选择方式，—表示未选择。

表 1　不同网络在 Shanghai Tech、UCF-QNRF、NWPU 数据集上的对比结果

Table 1　Comparison results of different networks on Shanghai Tech, UCF-QNRF, NWPU datasets

方法	训练方式		Part A		Part B		UCF-QNRF		NWPU	
	位置	数量	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN[7]	√	√	110.2	173.2	26.4	41.3	277.0	426.0	232.5	714.6
CSRNet[8]	√	√	68.2	115.0	10.6	16.0	121.3	208.0	190.6	491.4
DM-Count[20]	√	√	59.7	95.7	7.4	11.8	85.6	148.3	88.4	388.6
DM-Count[30]	√	√	66.9	125.6	12.3	17.9	119.2	213.3	105.8	445.3
MFP-Net[18]	√	√	65.5	112.5	8.7	13.8	112.0	190.7	90.3	458.0
TransCrowd[13]	√	√	66.1	105.1	9.3	16.1	97.2	168.5	117.7	451.0
DLMP-Net[31]	√	√	59.2	90.7	7.1	11.3	99.1	169.7	87.7	431.6
SC2Net[32]	√	√	58.9	97.7	6.9	11.4	98.5	174.5	89.7	348.9
FIDTM[33]	√	√	57.0	103.4	**6.9**	11.8	89.0	153.5	86.0	**312.5**
MSPT-Net(本文)	√	√	**53.1**	**88.3**	7.0	**11.1**	**82.5**	**139.4**	**73.9**	318.7
CCSLL[34]	—	√	104.6	145.2	12.3	21.2	133.4	216.4	164.5	532.1
MATT[35]	—	√	80.1	129.4	11.7	17.5	103.7	177.4	143.2	513.6
TransCrowd[13]	—	√	66.1	105.1	9.3	16.1	97.2	168.5	117.7	451.0
MPS[36]	—	√	71.4	110.7	9.6	15.0	~	~	~	~
MSPT-Net*(本文)	—	√	**65.5**	**97.1**	**7.8**	**12.8**	**94.3**	**162.7**	**94.3**	**412.5**

注：*表示弱监督方法。

在 ShanghaiTech 数据集上将本文方法与经典方法进行比较。在 Part A 部分，本文全监督方法与经典方法 CSRNet 相比，MAE 和 MSE 分别降低了 15.1 和 26.7，本文全监督方法与最新的方法 FIDTM 相比，MAE 和 MSE 分别降低了 3.9 和 15.1。在 Part B 部分，本文全监督方法与 CSRNet 相比，MAE 和 MSE 分别降低了 3.6 和 4.9，本文全监督方法与最新的方法 FIDTM 相比，MSE 降低了 0.7。本文弱监督方法甚至超越了一些全监督方法，本方法在 Part A 中获得 MAE 为 65.5、MSE 为 69.9 的优秀弱监督成绩，相比较于基准 TransCrowd 有着大幅度的提升。在 Part B 中，相较于最新弱监督方法 MPS，MAE 和 MSE 分别降低了 1.8 和 2.2。实验结果表明，本文方法相比于传统的卷积神经网络方法有着明显的优势。与主流弱监督方法相比，仍然有着很强的竞争力。

为了进一步验证本文所提模型的泛化能力，在 UCF-QNRF、NWPU 数据集上也进行了实验。在 UCF-QNRF 数据集上本文全监督方法与 CSRNet 相比，MAE 和 MSE 分别降低了 38.8 和 68.6，本文方法与最新方法 FIDTM 相比，MAE 和 MSE 分别降低了 6.5 和 14.1，本文方法再次优于其他方法。同时在弱监督方法上，本文方法相比于基线 TransCrowd 和最新弱监督方法 MPS 仍然有着大幅度的提升。这是因为本文设计的特征金字塔模块可以包含更多细节信息，有助于检测小物体。同时本文设计的多尺度空洞卷积模块与金字塔平均池化模块，可以更好地从 Transformer 中捕获多尺度特征和全局上下文信息来回归人群数量。

在 NWPU 数据集上全监督方法与 CSRNet 相比，MAE 和 MSE 分别降低了 116.7 和 172.7，本文方法与最新方法 FIDTM 相比，MAE 降低了 12.1。在弱监督方法上，本文方法有着更大的提升。实验结果表明，本文方法在处理大规模数据时表现出卓越的性能，同时相对于其他主流算法仍能够保持准确的计数结果和较高的鲁棒性。

