面向人群计数的 U 型 Transformer 模型

王 锐[1] 姚瑞玲[1] 席 茜[2] 张冬松[1] 毛凤翔[1]
WANG Rui YAO Ruiling XI Qian ZHANG Dongsong MAO Fengxiang

摘 要 在人群计数领域，可采用 U 型结构的全卷积神经网络模型将人群场景图映射为人群密度图。在映射的过程中，引入空间注意力机制和通道注意力机制，分别从空间维度和通道维度提取人群场景图的重要信息和抑制非重要信息。根据该思想，设计了一种基于通道和空间注意力机制的 U 型 Transformer 模型（SC U-Transformer）。SC U-Transformer 包含编码和解码过程，编码过程使用 Swin-Transformer 作为编码器，提取上下文特征并实现下采样；解码过程使用包含扩展图像块的对称 Swin-Transformer 模型，并添加了空间注意力模块和通道注意力模块，分别使模型更加关注前景和相关联的特征通道。根据 ShanghaiTech 数据集和 UCF_CC_50 的实验结果可知，SC U-Transformer 能够将人群场景图转换为高质量的人群密度图。

关键词 人群计数；人群场景图；人群密度图；Swin-Transformer；注意力机制

doi：10.3969/j.issn.1672-9528.2024.06.014

0 引言

世界人口增长速度加快，导致大规模的人群聚集现象频频出现，进而造成了一系列的踩踏事件。因此，及时准确地预测出其中的人群分布状况，指导人群控制，减少踩踏事件的发生是十分有意义的工作。

现有的基于深度学习的人群计数方法，主要使用深度神经网络（deep neural networks，DNN）将人群分布图映射为能反映人群分布情况的人群密度图。如 Gong 等人 [1] 设计了一种基于对抗学习的模型，其中包含了双层水平框架，能够实现任务驱动的数据和细粒度的特征对齐。Du 等人 [2] 通过引入密度分析的分层混合，设计了一种多尺度神经网络，该网络分层合并多尺度密度图以进行人群计数。Liang 等人 [3] 提出了一种无监督的视觉语言模型 CrowdCLIP，首次使用视觉语言知识来解决人群计数问题。

以上处理人群计数问题的方法均采用全卷积神经网络模型，虽然取得了较高的精度，但是模型性能仍然可以进一步增强。因此，本文试图构建一个新的模型，使用 Vision Transform 结构代替传统的 CNN 模型 [4]，并且能够较好地提取全局特征和深度的信息交互，同时还考虑了空间位置和通道信息。根据该观点，本文提出了一种面向人群计数的 U-Transformer 模型（SC U-Transformer）。该模型的启发来自 Swin-Unet 模型 [5]，它包含一个编码器和一个解码器。编码器使用拥有移位窗口的分层 Swin-Transformer 来充当，主要任务是提取上下文特征；解码器采用拥有图像块扩展层的对称 Swin-Transformer 进行上采样操作，主要任务是恢复特征图的空间分辨率。与 Swin-Unet 模型相似，SC U-Transformer 也使用两个 Swin-Transformer 模型分别作为编码器和解码器。每个人群场景图被分割成无重合的图像块，每个图像块则被视为一个标记，并输入 SC U-Transformer 的编码器中，以提取深层特征。编码器提取的深层特征再输入解码器中，由图像块扩展层实现上采样，并通过跳跃连接与来自编码器的多尺度特征融合，从而恢复特征图的空间分辨率，进一步进行分割预测。此外，在解码过程中还添加了 4 个空间注意模块和 4 个通道注意模块，分别对深层特征的感兴趣区域和重要通道进行关注。SC U-Transformer 与 Swin-Unet 的不同点主要表现在以下两个方面：（1）为了提取更深层次的特征，SC U-Transformer 的编码器比 Swin-Unet 的编码器多了两个 Swin-Transform 模块和 1 个 Patch Merging 层，解码器比 Swin-Unet 的解码器多了两个 Swin-Transform 模块和 1 个 Patch Expanding 层；（2）SC U-Transformer 的解码器中加入了 4 个空间注意模块和 4 个通道注意模块。

