第 58 卷 第 20 期/2021 年 10 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

多特征信息融合的人群密度估计方法

孟月波[1,2]，陈宣润[1]，刘光辉[1*]，徐胜军[1,2]

[1]西安建筑科技大学信息与控制工程学院，陕西 西安 710055；

[2]人工智能与数字经济广东省实验室(广州)，广东 广州 510000

摘要 人群密度估计在智能安全防范领域具有重要的应用价值。针对人群密度估计在二维图像中视角变化呈现较大差异、特征空间信息丢失、尺度特征和人群特征提取困难等问题，提出了一种多特征信息融合的人群密度估计方法。该方法通过注意力机制引导的空间注意力透视(Perspective of spatial attention，PSA)方法，对图像多视角信息进行了有效信息编码，获取了特征图的空间全局上下文信息，弱化了视角变化带来的影响；而后通过多尺度信息聚合(Multi-Scale Information Aggregation，MSIA)方法，利用多尺度非对称卷积与不同膨胀率的空洞卷积进行了有效融合，获取了较为全面的图像尺度及特征信息。最终通过细致语义特征嵌入融合的方式，补充了高层特征图的空间信息及低层特征图的语义信息，并使上下文信息与尺度信息相互补充，提高了模型的准确度与鲁棒性。采用 ShanghaiTech、Mall、Worldexpo'10 数据集进行了实验验证，实验结果表明，所提方法的性能较其他对比方法有一定的提升。

关键词 图像处理；卷积神经网络；人群密度；全局上下文信息；语义嵌入

中图分类号 O436 **文献标志码** A **doi**：10.3788/LOP202158.2010021

Crowd Density Estimation Method Based on Multi-Feature Information Fusion

Meng Yuebo[1,2], Chen Xuanrun[1], Liu Guanghui[1*], Xu Shengjun[1,2]

[1] *College of Information and Control Engineering, Xi'an University of Architecture and Technology, Xi'an, Shaanxi* 710055, *China*;

[2] *Guangdong Artificial Intelligence and Digital Economy Laboratory (Guangzhou), Guangzhou, Guangdong* 510000, *China*

Abstract Crowd density estimation has important application value in the field of intelligent security prevention. A crowd density estimation method with multi-feature information fusion is proposed to address the problems of large difference in viewpoint change of two-dimensional images, loss of feature spatial information, and difficulties in scale feature and crowd feature extraction. The proposed method encodes the multi-view information of images through the attention mechanism-guided perspective of spatial attention (PSA) method to obtain the spatial global contextual information of the feature map and weaken the influence of viewpoint change. Through the multi-scale information aggregation (MSIA) method, the multi-scale asymmetric convolution and the null convolution with different expansion rates are effectively integrated to obtain more comprehensive image scale and feature information. Finally, the spatial information of the high-level feature map and the semantic information of the low-level feature map are complemented by the detailed semantic feature embedding fusion, and the contextual information and scale information complement each other to improve the accuracy and robustness of the model. The experimental validation is carried out using the ShanghaiTech, Mall, and Worldexpo'10 datasets, and the

收稿日期：2021-03-04；**修回日期**：2021-03-12；**录用日期**：2021-03-23

基金项目：国家自然科学基金面上项目(51678470)、陕西省自然科学基础研究计划面上项目(2020JM-473，2020JM-472)

通信作者：*guanghuil@163.com

experimental results show that the performance of the proposed method has been improved compared with those of other comparative methods.

Key words image processing; convolutional neural network; crowd density; global context information; semantic embedding

OCIS codes 100.4996; 100.2000; 100.3008

1 引 言

随着国民经济的迅猛发展及城市化进程的加快，城市人口数量急剧增加，由此带来的社会问题也不断增加，人们因各种原因可能聚集在不同的场景下，易造成交通拥堵、人员踩踏等不安全事故的发生[1]。因此，人群密度估计在视频监控、公共安全、城市规划等诸多领域具有较高的应用价值[2-3]。

针对图像中的各种人群计数问题，研究者提出了各种各样的方法[4-6]，目前，人群计数方法可以分为基于检测、基于回归、基于深度学习三类[7]。基于检测的人群计数方法主要是通过类似滑动窗口探测器检测图像中人员全身或者脸、头等局部位置，但此方法对于遮挡较多的人群，存在计算量大、精度较差等问题[8-11]。基于检测的人群计数方法难以处理人群之间逐渐升级的遮挡问题，因此基于回归的人群计数方法被提出[12-13]。基于回归的人群计数方法通过学习图像低级特征与人群数量之间的映射关系，建立回归模型，预测人群人数[14-15]。但由于仅采用整个图像特征，图像空间信息缺失，研究人员在映射过程中加入上下文空间信息以提升计数精度[16-18]，但因人群遮挡的升级、特征提取能力弱及空间信息弱等问题，该方法已无法满足人群计数任务的精度要求。

近年来，卷积神经网络(Convolutional Neural Network，CNN)具有对图像深层次特征出色的提取及学习能力，被应用于人群计数领域中[19-21]。Wang 等[22]首次将 CNN 模型应用于人群密度估计中，将 AlexNet[23]网络末端链接全连接层以输出人群计数结果，但该模型部署于新场景中时，计数结果及精度明显下降，鲁棒性较差。Zhang 等[24]首次提出一种多列卷积神经网络模型，通过多列 CNN 结构并行提取多尺度信息，但多列结构的每一列都具有相似的学习功能，视角的变化使得多列卷积核的大小难以适用一些视角情况。孟月波等[25]提出了一个编码-解码的网络模型，利用空间金字塔结构改善多尺度特征融合的质量，但该网络在特征融合时忽略了上下文信息的获取，无法有效应对视角变化带来的影响。Li 等[7]提出了一个单列卷积神经网络(CSRNet)，该模型在网络后端添加空洞卷积以扩大感受野，同时缩减网络参数，但其对空间信息以及深度特征的提取能力较差。左静等[8]利用不同扩张率的空洞卷积模拟尺度特征，减少模型参数量，但因扩张率差距较小，构建的尺度特征模块的提取能力较差。Liu 等[26]提出了一个名为 CAN 的可感知尺度上下文的网络，该网络通过学习每个特征对图像位置的重要性，结合多特征信息结果，从而获取尺度上下文信息。但其应用于稀疏和较复杂场景时，因背景干扰及特征提取能力等问题，可能会出现错误的预测。

由上述分析可知，上下文信息、多列结构为多尺度的学习提供了一种有效手段，在一定程度上可以解决视角变化的问题，但上述方法仍无法解决视角变化导致的全局上下文信息提取能力差、特征融合不充分及特征空间信息丢失等问题。基于此，本文提出一种多特征信息融合的人群密度估计方法，该网络由骨架网络、空间注意力透视(Perspective of spatial attention，PSA)、多尺度信息聚合(Multi-Scale Information Aggregation，MSIA)以及多特征信息融合网络等组成。该方法首先利用骨架网络输出结果，得到高层语义信息；其次通过空间注意力透视机制来聚合图像的空间全局上下文信息，将空间信息引入到高层网络中，同时通过多尺度非对称卷积与不同膨胀率的空洞卷积的组合，增强提取到的语义信息与尺度信息的表达能力；最后在多特征信息融合网络中，使用语义嵌入的方法，将空间信息引入到表达更强的高层语义信息中，将高层语义信息引入到低层空间信息中，增强特征表达，并且将尺度信息与空间全局上下文信息融合，以获取高质量的密度图，更准确地预估人群人数。

