CSRNet：通过扩展的卷积神经网络来理解高度拥挤的场景

李玉红[1,2]张小凡[1]陈德明[1]
[1]伊利诺伊大学香槟分校
[2]北京邮电大学{雷伊，小凡3号，
dchen@illinois.edu

摘要

我们提出了一个名为CSRNet的拥挤场景识别网络，以提供一种数据驱动和深度学习的方法，可以理解高度拥挤的场景，并执行准确的计数估计，以及呈现高质量的密度图。所提出的CSRNet由两个主要组件组成：卷积神经网络（二维特征提取的CNN）作为前端，后端是扩展的CNN，它使用扩展的内核来提供更大的接收场并取代池化操作。CSRNet是一个易于训练的模型，因为它的纯卷积结构。我们在四个数据集上展示了CSRNet(上海技术数据集，UCF_抄送_50个数据集，WorldEXPO‘10个数据集，和UCSD数据集)，我们提供了最先进的性能。在上海科技部分_B数据集，CSRNet的平均绝对误差（MAE）比以前最先进的方法低47.3%。我们扩展了目标用于计算其他对象的应用程序，如TRANCOS数据集中的车辆。结果表明，CSRNet显著提高了输出质量，比之前的最先进的方法降低了15.4%。

1.介绍

越来越多的网络模型已经被开发出来了
提供[1,2,3,4,5]为人群流监测、装配控制和其他安全服务提供有前途的解决方案。目前的拥挤场景分析方法是从简单的人群计数（输出目标图像中的人数）到密度图呈现（显示人群分布的特征）[6]。这种发展遵循了现实生活应用程序的需求，因为相同数量的人可能有完全不同的人群分布(如图所示。1)，所以仅仅计算一下人群的数量是不够的。分布图帮助我们获得更准确和全面的信息，这对于在踩踏和暴乱等高风险环境中做出正确的决定至关重要。然而，生成准确的分布模式是一项挑战。一个主要的困难

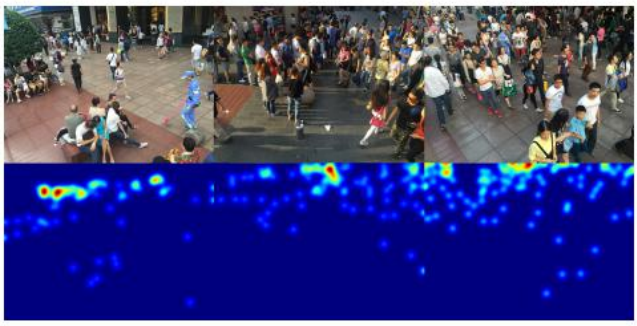

图1。第一行的图片显示了三张图片，全部包含95人在上海科技部分_B个数据集[18]，但它的空间分布完全不同。第二行的图片显示了它们的密度图。

来自于预测方式：由于生成的密度值遵循逐像素的预测，输出密度图必须包含空间相干性，这样才能呈现最近像素之间的平滑过渡。此外，多样化的场景，e。g.，不规则的人群集群和不同的摄像机视角，会使这项任务变得困难，特别是对于使用没有深度神经网络（DNNs）的传统方法。拥挤场景分析的发展以DNN语法为基础，因为它们在语义分割任务[7,8,9,10,11]方面取得了较高的准确性，并且在视觉显著性[12]方面取得了重大进展。使用DNNs的额外好处来自于热情的硬件社区，在那里，DNNs在gpu[13]、FPGAs [14,15,16]和ASICs [17]上被快速调查和实现。其中，低功耗、小尺寸的方案特别适合于在监控设备中部署拥挤的场景分析。

以往的拥挤场景分析工作大多基于多尺度架构[4,5,18,19,20]。他们在这一领域取得了很高的性能，但当网络更深入时，他们使用的设计也带来了两个显著的缺点：大量的训练时间和无效的分支结构（例如，[18]中的多列CNN（MCNN））。我们设计了一个实验来证明MCNN并比a

不

表现得更好

表1中的更深层次、更规则的网络。在[18]中使用MCNN的主要原因是不同列大小的卷积滤波器提供了灵活的接受域。直观地说，MCNN的每一列都致力于一定程度的拥挤场景。然而，使用MCNN的有效性可能并不显著。我们现在的图。2来说明在MCNN中通过三个独立的列（代表大、中、小的接受域）学习到的特征，并使用上海科技部分对其进行评估_[18]数据集。在50个测试用例中，图中的三条曲线具有非常相似的模式（估计错误率），这意味着这种分支结构中的每一列学习到几乎相同的特征。它违背了MCNN为学习每个列的不同特性而进行的设计的原始意图。

使用

在本文中，我们设计了一个更深的网络CSRNet用于计数人群和生成高质量的密度图。不像最新的工作，如[4,5]，它使用深度CNN作为辅助，我们专注于设计一个基于CNN的密度地图生成器。我们的模型使用纯卷积层作为骨干，以支持具有灵活分辨率的输入图像。为了限制网络的复杂性，我们使用了小规模的卷积滤波器(如 3x3)在所有的图层中。我们将VGG16 [21]的前10层作为前端，将扩展的卷积层作为后端，以扩大接受域，在不丢失分辨率的情况下提取更深的特征（因为不使用池化层）。利用这种创新结构，我们以上海科技[18]部分的7%、47.3%、10.0%和2.9%，超过了最先进的人群计数解决方案（基于MCNN的解决方案，称为CP-CNN [5]）_A、部分_b、UCF_抄送_50个[22]和世博会的10个[3]数据集。此外，我们在UCSD数据集[23]上实现了高性能。16 MAE.在将这项工作扩展到传输数据集[20]上的车辆计数后，我们实现的MAE比目前最好的FCN-HA [24]方法低15.4%。

sec

本文的其余部分的结构如下。秒2介绍了之前的人群计数和密度地图生成的工作。秒3介绍了我们的模型的体系结构和配置。4给出了在多个数据集上的实验结果。在秒。5、我们总结了这篇论文。

sec

2. 相关工作

遵循Loy等人提出的想法。[25]，人群场景分析的潜在解决方案可以分为三类：基于检测的方法、基于回归的方法和基于密度估计的方法。通过结合深度学习，基于cnn的解决方案在这项任务中表现出更强的能力，并优于传统的方法。