3.5　模型轻量化分析

为了客观衡量所提出模型的轻量化，对模型参数 (Params) 和计算量 (FLOPs) 的定量对比进行了实验验证，为保证实验数据的合理性，本文在基于 Transformer 的人群计数框架上进行重点对比试验分析，同时，对 CNN 网络模型也进行了数

据对比以作参考。其中, Params 和 FLOPs 越小越有利于轻量检测, GFLOPs 是在 256×256 的输入规模下计算的。实验结果如表 2 所示。

表 2　不同模型轻量化指标对比

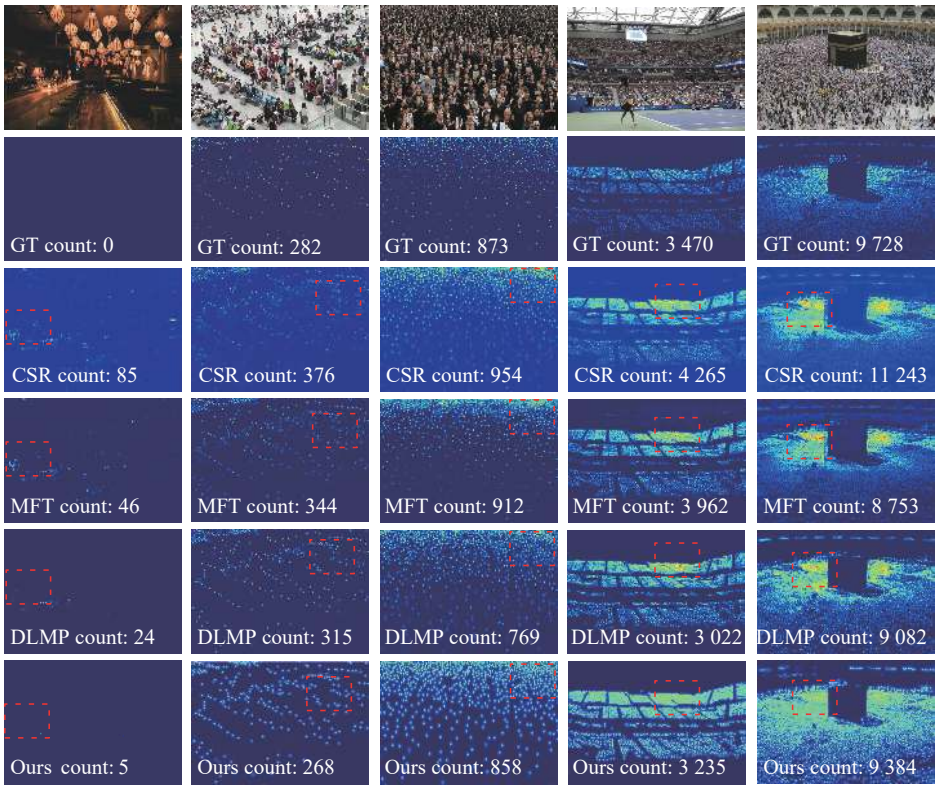
Table 2　Comparison of lightweight indexes of different models

方法	Param/MB	FLOPs/GB
MCNN[7]	**0.13**	56.21
CSRNet[8]	16.26	857.84
DM-Count[20]	21.5	60.8
TransCrowd[13]	86.8	49.3
PVT[23](backbone)	38.56	12.35
MSPT-Net(本文)	26.93	**9.63**

对表 2 中数据进行分析发现, 对于 MCNN、CSRnet、DM-Count 等均未使用 Transformer 的 CNN 网络模型, 参数量相对较低, 但 FLOPs 相对较高。本文方法的参数量为 26.93 MB, FLOPs 为 9.63 GB, 基准方法 TransCrowd 参数量为 86.8 MB, FLOPs 为 49.3 GB, 在性能提升的同时, 参数量以及计算量也有着显著的下降。