1 提出方法

由于卷积操作的局部性，全卷积神经网络在提取人群场景图的特征时，并不能较好提取全局特征和实现深度的信息交互，并且忽略了空间位置和通道信息。本节提出的 SC

1. 信阳学院 河南信阳 464000
2. 河南省信阳市南湾水库事务中心 河南信阳 464000

U-Transformer 采用 Swin-Transformer 结构 [7] 更好地提取全局特征和实现深度语义信息交互，并添加了注意力机制，进一步增强模型的分割性能。

1.1 SC U-Transformer

如图 1 所示，SC U-Transformer 的整体架构由编码器、解码器、跳跃连接组成。编码器中包含了划分图像块层（Patch Partition）、线性嵌入层（Linear Embedding）、Swin-Transformer 模块和图像块合并层（Patch Merging）。

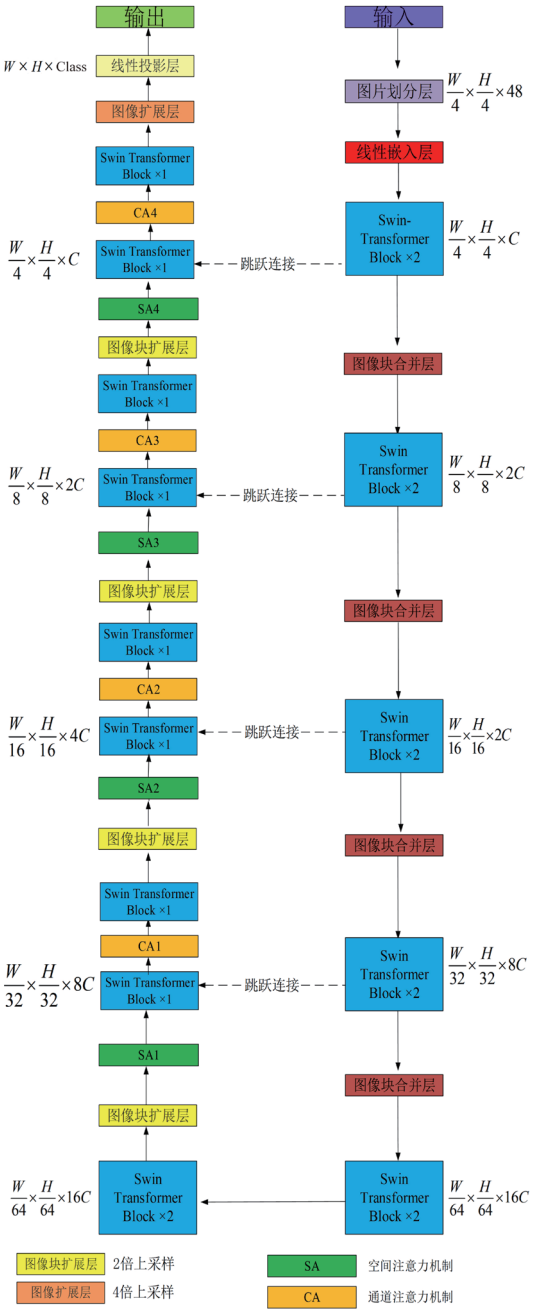

图 1 SC U-Transformer 结构图

人群场景图输入编码器后，首先被分割成大小为 4×4 的不重叠图像块并转换为序列嵌入，使每个图像块的特征维数变为 4×4×3 = 48。进一步，由线性嵌入层将图像块的特征维度投影到任意维度（设为 C）。投影后的图像块通过几个 Swin-Transformer 块和 Patch Merging 层来生成分级特征表示。其中，Patch Merging 层实现下采样和增加通道数，Swin-Transformer 块实现特征表示学习。解码器由 Swin-Transformer 块、空间注意力机制模块（SA）、通道注意力机制模块（CA）、图像块扩展层（Patch Expanding）和线性投影层（Linear Projection）组成。其中，空间注意力机制主要关注重要的像素点，消除非重要像素点的影响；通道注意力机制主要突出相关的特征通道，抑制不相关的通道；Patch Expanding 层实现上采样，将相邻维度的特征图的分辨率扩大 2 倍，最后一个 Patch Expanding 层进行 4 倍上采样 [9]，将特征图的分辨率恢复到输入特征图的分辨率大小，即 $W \times H$；Linear Projection 层将经过上采样的特征进行线性映射，输出分割结果。此外，所提取的上下文特征通过跳跃连接与编码器提取的多尺度特征融合，以弥补下采样造成的空间信息丢失。