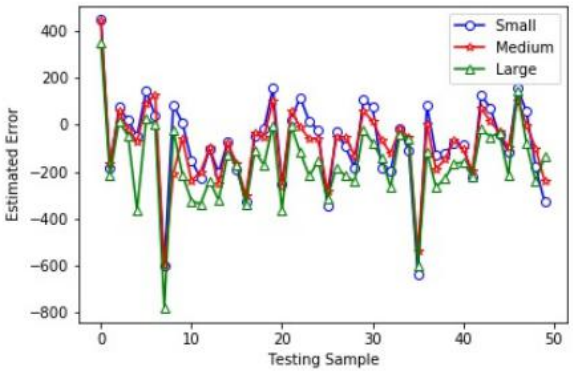

图2。来自上海科技部分测试集的50个样品的估计误差_由MCNN的三个预训练列生成的[18]。小、中、大分别代表MCNN中含有小、中、大核的柱。

方法	参数	更多的	MSE
关口MCNN的1	57.75k	141.2	206.8
关口MCNN的2	45.99k	160.5	239.0
关口3的MCNN	25.14k	153.7	230.2
MCNN总计	127.68k	110.2	185.9
更深层次的CNN	83.84k	**93.0**	**142.2**

表1。为了证明MCNN [18]可能不是最佳选择，我们设计了一个更深、参数更少的单列网络。建议的小型网络的结构为：CR（32、3）M-CR（64、3）M-CR（64、3）M-CR（32、3）CR（32、3）CR（1、1）。CR（m，n）表示具有m个滤波器的卷积层，其大小为nn，然后是ReLu层。M是最大池化层。结果表明，单列版本在上海科技部件上取得了较高的性能_一个MAE和均方误差（MSE）最小的数据集[18]

2.1. 基于检测的方法

大多数早期的研究集中在基于检测的方法上，使用一个类似移动窗口的探测器来检测人并计算他们的数量[26]。这些方法需要训练有素的分类器从整个人体中提取低级特征（如Haar小波[27]和HOG（面向直方图的梯度）[28]）。然而，它们在高度拥挤的场景中表现不佳，因为大多数目标物体都是模糊的。为了解决这个问题，研究人员检测特定的身体部位，而不是整个身体，以完成人群场景分析[29]。

2.2. 基于回归的方法

由于基于检测的方法不能适应于高度拥挤的场景，研究人员试图部署基于回归的方法来学习从裁剪后的图像斑块中提取的特征之间的关系，然后计算特定对象的数量。更多功能，

如前景和纹理特征，已被用于生成低级信息[30]。遵循类似的方法，Idrees等人。[22]提出了一种基于傅里叶分析和SIFT（标度不变特征变换）[31]兴趣点计数的特征提取模型。

.3.2.基于密度估计的方法

在执行基于回归的解决方案时，一个被称为显著性的关键特征被忽略了，它会导致在局部区域出现不准确的结果。兰皮茨基等人。[32]提出了一种通过学习局部特征与其目标密度映射之间的线性映射来解决这一问题。它整合了学习过程中的显著性信息。由于很难获得理想的线性映射，Pham等人。[33]使用随机森林回归来学习非线性映射，而不是线性映射。

.4.2基于cnn的方法

instead of

文献也侧重于基于cnn的方法来预测密度图，因为它在分类和识别[34,21,35]方面的成功。在Walach和Wolf [36]的工作中，演示了一种具有选择性采样和分层推进的方法。尚等人，~~而不是~~使用基于补丁的训练。[37]尝试了一种使用cnn的端到端回归方法，该方法将整个图像作为输入，并直接输出最终的人群计数。博米纳坦等人。[19]提出了第一个纯粹使用卷积网络和双列体系结构来生成密度图的工作。马斯登等人。[38]探索单列完全卷积网络，而Sindagi等人。[39]提出了一种利用高级先验信息来提高密度预测性能的CNN。Zhang等人提出了一种改进的结构。[18]引入了基于多列的架构（MCNN）用于人群计数。在Onoro和Sastre [20]中也显示了类似的想法，其中提出了一个尺度感知的、多列计数模型，称为Hydra CNN，用于目标密度估计。很明显，基于cnn的解决方案优于Sec中提到的之前的工作。2.1到2.3。

.5.2.最先进的方法的局限性

最近，Sam等人。[4]提出了使用开关~~ch~~-cnn的密度级分类器来为特定的输入补丁选择不同的回归器。辛达吉等人。[5]提出了一个上下文金字塔CNN，它使用CNN网络在不同的层次上估计上下文，以实现更低的计数误差和更好的质量密度图。这两种解决方案都实现了最先进的性能，并且都使用了基于多列的架构（MCNN）和密度级分类器。然而，我们观察到这些方法的几个缺点： (1)多列cnn很难按照中描述的训练方法进行训练工作[18]。这种膨胀的网络结构需要更多的时间来训练。(2)多列cnn引入了冗余性
结构，正如我们在章节中提到的。1.不同的列
似乎表现相似，没有明显的差异。(3)两种解决方案都需要密度级分类器
在MCNN中添加的图片。然而，在实时拥挤的场景分析中，由于密度水平的粒度不断变化，对象的数量有很大的范围。此外，使用细粒度的分类器意味着需要实现更多的列，这使得设计更加复杂，并导致更多的冗余。(4)这些工作花费了大量的密度级分类参数来标记输入区域，而不是将参数分配给最终的密度图生成。由于MCNN中的分支结构效率低下，因此缺乏生成密度图的参数，降低了最终的精度。考虑到这些缺点，我们提出了一种新的方法，集中于对拥挤场景中的深层特征进行编码，并生成高质量的密度图。