3.6　可视化

为了验证本文方法的有效性, 选择了 5 类有代表性的样本, 图 6 给出了在 NWPU 数据集使用不同方法上预测密度图的比较结果。在每张图片的右下角, 标注了预测密度图的计数结果。第 1 列是负样本, 样本信息中无人物出现, 其纹理信息与密集人群的纹理信息相似, 由于 CSRNet 直接融合所提取的特征而不进行区分, 导致如第 3 行第 1 列的图像所示的较差的预测结果。而本文提出的 MSPT-Net 使用金字塔 Transformer 去提取不同阶段的特征, 进而感知背景与细节信息, 可以抑制背景噪声, 提高模型的泛化能力。第 2、3 列为光照存在巨大差异的样本, 在光照不一的照明条件下, 本文模型仍实现了精细的预测结果, 这表明本文模型具有很强的鲁棒性。第 4 列为严重遮挡以及人头尺度不一的样本、第 5 列为高密集人群样本, 很明显, MSPT-Net 可以提供良好的预测结果, 由于 CSRNet 是单列网络没有融合多尺度信息、TransCrowd 只捕获全局信息丢失了局部信息导致预测图像人头丢失, 而本文提出的方法使用深度可分离自注意力模块能有效获取图像的局部和全局信息, 同时, 多尺度感受野的回归头模块能有效增强捕获的特征信息, 所以本文方法能够更好地解决人群计数面临的计数场景复杂、目标尺度变化较大等问题。这些可视化结果表明, 本文所提出的方法能够在不同场景、不同人群密度下有效地学习人群图像和人群密度图之间的映射关系, 表明本文方法具有很强的鲁棒性。



(a) 负样本　(b) 光照明亮　(c) 光照阴暗　(d) 严重遮挡　(e) 高密集

图 6　NWPU 数据集不同方法的部分可视化结果

Fig. 6　Partial visualization result plots of different methods for the NWPU dataset

3.7　消融实验

为了训练方便，将 Part A 中的图像随机裁剪成一些尺寸为 256×256 的小尺寸图像，Part B 中的图像随机裁剪成尺寸为 512×512 的图像。为了验证所设计的特征金字塔模块（FPM）、多尺度空洞卷积模块（MAC）和金字塔平均池化模块（PAP）的有效性，本文在 ShanghaiTech 数据集进行了消融实验，实验结果如表 3 和表 4 所示。

表 3　关于 FPM 不同阶段特征的消融实验

Table 3　Ablation experiments on the characteristics of different stages of FPM

方法			Part A		Part B	
F_1	F_2	F_3	MAE	MSE	MAE	MSE
√	—	—	79.6	123.8	9.4	14.8
—	√	—	58.1	96.6	7.9	12.5
—	—	√	56.7	92.6	7.4	11.8
√	√	—	57.2	95.6	7.8	12.3
√	—	√	55.6	91.5	7.2	11.6
—	√	√	54.3	90.4	7.2	11.5
√	√	√	**53.1**	**88.3**	**7.0**	**11.1**

表 4　关于 MAC 和 PAP 的消融实验

Table 4　Ablation experiments on MAC and PAP

方法	Part A		Part B	
	MAE	MSE	MAE	MSE
Baseline	56.8	94.4	7.8	13.1
Baseline+MAC	54.1	92.1	7.2	12.0
Baseline+PAP	55.2	92.9	7.4	12.3
MSPT-Net	**53.1**	**88.3**	**7.0**	**11.1**

表 3 为使用不同阶段的特征评估特征金字塔模块的性能。不同阶段的特征有着不同的语义特征，F_1、F_2、F_3分别为从浅到深 3 个阶段的特征。可以发现，当缺少其中任何一个阶段的特征时，性能都会下降，同时深层次的特征有着更多的语义特征，对人群计数也更为至关重要。聚合其他阶段的特征可以提供在深层阶段丢失的语义信息，从而提高网络对于不同密度场景的适应性和计数精度。结果表明，特征金字塔模块有助于融合图像深层和浅层语义特征信息，进而提高了人群计数的准确性。

表 4 为分别使用不同模块评估多尺度感受野的回归头中多尺度空洞卷积模块和金字塔平均池化模块的必要性。实验结果表明，在只有一种模块的情况下，模型精度会有所下降，进一步验证了多尺度空洞卷积模块和金字塔平均池化模块的必要性。同时，本文也对深度监督训练方式的辅助损失进行了验证。实验结果表明，每条支路的损失都对衡量模型起到正面作用，同时采用三分支不同权重的损失可提高网络性能，使其更好地收敛。结果如表 5 所示。

表 5　不同权重损失的消融实验

Table 5　Ablation experiments with different weight loss

损失	Part A		Part B	
	MAE	MSE	MAE	MSE
L_M	54.8	92.5	7.4	12.2
$L_M+n_1L_2$	54.5	91.6	7.3	12.0
$L_M+n_2L_2$	54.2	91.3	7.3	11.9
$L_M+n_3L_2$	53.8	90.4	7.1	11.5
$L_M+n_{1,2}L_2$	54.0	89.9	7.3	11.6
$L_M+n_{1,3}L_2$	53.8	89.5	7.2	11.5
$L_M+n_{2,3}L_2$	53.4	88.9	7.0	11.3
L_{oss}(MSPT-Net)	**53.1**	**88.3**	**7.0**	**11.1**