2 基于多特征信息融合的人群密度估计方法

2.1 密度图的制作

人群计数任务首先需要获得训练数据集的真值图像，真值密度图中特殊高亮部分代表人群分布密度的位置，积分值代表图像中的人数。在含有 N 个

人头标记点的图像 x 中，第 i 个坐标位置为 x_i 的人头标记点可以表示为

$$H(x)=\sum_{i=1}^{N}\delta(x-x_i), \tag{1}$$

式中：$\delta(\cdot)$为图像中人头位置的冲击函数。

在实际情况中，存在透视失真、像素与周边样本在不同场景区域中的尺度不一致等问题，因此为了准确估计密度，利用高斯卷积核 $\boldsymbol{G}_\sigma$[27] 对 $H(x)$进行卷积处理，可以得到密度函数 $F(x)$：

$$F(x)=\sum_{i=1}^{N}\delta(x-x_i)*\boldsymbol{G}_{\sigma_i(x)}, \tag{2}$$

式中：$\sigma_i(x)=\beta\bar{d}_i$，其中 $\bar{d}_i$ 是坐标 x_i 与其最近的 K 个人头之间的平均距离，且经大量实践验证，参数 $\beta=0.3$ 时生成的密度图质量最好[24]。

2.2　基于多特征信息融合的卷积神经网络结构

基于多特征信息融合的卷积神经网络结构包含基础骨架网络（VGG-16）、空间注意力透视、多尺度信息聚合以及多特征信息融合网络。首先利用基础骨架网络生成特征图，然后分别利用 PSA 和 MSIA，获取特征图的空间全局上下文信息及较为全面的图像尺度和特征信息，最终通过多特征信息融合网络的语义特征嵌入融合方式，网络上层将 MSIA 的结果与基础骨架网络的第 2、3 层的上采样结果融合后输入空洞卷积，补充低层特征图语义信息，网络下层将 PSA 与 MSIA 的结果融合后输入空洞卷积，补充高层特征图的空间信息，并使上下文信息与尺度信息相互补充，增强最终输出密度图的质量，提高模型的准确度与鲁棒性。

2.2.1　空间注意力透视

针对图像视角变化复杂引起的图像人员分布变化大、图像视角变化呈现较大差异、特征空间信息丢失等问题[28]，本文提出了一种空间注意力透视方法，利用注意力机制引导关注视角变化中的交并比（IOU）区域，减少背景噪声的影响，提升 PSA 中有效特征信息的提取能力，进而通过多角度的信息编码，聚合特征图的空间全局上下文信息，弱化视角变化带来的影响，提升输出密度图的质量。

PSA 结构如图 1 所示，由四个卷积类型（从左到右，从右到左，从上到下，从下到上）组成，分别处理四个方向，本文将聚合方向称作左（Left）、右（Right）、上（Up）、下（Down）四个方向。以 Left 方向为例，对卷积过程进行说明。F 为输入特征图，其大小为 $C\times H\times W$，其中 C 为上一卷积层的卷积核的个数，H 为特征图的高度，W 为特征图的宽度。将特征图 F 的宽度 W 均分为 N 等份，则特征图 F 可被分为 N 个大小为 $C\times H\times\frac{W}{N}$ 的特征块，用 F_N^i 表示第 i 个特征块，$i\in[1,N]$。Left 方向的卷积层由卷积核 c 与 ReLU 激活函数组成。进行 Left to Right 卷积运算：

$$D_N^i=\begin{cases}L(F_N^i), & i=1\\ L(F_N^i)+L(F_N^{i-1}), & i=2,3,\cdots,N\end{cases}, \tag{3}$$

式中：$L(\cdot)$为 Left 方向卷积层（Conv＋ReLU）运算。将特征块 F_N^1 送入 Left 方向卷积层后，生成一个与 F_N^1 同样大小的特征块，记作 D_N^1；将 D_N^1 与 F_N^2 相加送入 Left 方向卷积层得到 D_N^2；经过不断迭代，输出第 N 个特征块 D_N^N。最后，将 $D_N^1,\cdots,D_N^i,\cdots,D_N^N$ 连接起来，生成 Left 层输出的特征图 D，其大小为 $C\times H\times W$，与特征图 F 的尺寸一致。在 Left 方向卷积层图像经过最终融合后，生成的特征图分别进入空间定位网络。原始信号直接进入采样层

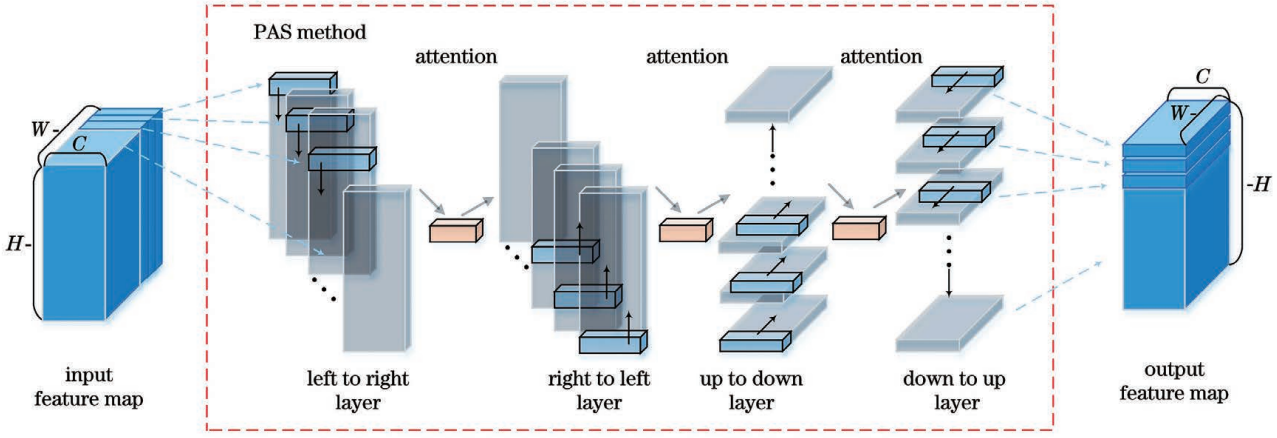

图 1　空间注意力透视结构

Fig. 1　Perspective structure of spatial attention

后，与定位信号进行融合，生成新的变化矩阵，变化矩阵与原始特征相乘之后得到新的包含注意力场景的特征，大小与原始特征大小一致，能够对上一层信号的关键信息进行关注，减少背景噪声的影响，提升有效信息的获取能力。其他三个方向 Right，Up，Down 的操作，除滑动方向不同外，计算与 Left 方向类似。