3.建议的解决方案

该设计的基本思想是部署一个更深层次的CNN，以捕获具有更大的接受域的高级特征，并在不严重扩大网络复杂性的情况下生成高质量的密度图。在本节中，我们首先介绍我们提出的体系结构，然后给出相应的训练方法。

.1.3 CSRNet体系结构

此段建议仔细读原文

根据[19,4,5]中类似的想法，我们选择VGG16 [21]作为CSRNet的前端，因为它具有很强的迁移学习能力，并且具有灵活的架构，易于连接后端生成密度图。×在CrowdNet [19]中，作者直接雕刻出VGG16的前13个层，并添加了1x1 个卷积层作为输出层。若没有修改会导致性能非常差。其他架构，如[4]，使用VGG16作为密度级分类器，在将输入图像发送到MCNN最合适的列之前标记它们，而CP-CNN [5]将分类结果与密度图生成器的特征结合起来。在这些情况下，VGG16作为一种辅助设备，而没有显著提高最终精度。在本文中，我们首先删除了VGG16的分类部分（全连接层），并在VGG16中构建了所提出的具有卷积层的CSRNet。这个前端网络的输出大小是原始输入大小的1/8。如果我们继续堆叠更多的卷积层和池化层（VGG16中的基本组件），输出大小将会进一步缩小，并且很难生成高质量的密度图。受[10,11,40]的启发，我们尝试部署扩展卷积层作为后端

提取更深层次的显著性信息，并保持输出的分辨率。

.1.13稀释卷积

我们设计的一个关键组件是扩展的卷积层。二维扩张卷积可以定义如下：

$$y(m,n)=\sum_{i=1}^{M}\sum_{j=1}^{N}x(m+r\times i,n+r\times j)w(i,j) \quad (1)$$

y（m、n）是输入x（m、n）和滤波器w（i、j）的展开卷积的输出，其长度和宽度分别为m和n。参数r是膨胀速率。如果r = 1，展开卷积变成正规卷积。

扩展卷积层已在分割任务中得到证明，显著提高了[10,11,40]的精度，是一种很好的替代方法。虽然池化层（如最大池化和平均池化）被广泛用于保持不变性和控制过拟合，但它们也显著降低了空间分辨率，意味着特征图的空间信息丢失。反卷积层[41,42]可以减轻信息的丢失，但是额外的复杂性和执行延迟可能并不适用于所有的情况。膨胀卷积是一个更好的选择，它使用稀疏核(如图所示。使用3)来交替使用池化层和卷积层。这个特征在不增加参数的数量或计算量(e。g.，添加更多的卷积层可以产生更大的接受域，但会引入更多的操作)。在放大卷积中，一个带有k-k滤波器的小尺寸核被放大到具有放大步幅r的k+（k-1）（r-1）。×因此，它允许在保持相同的分辨率的同时，灵活地聚合多尺度的上下文信息。这些例子如图中所示。×××3中，正常卷积得到3 3个感受野，两个扩张卷积分别得到5 5 个和7 7 个感受野。

扩展、膨胀、放大、是一个东西

为了保持特征图的分辨率，与使用卷积+池化+反褶积的方案相比，扩张型卷积具有明显的优势。我们在图中选择一个例子来说明。4.输入是人群的图像，通过两种方法分别进行处理，生成相同大小的输出。在第一种方法中，输入被因子为2的最大池化层降采样，然后将其传递到具有3 3 Sobel核的卷积层。×由于生成的特征图只有原始输入的1/2，因此需要通过反卷积层进行上采样（双线性插值）。在另一种方法中，我们尝试了扩展卷积，并将相同的3 3 Sobel核适应于具有a的扩展核×

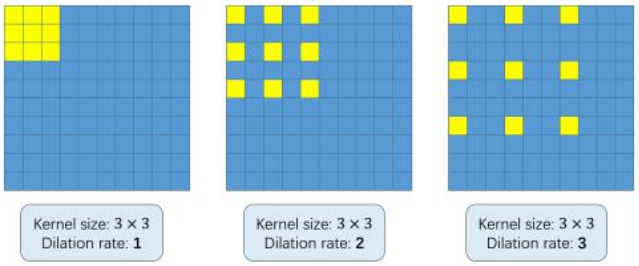

图3。×3 3 个卷积核分别为1、2、3。

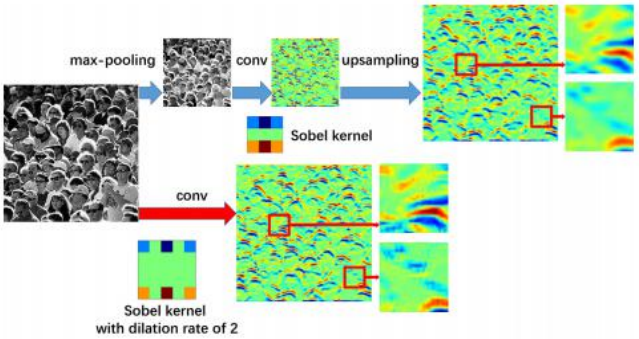

图4。扩张卷积与最大池化、卷积、上采样的比较。×这两种操作都使用了3 3 Sobel核，而膨胀率为2。

因素=2步幅。输出与输入共享相同的维度（这意味着不需要池化层和反卷积层）。最重要的是，扩展卷积的输出包含了更详细的信息（指的是我们放大的部分）。

.1.23网络配置

此段建议读原文

我们在表3中提出了四种CSRNet的网络配置，它们的前端结构相同，但后端的扩张速率不同。对于前端，我们采用了VGG16网络[21]（全连接层除外），只使用了3x3个内核。×根据[21]的研究，在针对相同大小的接受域时，使用具有较小内核的更多的卷积层比使用具有较大内核的更少的层更有效。