4　结束语

本文提出一种简单但高性能的基于多尺度金字塔 Transformer 人群计数方法（MSPT-Net），首先 MSPT-Net 能够同时捕获全局上下文信息和局部语义信息。其次为了融合深层和浅层语义特征信息，本文在网络中嵌入了特征金字塔模块，并通过多尺度的空洞卷积模块和金字塔平均池化模块增强特征来预测密度图。在网络的每个中间阶段都引入了额外的监督损失，以稳定训练过程。MSPT-Net 在多个公开数据集上表现出优异性能，其可以应用于人群计数领域，如人流量监控和城市规划等。未来，将继续改进我们的方法，以实现更轻量化的模型，使其可以迁移至移动设备中。同时使用无监督的学习策略，解决人群计数任务中人头位置标注繁琐等数据问题。

参考文献：

[1] LIU Hong, XU Bin, LU Dianjie, et al. A path planning approach for crowd evacuation in buildings based on improved artificial bee colony algorithm[J]. Applied soft computing, 2018, 68: 360–376.

[2] KHAN M A, MENOUAR H, HAMILA R. Revisiting crowd counting: state-of-the-art, trends, and future perspectives[J]. Image and vision computing, 2023, 129: 104597.

[3] DOLLAR P, WOJEK C, SCHIELE B, et al. Pedestrian detection: an evaluation of the state of the art[J]. IEEE transactions on pattern analysis and machine intelligence, 2011, 34(4): 743–761.

[4] 齐鹏宇, 王洪元, 张继, 等. 基于改进 FCOS 的拥挤行人

检测算法 [J]. 智能系统学报, 2021, 16(4): 811–818.
QI Pengyu, WANG Hongyuan, ZHANG Ji, et al. Crowded pedestrian detection algorithm based on improved FCOS[J]. CAAI transactions on intelligent systems, 2021, 16(4): 811–818.

[5] CHEN Ke, GONG Shaogang, XIANG Tao, et al. Cumulative attribute space for age and crowd density estimation[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2013: 2467−2474.

[6] LEMPITSKY V, ZISSERMAN A. Learning to count objects in images advances in neural information processing systems, 2010, 23: 1324–1332.

[7] ZHANG Yingying, ZHOU Desen, CHEN Siqin, et al. Single-image crowd counting via multi-column convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition . Piscataway: IEEE, 2016: 589−597.

[8] LI Yuhong, ZHANG Xiaofan, CHEN Deming. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1091−1100.

[9] SINDAGI V A, PATEL V M. Generating high-quality crowd density maps using contextual pyramid CNNs[C]// 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 1879−1888.

[10] HOSSAIN M, HOSSEINZADEH M, CHANDA O, et al. Crowd counting using scale-aware attention networks[C]// 2019 IEEE Winter Conference on Applications of Computer Vision . Piscataway : IEEE, 2019: 1280−1288.

[11] DAI Jifeng, QI Haozhi, XIONG Yuwen, et al. Deformable convolutional networks[C]//2017 IEEE International Conference on Computer Vision . Piscataway: IEEE, 2017: 764−773.

[12] SHEN Zan, XU Yi, NI Bingbing, et al. Crowd counting via adversarial cross-scale consistency pursuit[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 5245−5254.

[13] LIANG Dingkang, CHEN Xiwu, XU Wei, et al. TransCrowd: weakly-supervised crowd counting with transformers[J]. Science China information sciences, 2022, 65(6): 1–14.

[14] JIANG Xiaoheng, ZHANG Li, ZHANG Tianzhu, et al. Density-aware multi-task learning for crowd counting[J]. IEEE transactions on multimedia, 2020, 23: 443–453.

[15] CHAUDHARI S, MITHAL V, POLATKAN G, et al. An attentive survey of attention models[J]. ACM transactions on intelligent systems and technology, 2021, 12(5): 1–32.

[16] LIU Ning, LONG Yongchao, ZOU Changqing, et al. ADCrowdNet: an attention-injective deformable convolutional network for crowd understanding[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 3220−3229.

[17] JIANG Xiaoheng, ZHANG Li, XU Mingliang, et al. Attention scaling for crowd counting[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 4705−4714.

[18] LEI Tao, ZHANG Dong, WANG Risheng, et al. MFP-Net: multi-scale feature pyramid network for crowd counting[J]. IET image processing, 2021, 15(14): 3522–3533.

[19] LIAN Dongze, LI Jing, ZHENG Jia, et al. Density map regression guided detection network for RGB-D crowd counting and localization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1821−1830.