1.2 Swin-Transformer 模块

图 2 展示了两个连续的 Swin-Transformer 模块。每个 Swin-Transformer 模块包含 LayerNorm（LN）层、多头自注意力模块、残差连接和由非线性激活函数 GELU 激活的 MLP 层。其中，第一个 Swin-Transformer 模块的多头自注意力模块采用基于窗口的多头自注意（W-MSA）模块，第二个 Swin-Transformer 模块的多头自注意力模块采用基于移位窗口的多头自注意（SW-MSA）模块。

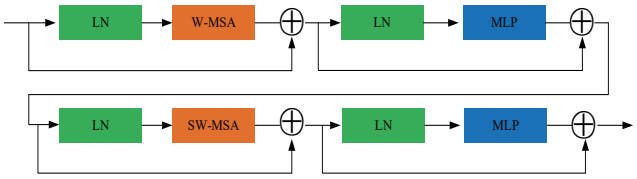

图 2 Swin Transformer 模块

1.3 Patch Merging 层

Patch Merging 层的主要功能是对输入的特征图进行上采样和扩展通道操作。输入特征图首先被划分为 4 个图像块，然后进行通道维度上的拼接。因此，经过划分与拼接处理后的特征图像块与输入图像块相比，分辨率将下降为原图的 1/2，通道个数扩大为输入图像块的 4 倍。拼接后的特征图再通过一个全连接层进行线性变换，使特征图的通道个数降低至输入特征图的 2 倍。所以，Patch Merging 层输出的特征图与输入特征图相比，高与宽缩小了 1/2，通道数扩大了 2 倍。

1.4 Patch Expanding 层

SC U-Transformer 模型的 Patch Expanding 层与 Swin-Unet 模型相同，其作用与 Patch Merging 层相反，将特征图输入到 Patch Expanding 层之后，宽与高均扩大了 2 倍，通道数减小 2 倍。具体过程为：首先，利用全连接层将输入特征

图进行线性映射，使其维数增加到原来的 2 倍；然后，使用重构操作将映射后的输入特征图的分辨率扩大 2 倍，并将通道数降低 4 倍。

1.5 跳跃链接

SC U-Transformer 模型使用的跳跃链接与 U-Net 模型相同，将来自编码器的低水平特征和解码器的高水平特征进行融合，进而减少由下采样造成的空间信息损失。

1.6 空间注意力机制

在解码器的每个 Patch Expanding 层后均连接一个空间注意力模块，分别为 SA_1、SA_2、SA_3 和 SA_4。每个 SA 均使用非局部区域块来获取所有像素之间的相互关联，从而更好地理解上下文信息，其具体结构如图 3 所示。

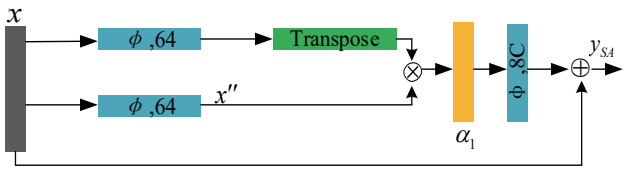

图 3 空间注意力模块 SA

以 SA_1 为例，其输入特征图尺寸为：$\frac{W}{32}\times\frac{H}{32}\times 8C$，其中 $8C$ 为输入通道数。首先使用两个平行卷积路径对输入特征进行降维，每个路径的卷积核个数为 64，尺寸为 1×1，分别可以得到两个形状相同的特征图 x' 和 x''；然后将两个特征图重构为二维矩阵，尺寸为$\frac{W}{32}\times\frac{H}{32}\times 64$。获得的空间注意系数图为：

$$\alpha_1=\sigma(x'^{\mathrm{T}}\cdot x'') \tag{1}$$

式中：T 表示矩阵的转置运算。$\alpha_1\in(0,1)^{\frac{W}{32}\times\frac{H}{32}\times\frac{W}{32}\times\frac{H}{32}}$是一个方阵，$\sigma$ 是一个逐行 Softmax 函数，使得方阵中每行之和等于 1.0。将 α_1 重构为 $\frac{WH}{32\times 32}\times 64$，并使用 $8C$ 个大小为 1*1 卷积核，即Φ^{8C}来改变 α_1 的通道数，使之与输入特征 x 的通道数相匹配。最后利用残差连接实现训练过程中的信息传播，因此得到 SA_1 的输出为：