通过删除全连接的层，我们试图确定从VGG16中需要使用的层数。最关键的部分是在准确性和资源开销（包括训练时间、内存消耗和参数的数量）之间的权衡。实验表明，保持VGG16 [21]的前10层只有3个池化层而不是5层，可以实现最佳的权衡，以抑制池化操作对输出精度的不利影响。由于CSRNet的输出（密度图）较小（输入大小的1/8），我们选择因子为8的双线性插值进行缩放，并确保输出共享相同

数据集	生成方法
上海科技部分 _A [18]	几何自适应内核
UCF抄送 50	
上海科技部分_B [18]	固定的内核：一个= 15
转子[44]	固定的内核：一个= 10
世界博览会的10[3]	固定的内核：一个=3
UCSD [23]	

表2。针对不同数据集的地面真值生成方法

分辨率作为输入图像。在相同的尺寸下，CSRNet生成的结果可以与使用PSNR（峰值信噪比）和SSIM（图像[43]中的结构相似性）的地面真实结果相比较。

训练

.2.3培训方法

在本节中，我们将提供CSRNet培训的具体细节。通过利用常规的CNN网络（没有分支结构），CSRNet易于实现，部署速度快。

3.2.1地面真值生成 Ground truth generation

根据在[18]中生成密度映射的方法，我们使用几何自适应内核来处理高度拥挤的场景。通过使用高斯核（归一化为1）模糊每个头部注释，我们考虑每个数据集所有图像的空间分布生成地面真实值。几何自适应内核的定义为：

$$F(\boldsymbol{x}) = \sum_{i=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_i) \times G_{\sigma_i}(\boldsymbol{x}),\ with\ \sigma_i = \beta \overline{d_i} \qquad (2)$$

__对于每个目标对象xi在地面真相6中，我们使用了i表示k个最近邻的平均距离。为了生成密度图，我们将6(x-xi)，具有参数为a的高斯核i（标准差），其中x为像素在图像中的位置。在实验中，我们遵循[18]中的配置，其中β = 0.3和k = 3。对于稀疏人群的输入，我们采用高斯核来适应平均头部大小来模糊所有的注释。对不同数据集的设置如表2所示。

3.2.2数据增强

我们从每个图像的不同位置裁剪9个补丁，大小为原始图像的1/4。前四个补丁包含图像的四分之四，没有重叠，而其他五个补丁是随机裁剪的
输入图像。在那之后，我们镜像补丁，使训练集加倍。

CSRNet的配置			
A	B	C	D
输入（固定分辨率彩色图像）			
前端的 （由VGG16微调）			
conv3-641 conv3-641			
最大池			
conv31281 conv31281			
最大池			
conv3-2561 conv3-2561 conv3-2561			
最大池			
conv3-5121 conv3-5121 conv3-5121			
后端（四种不同的配置）			
conv3-5121 conv3-5121 conv3-5121 conv3-2561 conv31281 conv3-641	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-2 conv3128-2 conv3-64-2	conv3-512-2 conv3-512-2 conv3-512-2 conv3-256-4 conv3128-4 conv3-64-4	conv3-512-4 conv3-512-4 conv3-512-4 conv3-256-4 conv3128-4 conv3-64-4
conv11- 1			

表3。CSRNet的配置。所有的卷积层都使用填充来保持以前的大小。卷积层的参数被表示为“conv-（内核大小）-（过滤器的数量）-（膨胀率）”，最大池化层在一个2 2 像素的窗口上进行，步幅为2。×

.2.33培训细节

我们使用一种简单的方法来将CSRNet训练为端到端结构。前10个卷积层是由一个训练有素的VGG16 [21]进行微调的。对于其他层，初始值来自于一个具有0.01个标准差的高斯初始化。在训练过程中，学习速率采用随机梯度下降（SGD）。此外，我们选择欧几里得距离来测量地面真实值和我们生成的估计密度图之间的差异，这与其他工作[19,18,4]相似。损失函数如下：

learning rate：1e-6

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \| Z(X_i; \Theta) - Z_i^{GT} \|_2^2 \qquad (3)$$

其中，N为训练批的大小，Z (Xi；Q)是CSRNet生成的输出，参数显示为Q。Xi表示输入的图像，而Z_i^{gt}是输入图像X的地面真实结果吗i.

4. 实验

我们在五个不同的公共数据集[18,3,22,23,44]中演示了我们的方法。与之前最先进的方法[4,5]相比，我们的模型更小、更准确，也更容易训练和部署。本节介绍了评价指标，然后对上海科技部分进行了消融研究_我们使用了一个数据集来分析我们的模型的配置（如表3所示）。随着消融研究的进行，我们评估和比较了我们提出的方法与所有这五个数据集的以前的最先进的方法。我们的模型的实现是基于Caffe框架[13]。

4.1. 评价指标

MAE和MSE被用于评价哪个
定义为：

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | C_i - C_i^{GT} | \tag{4}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} | C_i - C_i^{GT} |^2} \tag{5}$$

其中，N为一个测试序列中的图像数，而C_i^{gt}这才是计数的基本真理。C_i表示估计的计数，其定义如下：

$$C_i = \sum_{l=1}^{L}\sum_{w=1}^{W} z_{l,w} \tag{6}$$

L和W分别表示密度图的长度和宽度，而$z_{l,w}$是生成的密度图的（l，w）处的像素。C_i表示图像X的估计计数数i.