[20] WANG Boyu, LIU Huidong, SAMARAS D, et al. Distribution matching for crowd counting[EB/OL]. (2020−09−28)[2023−04−30]. 2009.13077. https://arxiv.org/abs/2009.13077.pdf.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000−6010.

[22] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[EB/OL]. (2020−10−22)[2023−04−30]. https://arxiv.org/abs/2010.11929.pdf.

[23] WANG Wenhai, XIE Enze, LI Xiang, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions[C]//2021 IEEE/CVF International Conference on Computer Vision . Piscataway: IEEE, 2022: 548−558.

[24] CHOLLET F. Xception: deep learning with depthwise separable convolutions[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition . Piscataway: IEEE, 2017: 1800−1807.

[25] 雷涛, 王洁, 薛丁华, 等. 差异特征融合的无监督 SAR 图像变化检测 [J]. 智能系统学报, 2021, 16(3): 595–604.
LEI Tao, WANG Jie, XUE Dinghua, et al. Unsupervised SAR image change detection based on difference feature fusion[J]. CAAI transactions on intelligent systems, 2021, 16(3): 595–604.

[26] YU F, KOLTUN V. Multi-scale context aggregation by

dilated convolutions[EB/OL]. (2015−11−23)[2023−04−30]. https://arxiv.org/abs/1511.07122.pdf.

[27] LEE Chenyu, XIE Saining, GALLAGHER P, et al. Deeply-supervised nets[EB/OL]. (2014−09−18)[2023−04−30]. https://arxiv.org/abs/1409.5185.pdf.

[28] IDREES H, TAYYAB M, ATHREY K, et al. Composition loss for counting, density map estimation and localization in dense crowds[C]//Computer Vision - ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part II. New York: ACM, 2018: 544−559.

[29] WANG Qi, GAO Junyu, LIN Wei, et al. NWPU-crowd: a large-scale benchmark for crowd counting and localization[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(6): 2141–2149.

[30] MENG Yanda, ZHANG Hongrun, ZHAO Yitian, et al. Spatial uncertainty-aware semi-supervised crowd counting[C]//2021 IEEE/CVF International Conference on Computer Vision . Piscataway: IEEE, 2022: 15529−15539.

[31] CHEN Qi, LEI Tao, GENG Xinzhe, et al. DLMP-net: a dynamic yet lightweight multi-pyramid network for crowd density estimation[C]//Chinese Conference on Pattern Recognition and Computer Vision . Cham: Springer, 2022: 27−39.

[32] LIANG Lanjun, ZHAO Huailin, ZHOU Fangbo, et al. SC2Net: scale-aware crowd counting network with pyramid dilated convolution[J]. Applied intelligence, 2023, 53(5): 5146–5159.

[33] LIANG Dingkang, XU Wei, ZHU Yingying, et al. Focal inverse distance transform maps for crowd localization[J]. IEEE transactions on multimedia, 2023, 25: 6040–6052.

[34] YANG Yifan, LI Guorong, WU Zhe, et al. Weakly-supervised crowd counting learns from sorting rather than locations[C]//Vedaldi A, Bischof H, Brox T, et al. European Conference on Computer Vision. Cham: Springer, 2020: 1−17.

[35] LEI Yinjie, LIU Yan, ZHANG Pingping, et al. Towards using count-level weak supervision for crowd counting[J]. Pattern recognition, 2021, 109: 107616.

[36] ZAND M, DAMIRCHI H, FARLEY A, et al. Multiscale crowd counting and localization by multitask point supervision[C]//ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing . Piscataway: IEEE, 2022: 1820−1824.

作者简介：

张少乐，硕士研究生，主要研究方向为计算机视觉、机器学习。E-mail：210612054@sust.edu.cn。

雷涛，教授，博士生导师，陕西科技大学电子信息与人工智能学院副院长、IEEE 高级会员，主要研究方向为计算机视觉、机器学习。主持国家自然科学基金项目 5 项、陕西省重点研发计划、中国博士后科学基金等 6 项，授权发明专利 15 项，获陕西省科学技术二等奖 1 项（自然科学奖）。发表学术论文 90 余篇。E-mail：leitao@sust.edu.cn。

王营博，讲师，主要研究方向为散射环境下图像复原与场景感知。参与国家自然科学基金面上项目、高分重大专项等项目 5 项，授权发明专利 8 项，授权软件著作权 1 项。发表学术论文 20 余篇。E-mail：wangyingbo@sust.edu.cn。