在人群场景中，图像视角变化的方向不一致，导致透视现象不同，进而使得不同方向的特征信息不一致；在 PSA 中，特征块之间相互融合，F_N^1 结果影响 F_N^{i+1} 结果，因此可将其中一个方向的输出结果视为图像特征的一个聚合表示。对于不同的列，由于计算顺序不一致，每列聚合信息不同，这与视角变化是一一对应的，因此视角变化可被视为空间全局上下文信息的变化。

2.2.2　多尺度信息聚合

由于人群图像存在较大的尺度变化，因此图像存在全局上下文信息提取困难、人群特征较难提取等问题。针对此问题本文设计了一种多尺度信息聚合方法，利用多尺度非对称卷积[29]与不同膨胀率的空洞卷积进行有效融合，进而使得网络拥有更佳的人员特征信息提取性能。如图 2 所示，多尺度信息聚合结构主要由两部分组成，一部分采用 1×1、3×3、5×5 三种卷积核尺寸的多分支非对称卷积层[30]，另一部分是对应不同卷积核尺寸的不同扩张率(rate)的空洞卷积。两者的对应关系是为了模拟神经学中[31]人眼的感受野与离心率的变化，重塑最终的表达，增强特征表达能力。多尺度卷积核负责捕获多尺度特征，不同扩张率的空洞卷积负责扩大群感受野，在减少参数量的同时，保留多尺度特征和图像上下文信息。

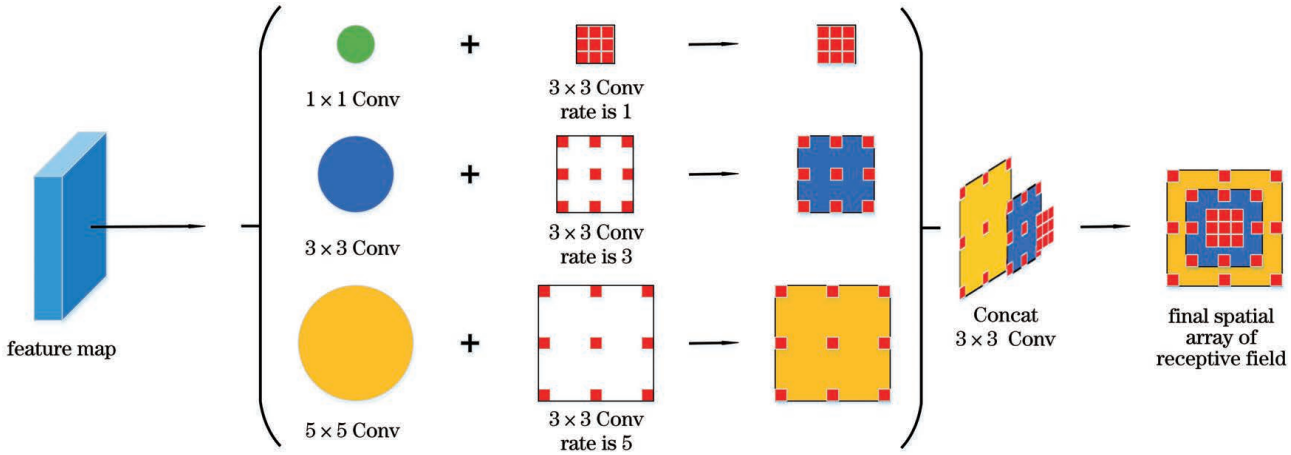

图 2　多尺度信息聚合结构

Fig. 2　Multi-scale information aggregation structure

非对称卷积增强信息熵较大位置的信息提取能力，从而增大平方卷积核，提升获取图像特征的能力。非对称卷积的本质是一种提升特征表达的方法，分为训练和部署两个阶段，训练阶段在于强化特征提取，实现效果提升；部署阶段在于卷积核融合，同时不会带来额外的计算量，如图 3 所示。

在训练阶段中，将现有网络每一个 3×3 卷积层替换成 3×3、3×1、1×3 三个卷积层，最后将三个卷积层的计算结果进行融合，得到卷积层的输出。在融合阶段中，主要是将三个卷积核进行融合，对输入先卷积后融合所得到的结果，与先融合卷积核再对输入进行卷积的结果是一样的。例如：

$$\boldsymbol{I}*\boldsymbol{K}^{(1)}+\boldsymbol{I}*\boldsymbol{K}^{(2)}=\boldsymbol{I}*(\boldsymbol{K}^{(1)}\oplus\boldsymbol{K}^{(2)}),\quad(4)$$

式中：$\boldsymbol{I}$ 表示一个二维矩阵输入；$\boldsymbol{K}^{(1)}$ 和 $\boldsymbol{K}^{(2)}$ 分别表示两个二维卷积，两个卷积核的高和宽一致。$\boldsymbol{K}^{(1)}$ 和 $\boldsymbol{I}$ 的卷积运算结果与 $\boldsymbol{K}^{(2)}$ 和 $\boldsymbol{I}$ 的卷积运算结果进行相加，或先将 $\boldsymbol{K}^{(1)}$ 和 $\boldsymbol{K}^{(2)}$ 逐点相加再与 $\boldsymbol{I}$ 卷积，两种方法最终得到的结果是一致的。在实际操作中，首先对卷积核进行额外的参数训练，利用训练后的卷积核参数初始化网络，使提取特征的能力更强，即训练阶段强化特征提取的能力，部署阶段通过融合卷积核，达到结构不改变、不增加计算量、提升特征提取能力的目的。

空洞卷积是在标准的卷积核中添加空洞，即空洞位置参数为 0，以扩大卷积核尺寸，增大感受野，减少参数运算量。空洞卷积算法定义为

$$(a*_l w)(i)=\sum_{k=1}^{k}a[i+kl]w[k],\quad(5)$$

式中：a 为输入特征；w 表示卷积核；k 表示卷积核尺寸；$w[k]$ 表示大小为 k 的卷积核；$a[i]$ 表示第 i 个输入；$*_l$ 表示空洞卷积运算；l 表示扩张率，描述卷积核处理数据时采样的步幅，调整 l 可自适应