我们还使用PSNR和SSIM来评估上海技术部分的输出密度图的质量_一个数据集。为了计算PSNR和SSIM，我们遵循[5]给出的预处理，其中包括调整密度图的大小（与原始输入的相同大小），并对地面真实值和预测密度图进行插值和归一化。

.2. 对上海科技部分的4次消融_A

在本小节中，我们进行了消融研究，分析了上海科技部分的CSRNet的四种配置_数据集[18]是一个新的大规模人群计数数据集，包括482张拥挤场景的图像，241,667名注释人员。从这些图像中计算出来是具有挑战性的，因为极其拥挤的场景，不同的视角，和不固定的分辨率。这四种配置如表3所示。
CSRNet A是所有扩张速率均为1的网络。CSRNet B和D保持了2英寸和4英寸的膨胀率

架构	更多的	MSE
CSRNet A	69.7	116.0
CSRNet B	**68.2**	**115.0**
CSRNet C	71.91	120.58
CSRNet D	75.81	120.82

表4。上海科技部分的体系结构比较_一个数据集

而CSRNet C分别结合了2和4的扩张率。这四种模型的参数数均为16.26M。我们打算用不同的膨胀速率来比较结果。在经过了关于上海部分的培训后_一个使用第二节中提到的方法的数据集。3.2，我们执行在Sec中定义的评估指标。4.1.我们尝试了辍学[45]来防止潜在的过拟合问题，但没有明显的改进。所以我们在模型中不包括辍学。详细的评估结果如表4所示，其中CSRNet B的误差最低（精度最高）。因此，我们使用CSRNet B作为所提出的CSRNet来进行以下实验。

4.3. 评价与比较

.3.14上海科技数据集

上海科技人群计数数据集包含1198张标注图像，共计330,165人，[3]。这个数据集由两部分组成_A包含482张图片和高度拥挤的场景随机下载从互联网上，而部分_B包括716张来自上海街道的相对稀疏的人群场景。并将我们的方法与其他六种方法进行了评估和比较
最近的工作和结果见表5。它表示
我们的方法部分达到了最低的MAE（最高的精度）_与其他方法相比，我们得到的MAE比最先进的解决方案CP-低7%中心体CSRNet还部分降低了47.3%的MAE_B与CP-CNN相比。为了评估生成的密度图的质量，我们将我们的方法与MCNN和CP-CNN进行了比较_一个数据集，我们遵循Sec中的评估指标。.2.3个测试用例的样本见图5。结果如表6所示，表明CSRNet的SSIM和PSNR值最高。我们
并报告了表11中上海技术数据集的质量结果。

.3.24 UCF_抄送_50个数据集

超离心分离机_抄送_50个数据集包括50张不同透视图和分辨率的图像，[22]。每幅图像标注的人数为94到4543人，平均人数为1280人。按照[22]中的标准设置进行5倍交叉验证。结果比较

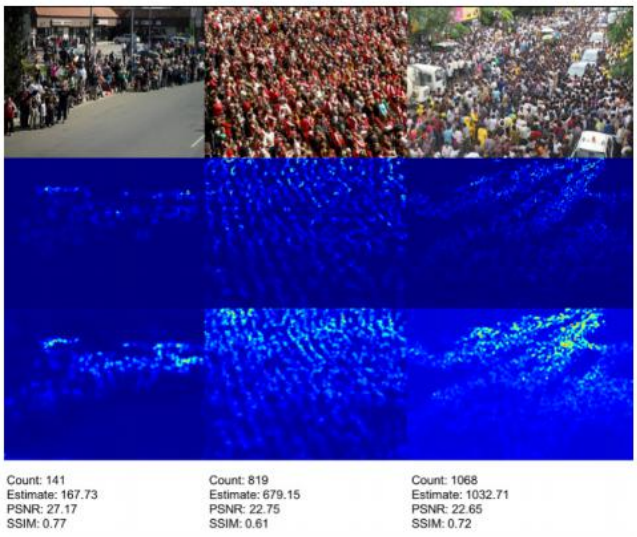

图5。第一行为上海科技部分测试集的样本_一个数据集。第二行显示了每个样本的地面真实情况，而第三行显示了由CSRNet生成的密度图。

	部 A 分		部 _B 分	
方法	更多的	MSE	更多的	MSE
张等人。[3]	181.8	277.7	32.0	49.8
马斯登等人。[38]	126.5	173.5	23.8	33.1
MCNN [18]	110.2	173.2	26.4	41.3
级联-MTL[39]	101.3	152.4	20.0	31.1
交换机-CNN[4]	90.4	135.0	21.6	33.4
CP-CNN [5]	73.6	**106.4**	20.1	30.1
CSRNet（我们的）	**68.2**	115.0	**10.6**	**16.0**

表5。上海科技数据集的估计误差

方法	信号-噪音功率比	斯西姆
MCNN [18]	21.4	0.52
CP-CNN [5]	21.72	0.72
CSRNet（我们的）	**23.79**	**0.76**

表6。上海科技部分密度质量图_一个数据集

MAE和MSE列于表7，生成的密度图的质量见表11。

.3.3世界博览会的10个数据集

世博会的10个数据集[3]由108个不同监控摄像机拍摄的视频序列的1132个视频序列的3980个注释帧组成。这个数据集被分为一个训练集（3380帧）和一个来自五个不同场景的测试集（600帧）。为整个数据集提供了感兴趣的区域（ROI）。在预处理过程中，每一帧及其点图都被ROI掩盖，我们按照Sec中给出的指令来训练我们的模型。3.2.结果

方法	更多的	MSE
伊德里斯等人。[22]	419.5	541.6
张等人。[3]	467.0	498.5
MCNN [18]	377.6	509.1
Onoro等。[20]Hydra-2	333.7	425.2
Onoro等。[20]九头蛇-3s	465.7	371.8
沃拉赫等人。[36]	364.4	341.4
马斯登等人。[38]	338.6	424.5
级联-MTL[39]	322.8	397.9
交换机-CNN[4]	318.1	439.2
CP-CNN [5]	295.8	**320.9**
CSRNet（我们的）	**266.1**	397.5