$$y_{SA_1}=\Phi^{8C}(\alpha_1)+x \tag{2}$$

S_2 ～ S_4 模块的输出与 S_1 模块类似。

1.7 通道注意力机制

如图 1 所示，在 SC U-Transformer 的解码器中，两个 Swin-Transformer 模块之间添加了一个通道注意力机制（CA）。通道注意力机制能够突出相关的特征通道，抑制不相关的通道，从而将来自编码器的低级信息和来自解码器的语义信息的通道特征更好地利用。图 4 详细展示了 CA 模型的结构，设输入特征图 x 拥有 N 个通道，首先使用全局最大池化 $P_{\max}$ 来获取每个通道的全局信息，输出表示为 $P_{\max}(x)\in\mathbf{R}^{N\times1\times1}$。使用多层感知机（M LP）映射得到通道关注系数 $\beta\in[0,1]^{N\times1\times1}$。MLP 是一个全连接神经网络，其中第一层的神经元个数为 N/r，激活函数为 ReLU，第二层的神经元个数为 N，性能和计算成本的权衡 r 设置为 2。得到的通道关注系数 β 与输入特征图 x 相乘，从而为每个通道进行加权，得到的特征图记为 α。因此，通道注意力模块的输出结果为：

$$y_{CA}=x\cdot\alpha+x \tag{3}$$

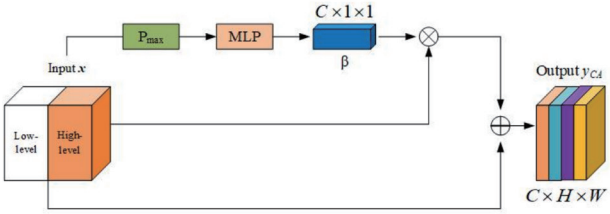

图 4 通道注意力机制

1.8 评价标准

平均绝对误差（MAE）和最小均方误差（MSE）是常用于评估人群计数方法的指标，其中 MAE 可以表示预测的准确性，MSE 表示预测的鲁棒性，具体定义如下：

$$\mathrm{MAE}=\frac{1}{N}\sum_{i=1}^{N}|y_i-pre_i| \tag{4}$$

$$\mathrm{MSE}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i-pre_i)^2} \tag{5}$$

式中：N 为测试图像的个数，y_i 和 pre_i 分别为第 i 组图像的实际计数和估计计数。本文使用这两个指标评估人群计数方法性能。

2 实验

2.1 实验设置

SC U-Transformer 是在 Python 3.10 和 PyTorch 1.4.0 环境下实现的。对于训练集上的样本，使用翻转和旋转操作进行数据增强。输入的图像块大小设置为 4。在训练过程中，批次大小为 1，使用动量为 0.9 和权重衰减为 1e-5 的 Adam 优化器优化模型。本次实验采用 ShanghaiTeac 数据集 [11] 和 UCC_CF50 数据集进行模型的训练和测试。

2.2 消融实验

为了探究不同因素对模型性能的影响，本文利用 ShangHaiTech 数据集进行了消融实验，主要研究空间注意力和通道注意力对模型的影响。表 1 展示了无注意力机制的 U-Transformer、添加空间注意力的 U-Transformer（S U-Transformer）、添加通道注意力的 U-Transformer（C U-Transformer），以及同时添加空间注意力和通道注意力的 U-Transformert（SC U-Transformer）。实验结果表明，注意力机制能够有效提高 U-Transformer 模型的分割性能。具体表现为：（1）添加空间注意力机制的 U-Transformer 的 MAE 和 MSE 比未添加注意力机制的 U-Transformer 分别降低了 1.3 和 2.9；（2）添加通道注意力机制的 U-Transformer

的 MAE 和 MSE 比未添加注意力机制的 U-Transformer 分别降低了 2.5 和 2.6；同时添加通道注意力机制和通道注意力机制的 U-Transformer 的 MAE 和 MSE 比未添加注意力机制的 U-Transformer 降低了 4.9 和 4.1。

表 1 注意力机制的影响

模型	MAE	MSE
U-Transformer	60.1	100.2
S U-Transformer	56.8	97.3
C U-Transformer	55.6	97.6
SC U-Transformer	53.2	96.1