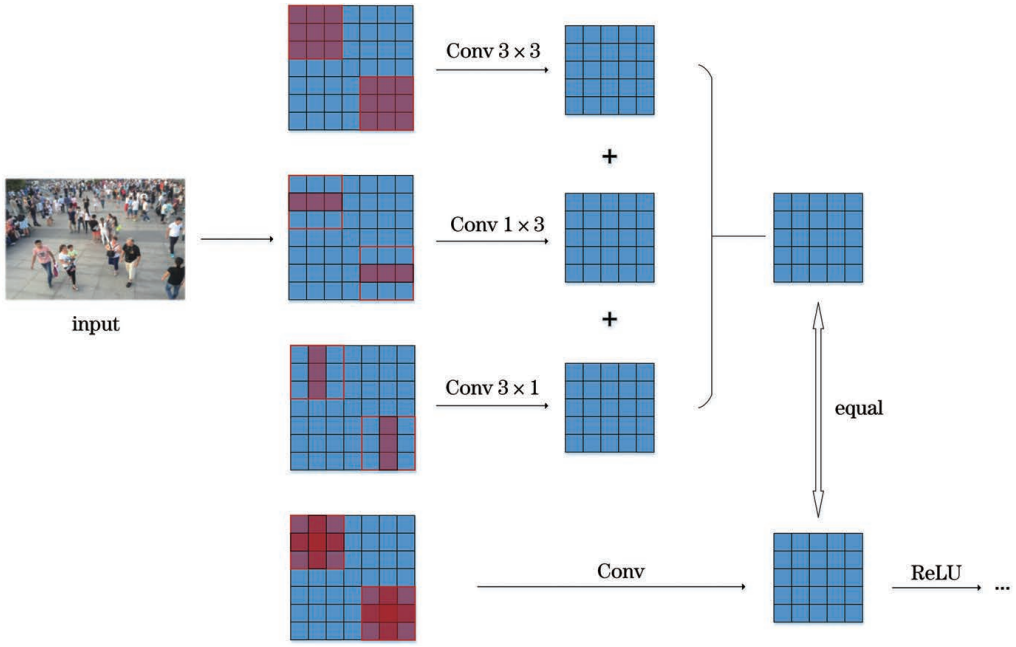

图 3　非对称卷积结构

Fig. 3　Asymmetric convolution structure

地调整感受野大小。

不同空洞率的空洞卷积如图 2 所示，其中 rate is 1 表示标准的 3×3 卷积，其感受野仅为 3×3；rate is 3 表示扩张率为 3 的 3×3 扩张卷积，其感受野可达 9×9；rate is 5 表示扩张率为 5 的 3×3 扩张卷积，其感受野可达 19×19。

2.2.3　多特征信息融合网络

随着卷积层网络的加深，卷积过程中会丢失一部分特征信息，降低最终生成的特征图质量。因此，通常使用特征融合的方式，将低层特征与高层特征进行拼接（Concat）或相加（Add），但因语义层次和空间信息的不充分融合，单纯的 Concat 或 Add 无法将特征进行有效融合，从而导致特征图质量下降，精度降低。文献[32]通过实验验证，通过在高层语义特征信息中引入空间信息，在低层特征中引入语义信息，可有效融合特征信息，提升输出特征的质量，如图 4 所示。虽然在特征融合时，一般都是利用低层的残差特征信息来填充空间细节，但是低层的特征信息包含的信息较少且含有噪声，对高层信息的补充作用不大。

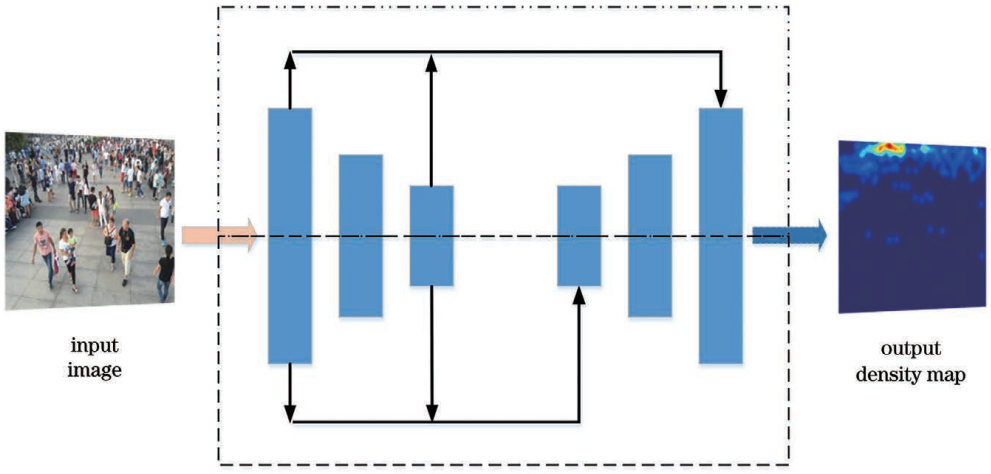

图 4　特征融合过程图

Fig. 4　Diagram of feature fusion process

基于此，本文利用语义嵌入的方式将高层富含丰富语义信息的多尺度特征信息与高层 PSA 输出的含有空间语义信息的特征相融合，不从低层特征中引入含有噪声的空间信息，可有效减少由

噪声带来的影响，增加网络可学习特征量。同时，为了增强输出密度图的质量，本文利用多尺度的结构，将低层特征与多尺度结构的特征进行自下向上的融合，增加低层特征的语义信息，改进低层的语义信息。

本文采用改进后的语义嵌入特征融合算法：

$$U_y = U_{\text{upsample}}(m_{y+1}) + F(m_y, m_{y+1}, \cdots, m_Y), \quad (6)$$

式中：U_y 为融合后第 y 层信息；$U_{\text{upsample}}(m_{y+1})$ 为对第 $y+1$ 层的特征进行上采样；m_y 为第 y 层的特征；y 为特征的层数；Y 为特征级别的数量；$F(m_y)$ 为低层特征函数。利用此方法从高级特征中引入更多的语义信息来提升特征融合的质量，详细设计如图 5 所示，其中 × 表示逐元素相乘，SE 表示语义嵌入。

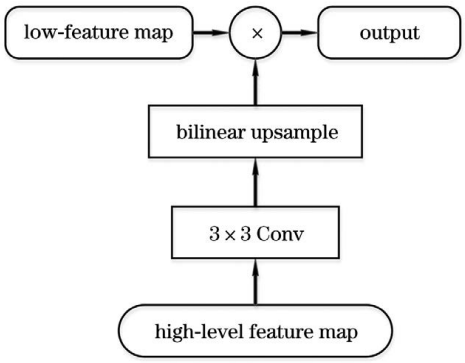

图 5　语义嵌入特征融合算法的结构

Fig. 5　Structural diagram of semantic embedding and feature fusion

图 6 所示为多特征信息融合网络结构。基于语义嵌入特征融合方法，本文在多特征信息融合卷积神经网络的下层（短线）特征融合过程中，在高层卷积后利用 PSA 提取图像的空间信息，并与低层自下而上包含丰富语义信息的高层多尺度特征进行融合，从而减少低层特征带来的噪声及层次信息影响；而在低层特征中，网络则希望加入更多的高级语义信息与低层特征相融合，因此本文在上层（点线）特征融合过程中，利用语义嵌入的方式将包含丰富语义信息的高层多尺度特征融合至上层特征中，改进低层的语义信息，丰富上层特征融合过程中的语义信息。综上可知，本文网络在下层中利用 PSA 补充高层的空间信息，在上层中利用语义嵌入的方式增加语义信息，从而两者结合输出特征图，最终提高输出密度图的质量。整体算法的程序框图如图 7 所示。