表7。UCF估计误差_抄送_50个数据集

如表8所示。所提出的CSRNet在5个场景中的4个场景中提供了最好的精度，并且它平均达到了最好的精度。

方法	Sce。1	Sce。2	Sce。3	Sce。4	Sce。5	平均。
陈等人。[46]	2.1	55.9	9.6	11.3	3.4	16.5
张等人。[3]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [18]	3.4	20.6	12.9	13.0	8.1	11.6
尚等人。[37]	7.8	15.4	14.9	11.8	5.8	11.7
交换机-CNN[4]	4.4	15.7	10.0	11.0	5.9	9.4
CP-CNN [5]	2.9	14.7	10.5	**10.4**	5.8	8.86
CSRNet（我们的）	**2.9**	**11.5**	**8.6**	16.6	**3.4**	**8.6**

表8。对世博会的10个数据集的估计误差

4.3.4UCSD数据集

加州大学圣迭戈分校的数据集[23]有2000帧被监控摄像机捕获。这些场景包含11到46人的稀疏人群。同时，还提供了感兴趣的区域（ROI）。×由于每一帧的分辨率都是固定的且较小的（238 158），因此在频繁的池化操作后，很难生成高质量的密度图。因此，我们利用双线性插值法对帧进行预处理，将其调整为952 632。×在2000帧中，我们使用601到1400帧作为训练集，其余的根据[23]作为测试集。在模糊我们在Sec中提到的注释之前。3.2，所有的帧和相应的点图都被ROI掩盖。运行UCSD数据集的准确性如表9所示，我们在MAE类别中除了MCNN之外的性能优于之前的大多数方法。结果表明，该方法不仅可以对极其密集的人群进行计数，还可以对相对稀疏的场景进行计数。此外，我们还在表11中提供了生成的密度图的质量。

4.3.5跨型数据集

除了人群计数之外，我们还在交通系统数据集[44]上设置了一个实验，用于车辆计数

方法	更多的	MSE
张等人。[3]	1.60	3.31
CCNN [20] CCNN	1.51	–
交换机-CNN[4]	1.62	2.10
FCN-rLSTM [24]	1.54	3.02
CSRNet（我们的）	1.16	1.47
MCNN [18]	**1.07**	**1.35**

表9。在UCSD数据集上的估计错误

证明了我们的方法的鲁棒性和通用性。TRANCOS是一个公共交通数据集，包含1244张不同拥堵交通场景的图像，由监控摄像头和46796辆带注释的车辆捕获。此外，还提供了感兴趣区域（ROI）。图像的视角不是固定的，而且图像是从非常不同的场景中收集出来的。本测试采用网格平均平均绝对误差（GAME）[44]进行评估。“游戏”的定义如下：

$$GAME(L) = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{l=1}^{4^L}\left|D_{I_n}^{l} - D_{I_n^{gt}}^{l}\right|\right) \qquad (7)$$

其中，N为测试集中的图像数，和$D_{I_n}^{l}$是输入图像n在区域l内的估计结果。$D_{I_n}^{l^{gt}}$为对应的地面真实结果。对于特定的级别L，GAME(L)使用4对图像进行细分^{L}覆盖完整图像的非重叠区域，
误差计算为每个区域的MAE之和。当L=为0时，游戏相当于
MAE度量。

我们将我们的方法与之前的最先进的方法[47,32,20,24]进行了比较。[20]中的方法使用类MCNN网络生成密度图，而[24]中的模型部署了全卷积神经网络（FCN）和长短期记忆网络（LSTM）的组合。结果如表10所示，其中三个例子如图所示。6.我们的模型在四种不同的游戏指标上取得了显著的改进。与[20]的结果相比，CSRNet的GAME (0)低67.7%，60。GAME低1%(1)，GAME (2)降低48.7%，GAME低22.2%(3)，这是最好的解决方案。我们还在表11中显示了生成的密度图的质量。

方法	游戏0	游戏1	第二场比赛	第三场比赛
Fiaschi等。[47]	17.77	20.14	23.65	25.99
兰皮茨基等人。[32]	13.76	16.72	20.72	24.36
九头蛇-3s[20]	10.99	13.75	16.69	19.32
FCN-HA [24]	4.21	–	–	–
CSRNet（我们的）	**3.56**	**5.49**	**8.57**	**15.04**

表10。在TRANCOS数据集上的游戏

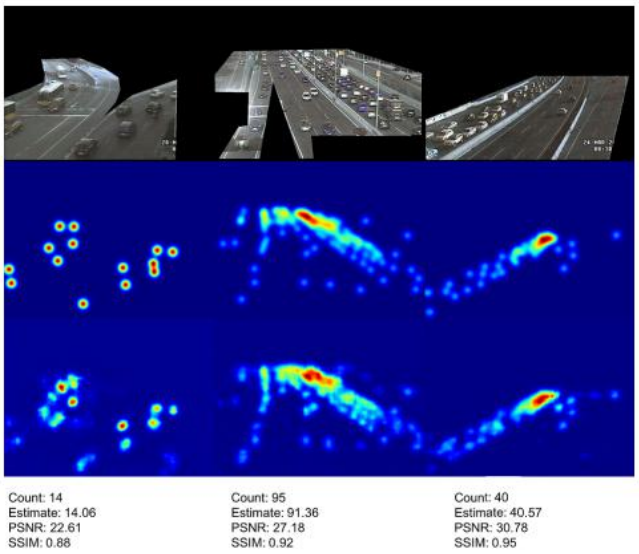

图6。第一行显示了具有ROI的TRANCOS [44]数据集中的测试集样本。第二行显示了每个样本的基本真相。第三行显示了由CSRNet生成的密度图。

数据集	信号-噪音功率比	斯西姆
上海科技部分 _A [18]	23.79	0.76
上海科技部分_B [18]	27.02	0.89
超离 _C C_50 [22] 心分 离机	18.76	0.52
世界博览会的10[3]	26.94	0.92
UCSD [23]	20.02	0.86
转子[44]	27.10	0.93