2.3 ShangHaitech 数据集

ShangHaitech 数据集是文献 [6] 的数据集，其规模较大，包含了 Part_A 和 Part_B 两部分，每一部分又各自包含训练集和测试集。Part_A 和 Part_B 总共有图片 1198 张，标记的人头数目为 330 165。其中，Part_A 是在网络上随机抽取的高密度人群图，共有 482 张图片，平均每张图片 501 人，人数最多的图片中有 3139 人，人数最少的图片有 33 人；Part_B 是在上海街头抓拍的图片，相比于 Part_A 中的人数密度较稀疏，人数最少的图片仅有 9 人，最多的有 578 人。

表 2 给出了 SC U-Transformer 模型在 ShangHaitech 数据集上的性能。从表 2 可以看出，SC U-Transformer 模型与目前较先进的模型比性能较好，尤其在 Part_B 上 MAE 与 MSE 均低于其他模型。

表 2 ShangHaitech 数据集上的误差

方法	Part_A		Part_B	
	MAE	MSE	MAE	MSE
DKPNet[7]	55.6	91.0	6.6	10.9
UEPNet[8]	54.6	91.1	6.4	10.9
D2CNet[9]	59.6	100.7	6.7	10.7
S-DCNet[10]	59.8	100.0	6.8	11.5
HMoDE+REL	54.4	**87.4**	6.2	9.8
SC U-Transformer	**53.2**	96.1	**6.0**	**9.4**

3 结语

本文提出一种基于空间注意力和通道注意力机制的 U 型 Transformer 模型（SC U-Transformer），以提高人群图映射密度图的准确性。SC U-Transformer 由编码器、解码器和跳跃连接组成。编码器包含多个具有移位窗口的分层 Swin Transformer 和减小图片尺寸的 Patch Emerge 层，提取上下文特征；解码器包含多个 Swin Transformer 模型和使图片尺寸增大的 Patch Expanding 层，并添加了空间注意力模块和通道注意力模块。本文在 ShangHaitech 数据集上的实验结果表明，SC U-Transformer 模型与现有的模型相比，计数的误差更低且鲁棒性更强。

参考文献：

[1]SHEN J, SHAN S, JIAN Y, et al.Bi-level alignment for cross-domain crowd counting[EB/OL].(2022-05-12)[2024-02-23]. https://doi.org/10.48550/arXiv.2205.05844.

[2]ZHI P, MIAO J, JIAN K, et al.Redesigning multi-scale neural network for crowd counting[J].IEEE transactions on image processing, 2023,32:3664-3678.

[3]DING K, JIA H, ZHI K, et al.CrowdCLIP:unsupervised crowd counting via vision-language model[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR).Piscataway:IEEE,2023:2893-2903.

[4]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)[2024-02-13]. https://doi.org/10.48550/arXiv.2010.11929.

[5]CAO H, WANG Y, CHEN J, et al.Swin-Unet:Unet-Like pure transformer for medical image segmentation[C]//Computer Vision – ECCV 2022 Workshops,Part 3. Cham: Springer, 2023: 205-218.

[6]ZHANG Y, ZHOU D, CHEN S, et al.Single-image crowd counting via multi-column convolutional neural network[C]//29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway:IEEE,2016:589-597.

[7]CHEN B, YAN Z, LI K, et al.Variational attention:propagating domain-specific knowledge for multidomain learning in crowd counting[EB/OL].(2021-08-18)[2024-02-25].https://doi.org/10.48550/arXiv.2108.08023.

[8]WANG C, SONG Q, ZHANG B, et al. Uniformity in heterogeneity:diving deep into count interval partition for crowd counting[C]//2021 IEEE/CVF International Conference on Computer Vision,[v.1].Piscataway:IEEE,2021:3214-3222.

[9]CHENG J, XIONG H, CAO Z, et al.Decoupled two-stage crowd counting and beyond[C]//IEEE Transactions on Image Processing.Piscataway:IEEE,2021:2862-2875.

[10]XIONG H, YAO A.Discrete-constrained regression for local counting models[C]//Computer Vision-ECCV 2022,p.24. Cham: Springer, 2022:621-636.

[11]HAROON I, IMRAN S, CODY S, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2013: 2547-2554.

【作者简介】

王锐（1995—），通信作者（email：ruiwangjsj@163.com），男，河南信阳人，助教，硕士，研究方向：机器学习。

（收稿日期：2024-03-21）