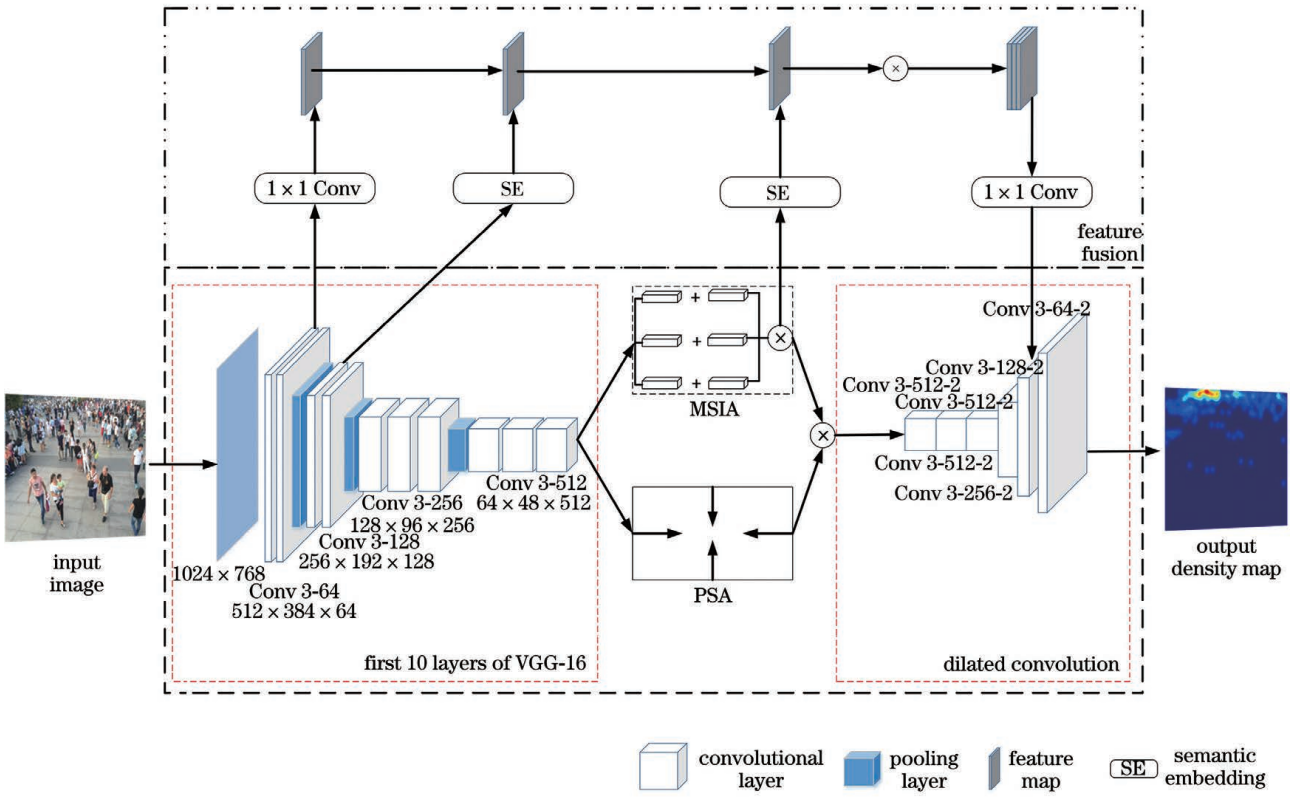

图 6　多特征信息融合网络结构

Fig. 6　Structural diagram of multi-feature information fusion network

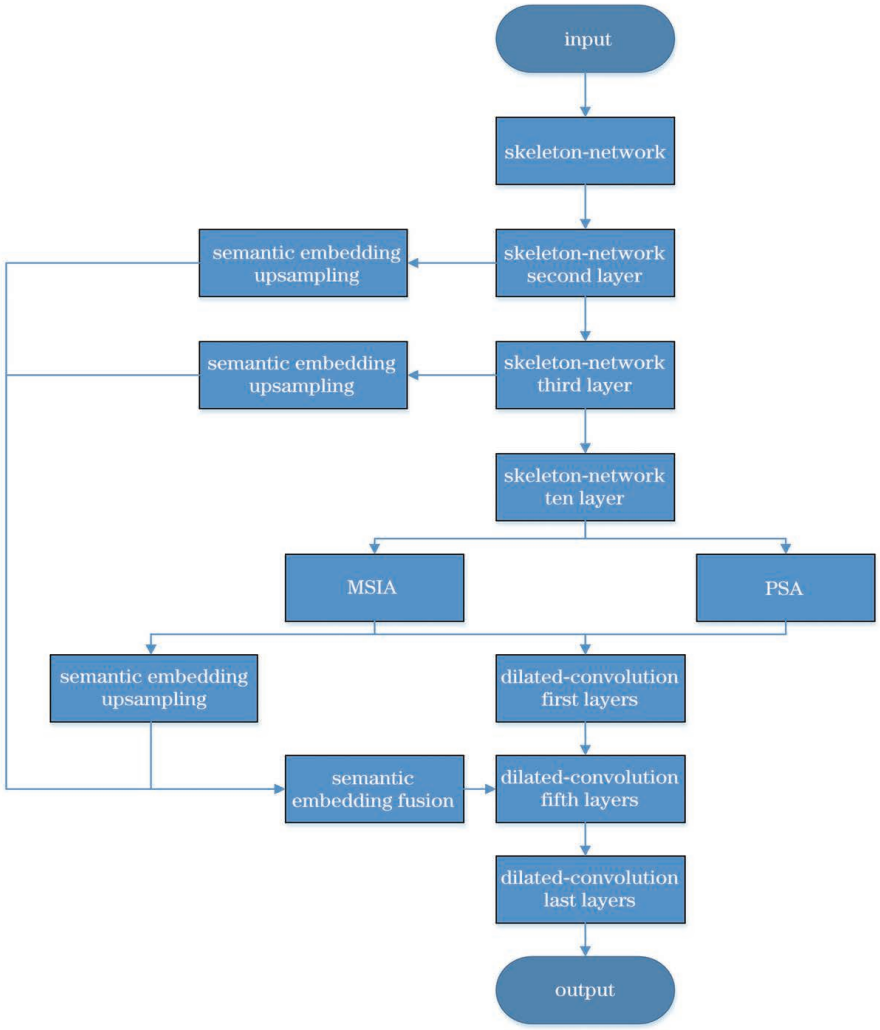

图 7　多特征信息融合网络算法的程序框图

Fig. 7　Flow chart of multi-feature information fusion network algorithm

3　实验分析与讨论

本文在 ShanghaiTech、Mall 数据集及 Worldexpo'10 数据集中进行了实验，并对实验结果进行分析。对比实验均在 Ubuntu 系统下进行，GPU 型号为 RTX2080Ti，环境配置为 CUDA9.0＋anaconda3＋python3＋tensorflow1.8.0。本文算法中的所有层均使用标准差为 0.01 的高斯分布进行初始化处理，网络初始训练学习率为 10^{-5}，迭代次数为 1200。此外，为了使模型充分训练，本文采用数据增强方法，对样本图像随机进行裁剪、旋转、放缩等操作，扩充数据集的样本数量，增强 CNN 模型的鲁棒性。