表11。CSRNet在5个数据集中生成的密度图的质量

5. 结论

在本文中，我们提出了一种新的架构CSRNet，用于人群计数和高质量的密度地图生成，采用易于训练的端到端方法。我们使用扩展的卷积层来聚合拥挤场景中的多尺度上下文信息。通过利用扩展的卷积层，CSRNet可以在不失去分辨率的情况下扩展接受域。我们在四个人群计数数据集中以最先进的性能展示了我们的模型。我们还将我们的模型扩展到车辆计数任务，我们的模型也取得了最好的精度。

6. 致谢

这项工作得到了IBM-伊利诺斯州认知计算系统研究中心（C3SR）的支持，这是作为IBM AI视野网络的一部分的一个研究合作项目。

参考文献

[1]贝贝，多萝西，保罗雷马尼诺，塞
和徐立群。人群分析：一项调查。机器视觉和应用程序，19（5-6）：345-357,2008。

李[2]，桓昌，王孟，倪冰冰，瑞昌
洪和水成燕。拥挤的场景分析：一项调查。IEEE关于视频技术电路和系统的交易，25(3)：367-386,2015。

张聪聪、李宏生、王小刚、杨小康。通过深度卷积神经网络进行跨场景人群计数。*在IEEE会议会议记录中《计算机视觉和模式识别》，第833-841页，2015年。*

[4] Deepak先生萨姆，希夫苏里亚，和R文卡特什先生。
切换卷积神经网络，用于人群计数。*发表在IEEE计算机视觉和模式识别会议论文集，第1卷，第6页，2017年。*

[5]，辛达吉和帕特尔。产生高
使用上下文cnn金字塔。*发表在IEEE计算机视觉和模式识别会议论文集，1861-1870页，2017年。*

张[6]聪、康凯、李宏生、王小刚、解荣、杨小康。数据驱动的人群理解：一个大规模的人群数据集的基线。*IEEE《多媒体交易》，18(6)：1048-1061,2016。*

[7]，乔纳森·朗，埃文·谢尔哈默，和特雷弗·达雷尔。用于语义分割的全卷积网络。*发表在IEEE计算机视觉和模式识别会议论文集上，第3431-3440页，2015年。*

魏云超[8]、梁小丹、陈云鹏、沈晓辉、
程明、冯、姚赵、水城燕。Stc：一个简单到复杂的弱监督语义分割框架。*IEEE关于模式分析和机器智能的交易，39（11）：2314-2320,2017。*

魏云超[9]，冯，梁小丹，明代师
程、姚赵、水城燕。具有对抗性擦除的对象区域挖掘：一种简单的语义分割分类方法。在IEEE CVPR，2017年。

[10] Fisher Yu和弗拉德伦·科尔顿。通过扩展卷积实现的多尺度上下文聚合。在ICLR，2016年。

[11] L.C.陈，G。帕潘德里欧，我。KokkinosK。墨菲和A.L.尤伊尔Deeplab：具有深度卷积网、无对称卷积和全连通crfs的语义图像分割。*IEEE《模式分析与机器智能交易报》，PP（99）：1-1,2017。*

[12]容潘，伊丽莎赛罗，泽维尔吉罗涅托，凯文
麦吉尼斯和诺埃尔·奥康纳。用于显著性预测的浅层和深度卷积网络。*发表在IEEE计算机视觉和模式识别会议论文集上，第598-606页，2016年。*

[13]杨庆贾，埃文谢尔哈默，杰夫多纳休，谢尔盖
卡拉耶夫，乔纳森·朗，罗斯·格尔希克，塞尔吉奥·瓜达拉马，
和特雷弗达雷尔。卷积体系结构
功能嵌入。*在第22届ACM会议程序中全国多媒体会议，第675-678页。*ACM，2014年。

[14]邱建涛、王杰、宋耀、郭开元、李博勋、
周二金、晋城Yu、天启唐、徐宁毅、宋森、余王、华中阳。深入使用嵌入式卷积神经网络FPGA平台。在*2016年ACM/SIGDA现场可编程门阵列国际研讨会论文集，FPGA '16，第26-35页，纽约，纽约，美国，2016年。*ACM。

[15]，张小凡，刘新恒，阿南德·拉马钱德兰，
川浩诸葛、汤世斌、欧阳鹏、郑祖福、凯儿、陈德明。高性能视频内容识别与长期循环卷积网络的FPGA。*在现场可编程逻辑和应用（FPL），2017年第27届国际会议，第1-4页。*2017年IEEE。

[16]，张小凡，阿南德·拉马钱德兰，诸葛亮，
地和、魏左、郑祖福、罗普诺、陈德铭。在fpga上的机器学习来面对物联网革命。*在计算机辅助设计（ICCAD）中，2017年IEEE/ACM国际会议，第819-826页。*2017年IEEE。

[17]·伦佐·安德里，卡维格利，罗西和贝尼尼。瑜伽导航仪：一种基于二进制权值的超低功率卷积神经网络加速器。*在VLSI（ISVLSI），2016年IEEE计算机协会年度研讨会，第236-241页。*2016年IEEE。

张[18]莺、周德森、陈思勤、高盛华、马易。通过多列卷积神经网络进行单图像群体计数。*发表在IEEE计算机视觉和模式识别会议论文集上，第589-597页，2016年。*

[19]洛克什·布米纳坦，克鲁蒂文蒂，和
R文卡特什先生。Crowdnet：一个针对密集人群统计的深度卷积网络。*在2016年ACM关于多媒体会议的会议记录中，第640-644页。*ACM，2016年。

[20]丹尼尔·奥诺罗-卢比奥和罗伯托·勒佩兹-萨斯特雷。6通过深度学习来实现无视角的对象计数。*在欧洲计算机视觉会议上，第615-629页。*施普林格，2016年。

[21]凯伦·西蒙扬和安德鲁·齐塞尔曼。用于大规模图像识别的深度卷积网络。*arXiv预印本，arXiv：1409.1556,2014年。*

[22] Haroon Idrees，伊姆兰·萨莱米，科迪·塞贝特和穆巴拉克·沙阿。多源多尺度计数在极其密集的人群图像。*发表在IEEE计算机视觉和模式识别会议论文集上，第2547-2554页，2013年。*