3.1　评价指标

人群密度估计领域的大部分研究均采用平均绝对误差（Mean Absolute Error，MAE）和均方误差（Mean Square Error，MSE）作为评价指标。为了能够较好地进行实验对比分析，本文使用平均绝对误差、均方误差作为评价指标。

MAE 反映网络预测人数与图像真值人数之间的误差，其定义为

$$f_{\mathrm{MAE}}=\frac{1}{N'}\sum_{i'=1}^{N'}\left|C_{i'}-C_{i'}^{\mathrm{GT}}\right|,\tag{7}$$

MSE 描述网络预测人数与图像真值人数之间的差异程度，其定义为

$$f_{\mathrm{MSE}}=\sqrt{\frac{1}{N'}\sum_{i'=1}^{N'}\left(C_{i'}-C_{i'}^{\mathrm{GT}}\right)^{2}},\tag{8}$$

式中：N'为测试图像数量；$C_{i'}$为第 i'幅测试图像的预测人数；$C_{i'}^{\mathrm{GT}}$为第 i'幅测试图像的真实人数。

3.2　在 ShanghaiTech 数据集中的实验与分析

ShanghaiTech 数据集共包含 1198 幅图像，数据集共分为两部分，即 Part_A 和 Part_B。Part_A

包含 482 幅图像，来源于互联网；Part_B 包含 716 幅图像，来源于上海的街道。本文将 Part_A 的 300 幅和 Part_B 的 400 幅图像用于训练，其余用于测试。ShanghaiTech 数据集的单幅图像实验结果如图 8 所示，多算法性能指标的对比如表 1 所示。

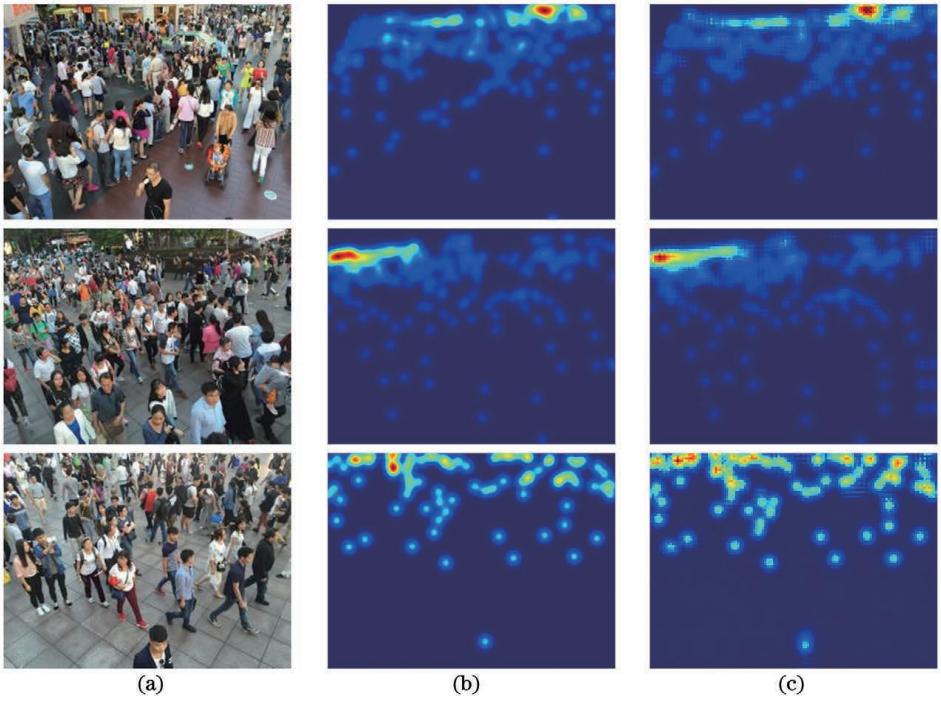

图 8　ShanghaiTech 数据集中的实验结果。(a)原图；(b)真值图；(c)预测结果

Fig. 8　Experimental results in ShanghaiTech dataset. (a) Original graphs; (b) true-value graphs; (c) prediction results

表 1　ShanghaiTech 数据集中算法性能的对比

Table 1　Performance comparison among algorithms for ShanghaiTech dataset

Method	Part_A		Part_B	
	MAE	MSE	MAE	MSE
Algorithm in Ref. [33]	181.8	277.7	32.0	49.8
MCNN[24]	110.2	173.2	26.4	41.3
Switch-CNN[34]	90.4	135.0	21.6	33.4
MSCNN[35]	83.8	127.4	17.7	30.2
CSRNet[7]	68.2	115.0	10.6	16.0
SANet[36]	67.0	104.5	8.4	13.6
Proposed algorithm	**63.2**	**102.8**	**8.0**	**12.8**

由 ShanghaiTech 数据集中的实验结果可知，在 Part_A 中，本文算法的 MAE 与文献[36]相比下降了 3.8，MSE 下降了 1.7；在 Part_B 中，与文献[36]相比，MAE 下降了 0.4，MSE 下降了 0.8。这是因为 Part_B 以稀疏区域人群为主，含有较多的背景干扰因素，所以精度有所降低。

3.3　在 Mall 数据集中的实验与分析

Mall数据集是使用安装在购物中心的监视摄像机收集到的数据集。该数据集具有人群密度变化大、活动模式多、透视畸变以及遮挡严重等特点。实验选用 1000 幅图像作为训练集，1000 幅图像作为测试集。Mall 数据集的单幅图像实验结果如图 9 所示，多算法性能指标的对比如表 2 所示。

表 2　Mall 数据集中算法性能的对比

Table 2　Performance comparison among algorithms for Mall dataset

Method	MAE	MSE
Algorithm in Ref. [33]	3.15	15.7
MCNN[24]	2.21	7.33
Switch-CNN[34]	2.01	6.25
MSCNN[35]	2.12	7.04
DecideNet[37]	1.52	1.90
Proposed algorithm	**1.43**	**1.72**

由实验结果可得，在 Mall 数据集中，本文所提模型的 MAE 与文献[37]相比下降了 0.9，MSE 下降了 0.8，相较于其他算法均有不错的表现。表明本文提出的模型在复杂多变的室内场景下也有较好的精度，同时所提模型在 MSE 上优于对比算法，证明本文算法具有较高的鲁棒性。

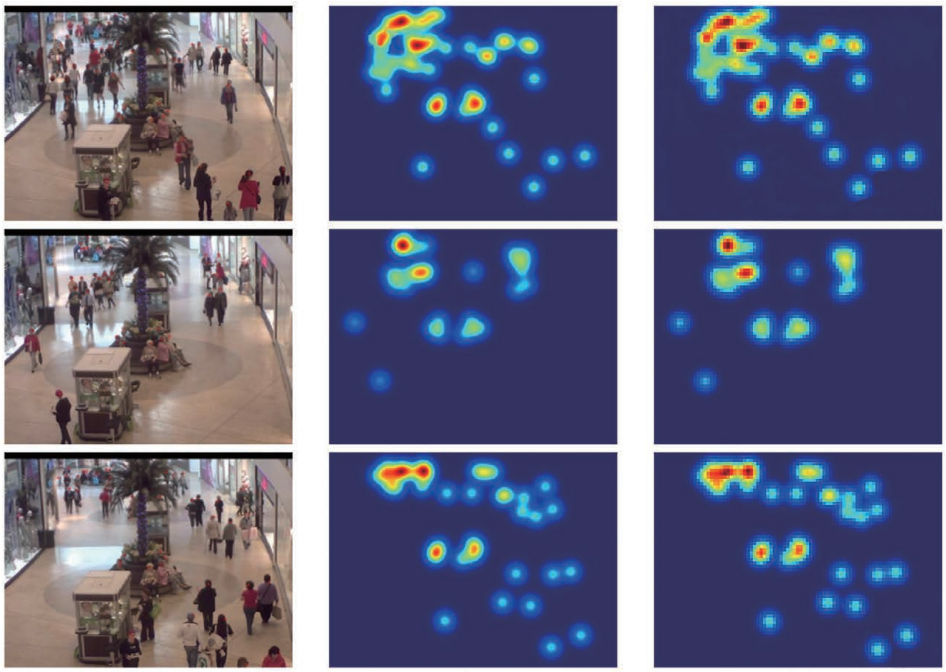

图 9　Mall 数据集中的实验结果。(a)原图;(b)真值图;(c)预测结果

Fig. 9　Experimental results in Mall dataset. (a) Original graphs; (b) true-value graphs; (c) prediction results

3.4　在 Worldexpo'10 数据集中的实验与分析

Worldexpo'10 的 10 个数据集是从 2010 年世界博览会的 103 个不同场景中收集的 1132 个带注释的视频序列。这些数据共有 3980 frame,统一归一化至 576×720 大小。本文用于训练的图像尺寸是 144×144,训练部分使用 3380 frame,剩余的数据用于测试。由于测试场景(S1～S5)提供了感兴趣区域(Region of Interest, ROI),由数据集常用评估结果可知,均只统计 ROI 区域内的人员。本文使用 MAE 进行评估,多算法性能指标的对比如表 3 所示。

表 3　Worldexpo'10 数据集中算法性能的对比

Table 3　Performance comparison among algorithms for Worldexpo'10 dataset

Method	S1	S2	S3	S4	S5	Average MAE
MCNN[24]	3.4	20.6	12.9	13.0	8.1	11.6
MSCNN[35]	7.8	15.4	14.9	11.8	5.8	11.7
Switch-CNN[34]	4.4	15.7	10.0	11.0	5.9	9.4
DecideNet[37]	2.0	13.14	8.9	17.4	4.75	9.23
CSRNet[7]	2.9	11.5	8.6	16.6	3.4	8.6
Proposed algorithm	2.6	11.2	8.9	14.2	3.6	8.1

从表 3 中可以看出,本文所提算法在所有 5 个场景中的平均 MAE 达到 8.1。与文献[7]相比,本文所提算法的性能较优,MAE 下降了 0.5。在 5 个场景中,不同方法预测结果的变化较大,不同方法对特定场景有各自的优势,与其他算法相比,本文所提算法得到了三个最小误差。表明在不同的场景下,本文所提算法具有良好的鲁棒性。

由上述三个数据集中的实验结果可知,本文算法无论是在 MAE 指标还是在 MSE 指标上均优于对比算法,分析原因主要有以下两点:1)对比方法在空间信息的补充以及特征提取能力上有所欠缺,而本文方法通过空间信息聚合以及多尺度特征提取,在融合图像特征信息时生成了较高质量的特征图;2)对比方法使用高级特征,对人群有更本质的描述,但损失了纹理、轮廓等底层特征,而本文方法采用的特征融合结构保留了底层特征,实现了更精确的预测密度图生成。

3.5　算法复杂度的对比分析

表 4 为算法复杂度的对比分析。由表 4 可以看出,Switch-CNN[34]模型的网络结构最大,运行速度

表 4 算法复杂度的对比分析

Table 4 Comparative analysis of algorithm complexity

Method	Size /MB	Average running speed of test image /s		
		ShanghaiTech	Mall	Worldexpo'10
Algorithm in Ref. [33]	7.1	2.36	0.32	1.22
MCNN[24]	19.2	2.31	0.32	1.15
Switch-CNN[34]	32.2	2.71	0.43	1.35
MSCNN[35]	22.2	2.34	0.32	1.14
CSRNet[7]	16.26	1.97	0.26	0.93
Proposed algorithm (MSIA)	17.39	2.11	0.29	1.02
Proposed algorithm (MSIA+PSA)	21.4	2.32	0.31	1.11

也最慢;文献[33]中的模型较小,但其采用全连接层,速度较慢;MCNN[24]、MSCNN[35]模型采用了多列结构,同时使用了尺寸较大的卷积核,导致模型参数量较大,运行速度较慢。

本文模型相对较小,且模型运行速度较快。分析原因有以下三点:1)本文所提模型运用了空洞卷积思想,在扩大感受野的基础上大大减少了参数量;2)本文的多尺度信息模块部分,虽采用额外训练时间增强卷积核的特征提取能力,但不影响模型速度,并且采用组合卷积核的模式,在保留上下文信息的同时减少了参数量;3)本文模型中出现速度慢及模型增大的原因是空间注意力透视方法,其在获取图像的空间全局上下文信息时,卷积递进的过程增大了模型的大小、降低了模型的速度。

4 结 论

提出了一种基于多特征信息融合卷积神经网络的人群密度估计方法。网络首先通过空间注意力透视方法,获取特征图的空间全局上下文信息,然后通过多尺度信息聚合方法,增强网络对图像尺度特征的提取能力以及细节特征的提取能力,最后通过细致语义特征嵌入融合方式,补充高层特征图的空间信息及低层特征图的语义信息,并使上下文信息与尺度信息相互补充,解决了特征空间信息丢失、全局上下文信息提取困难、人群特征较难提取等问题。多个数据集的实验结果表明:1)所提模型在不同人群场景下都具有较好的密度估计性能,泛化能力较强;2)所提方法对人群计数任务给出了有益尝试。但所提网络仍有不足,在处理更高遮挡场景时效果依然不理想,如何进一步弱化视角变化带来的影响、提升网络的速度并降低模块大小是后续改进的重点。

参 考 文 献

[1] Zhao L, He Z H, Cao W M, et al. Real-time moving object segmentation and classification from HEVC compressed surveillance video[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018, 28(6): 1346-1357.

[2] Coşar S, Donatiello G, Bogorny V, et al. Toward abnormal trajectory and event detection in video surveillance[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 683-695.

[3] Li X, Chen M, Nie F, et al. A multiview-based parameter free framework for group detection[J]. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, 2017: 4147-4153.

[4] Yu C Y, Xu Y, Gou L S, et al. Crowd counting based on single-column deep spatiotemporal convolutional neural network[J]. Laser & Optoelectronics Progress, 2021, 58(8): 0810011.
鱼春燕, 徐岩, 缑丽莎, 等. 基于单列深度时空卷积神经网络的人群计数[J]. 激光与光电子学进展, 2021, 58(8): 0810011.

[5] Kilambi P, Ribnick E, Joshi A J, et al. Estimating pedestrian counts in groups[J]. Computer Vision and Image Understanding, 2008, 110(1): 43-59.

[6] Idrees H, Saleemi I, Seibert C, et al. Multi-source multi-scale counting in extremely dense crowd images[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 2547-2554.

[7] Li Y H, Zhang X F, Chen D M. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes [C] // 2018 IEEE/CVF

Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 1091-1100.

[8] Zuo J, Ba Y L. Population-depth counting algorithm based on multiscale fusion[J]. Laser & Optoelectronics Progress, 2020, 57(24): 241502.
左静, 巴玉林. 基于多尺度融合的深度人群计数算法[J]. 激光与光电子学进展, 2020, 57(24): 241502.

[9] Zhao J M, Li X D, Li B S. Research on the algorithm of sheep dense counting based on UAV images[J]. Laser & Optoelectronics Progress, 2021, 58(22): 2210013.
赵建敏, 李雪冬, 李宝山. 基于无人机图像的羊群密集计数算法研究[J]. 激光与光电子学进展, 2021, 58(22): 2210013.

[10] Liu J, Gao C Q, Meng D Y, et al. DecideNet: counting varying density crowds through attention guided detection and density estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 5197-5206.

[11] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1627-1645.

[12] Chan A B, Vasconcelos N. Counting people with low-level features and Bayesian regression[J]. IEEE Transactions on Image Processing, 2012, 21(4): 2160-2177.

[13] Shi Z L, Ye Y D, Wu Y P, et al. Crowd counting using rank-based spatial pyramid pooling network[J]. Acta Automatica Sinica, 2016, 42(6): 866-874.
时增林, 叶阳东, 吴云鹏, 等. 基于序的空间金字塔池化网络的人群计数方法[J]. 自动化学报, 2016, 42(6): 866-874.

[14] Chen K, Gong S G, Xiang T, et al. Cumulative attribute space for age and crowd density estimation[C]//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE Press, 2013: 2467-2474.

[15] Idrees H, Tayyab M, Athrey K, et al. Composition loss for counting, density map estimation and localization in dense crowds[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11206: 544-559.

[16] Zhu L, Zhao Z J, Lu C, et al. Dual path multi-scale fusion networks with attention for crowd counting[EB/OL]. (2019-02-04)[2021-02-25]. https://arxiv.org/abs/1902.01115.

[17] Rodriguez M, Laptev I, Sivic J, et al. Density-aware person detection and tracking in crowds[C]//2011 International Conference on Computer Vision, November 6-13, 2011, Barcelona, Spain. New York: IEEE Press, 2011: 2423-2430.

[18] Arteta C, Lempitsky V, Noble J A, et al. Interactive object counting[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8691: 504-518.

[19] Ren S Q, He K M, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[20] Cao Y L, Ming T F, He G et al. Artificial recognition of centrifugal pump cavitation status based on deep learning[J]. Journal of Xi'an Jiaotong University, 2017, 51(11): 165-172.
曹玉良, 明廷锋, 贺国, 等. 基于深度学习的离心泵空化状态识别[J]. 西安交通大学学报, 2017, 51(11): 165-172.

[21] Chang L, Deng X M, Zhou M Q, et al. Convolutional neural networks in image understanding[J]. Acta Automatica Sinica, 2016, 42(9): 1300-1312.
常亮, 邓小明, 周明全, 等. 图像理解中的卷积神经网络[J]. 自动化学报, 2016, 42(9): 1300-1312.

[22] Wang C, Zhang H, Yang L, et al. Deep people counting in extremely dense crowds[C]//Proceedings of the 23rd ACM international conference on Multimedia, October 13, 2015, Brisbane, Australia. New York: ACM, 2015: 1299-1302.

[23] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84-90.

[24] Zhang Y Y, Zhou D S, Chen S Q, et al. Single-image crowd counting via multi-column convolutional neural network[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 589-597.

[25] Meng Y B, Ji T, Liu G H, et al. Encoding-decoding multi-scale convolutional neural network for crowd counting[J]. Journal of Xi'an Jiaotong University, 2020, 54(5): 149-157.
孟月波, 纪拓, 刘光辉, 等. 编码-解码多尺度卷积神经网络人群计数方法[J]. 西安交通大学学报, 2020,

54(5): 149-157.

[26] Liu W Z, Salzmann M, Fua P. Context-aware crowd counting[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 15-20, 2019, Long Beach, CA, USA. New York: IEEE Press, 2019: 5094-5103.

[27] Lempitsky V S, Zisserman A. Learning to count objects in images[C]//Advances in Neural Information Processing Systems 23: 24th Annual Conference on Neural Information Processing Systems 2010, December 6-9, 2010, Vancouver, British Columbia, Canada. Spain: Curran Associates, Inc., 2010: 1324-1332.

[28] Wang Q, Gao J Y, Lin W, 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.

[29] Ding X H, Guo Y C, Ding G G, et al. ACNet: strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV), October 27-November 2, 2019, Seoul, Korea (South). New York: IEEE Press, 2019: 1911-1920.

[30] Wang L J, Wen H L, Qin J W, et al. Asymmetric convolution with densely connected networks[J]. International Journal of Computing Science and Mathematics, 2020, 12(3): 274-284.

[31] Liu S T, Huang D, Wang Y H. Receptive field block net for accurate and fast object detection[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11215: 404-419.

[32] Zhang Z L, Zhang X Y, Peng C, et al. ExFuse: enhancing feature fusion for semantic segmentation[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11214: 273-288.

[33] Zhang C, Li H S, Wang X G, et al. Cross-scene crowd counting via deep convolutional neural networks[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 7-12, 2015, Boston, MA, USA. New York: IEEE Press, 2015: 833-841.

[34] Sam D B, Surya S, Babu R V. Switching convolutional neural network for crowd counting[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 4031-4039.

[35] Zeng L K, Xu X M, Cai B L, et al. Multi-scale convolutional neural networks for crowd counting[C]//2017 IEEE International Conference on Image Processing (ICIP), September 17-20, 2017, Beijing, China. New York: IEEE Press, 2017: 465-469.

[36] Cao X K, Wang Z P, Zhao Y Y, et al. Scale aggregation network for accurate and efficient crowd counting[M]//Ferrari V, Hebert M, Sminchisescu C, et al. Computer vision-ECCV 2018. Lecture notes in computer science. Cham: Springer, 2018, 11209: 757-773.

[37] Liu J, Gao C Q, Meng D Y, et al. DecideNet: counting varying density crowds through attention guided detection and density estimation[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 5197-5206.