[23] A.B.陈，张胜梁和N。瓦斯孔塞卢斯保护隐私的人群监控：计算没有人模型或跟踪。*2008年IEEE计算机视觉和模式识别会议，第1-7页，2008年6月。*

张上杭[24]、吴冠亨、P、何拉。Fcn-rlstm：用于城市摄像机中车辆计数的深度时空神经网络。*发表在IEEE计算机视觉和模式识别会议论文集上，第3667-3676页，2017年。*

[25]陈换洛伊、陈可、龚钢刚、陶翔。人群计数和分析：方法和评估。*关于人群的建模、模拟和可视化分析，第347-382页*。施普林格，2013年。

[26] Piotr美元，克里斯蒂安·沃耶克，伯恩特·席勒和彼得罗·佩罗纳。行人检测：对最新技术水平的评估。*IEEE关于模式分析和机器智能的交易，34(4)：743-761,2012。*

[27]的保罗·维奥拉和迈克尔·J·琼斯。健壮的实时人脸检测。*国际计算机视觉杂志，57(2)：137-154,2004。*

[28] Navneet·达拉尔和比尔·特里格斯。用于人体检测的定向梯度的直方图。*在计算机视觉和模式识别方面，2005年。CVPR 2005.IEEE计算机学会会议，第1卷，第886-893页*。IEEE，2005年。

[29]佩德罗·费尔森斯兹瓦尔布，罗斯·吉尔希克，大卫·麦卡利斯特，
和德瓦拉马南。用有区别训练的基于部分的模型进行对象检测。*IEEE关于模式分析和机器智能的交易，32(9)：1627-1645,2010。*

[30] Antoni B陈和努努·瓦斯康塞洛斯。人群计数的贝叶斯泊森回归。*在计算机视觉方面，2009年的IEEE第12届国际会议，第545-551页*。2009年IEEE。

[31] D.G.洛的变体从局部尺度不变特征进行的目标识别。*发表在第七届IEEE计算机视觉国际会议论文集，第2卷，第1150-1157页，第2卷，1999年。*

[32]，维克多·兰皮茨基和安德鲁·齐塞尔曼。学习计算图像中的对象。*《神经信息处理系统的进展》，第1324-1332页，2010年。*

[33]越南，田夫，山口先生和冈田龙佐。计数森林：使用随机森林对不确定的目标数量进行人群密度估计。在*计算机视觉（ICCV），2015年IEEE国际会议，第3253-3261页*。2015年IEEE。

[34]亚历克斯·克里热夫斯基，伊利亚·苏茨克弗和杰弗里·辛顿。
基于深度卷积神经网络的图像集分类。*《神经信息处理系统的进展》，第1097-1105页，2012年。*

[35] F.球团。具有深度可分离卷积的深度学习。*2017年IEEE计算机视觉与模式识别会议（CVPR），第1800-1807页，2017年7月。*

[36] Elad Walach和Lior Wolf。学会用CNN的推进来计数。*在欧洲计算机视觉会议上，第660-676页*。施普林格，2016年。

[37]冲商，海州爱，和波白。通过联合学习本地和全球计数来进行端到端人群计数。*在图像处理（ICIP）中，2016年IEEE国际会议，第1215-1219页*。2016年IEEE。

[38]马克马斯登，凯文麦吉尼斯，苏珊娜利特尔，和
诺埃尔E奥康纳。完全卷积的人群指望着高度拥挤的场景。*arXiv预印arXiv：1612.00220，*
2016.

[39]，辛达吉和帕特尔。基于cnn的级联多任务学习和密度估计的人群计数。*在基于视频和信号的高级监控（AVSS），2017年第14届IEEE国际会议上，第1-6页*。2017年IEEE。

[40] Liang-Chieh陈，乔治·帕潘德里欧，弗洛里安·施罗夫，和
哈特维格亚当。语义图像分割的重构式卷积。CoRR，abs/1706.05587,2017年。

[41]马修·泽勒，迪利普·克里希南，格雷厄姆·泰勒，和
罗布·费格斯。解卷积网络。*在计算机视觉和模式识别（CVPR）中，2010年IEEE会议，第2528-2535页*。2010年IEEE。

[42]庆宇、洪和韩博贤。
学习语义分割的反褶积网络。*发表在IEEE国际计算机视觉国际会议论文集，第1520-1528页，2015年。*

王[43]，艾伦博维克，哈米德谢赫和西蒙切利。图像质量评估：从误差可见性到结构相似性。*IEEE交易，13(4)：600-612,2004。*

[44]罗伯托-里卡多
格尔梅多，比奥特丽斯，托雷，吉姆内斯和丹尼尔·奥罗-卢比奥。极度重叠的车辆数量。在*伊比利亚模式识别和图像分析会议（IbPRIA），2015年。*

[45] Nitish斯里瓦斯塔瓦，杰弗里·辛顿，亚历克斯·克里热夫斯基，伊利亚
苏茨科弗和鲁斯兰·萨拉库迪诺夫。辍学生：一种防止神经网络过拟合的简单方法。*机器学习研究杂志，15：1929-1958,2014。*

陈[46]可、龚少刚、陶翔、陈
改变洛伊。用于年龄和人群密度估计的累积属性空间。*发表在IEEE计算机视觉和模式识别会议论文集上，第2467-2474页，2013页。*

[47] L.Fiaschi，美国科特，R。Nair和F。A.汉普雷希特。学习使用回归森林和结构化标签进行计数。在*第21届模式识别国际会议论文集（ICPR2012），第2685-2688页，2012年11月。*

7.附件：补充材料

在本附录中，由CSRNet从五个数据集(上海科技[18]，UCF_抄送_50 [22]，世博会的10[3]，UCSD[23]，和TRANCOS [44])被提出，以证明了我们的设计的有效性。使用两个标准作为PSNR（峰值信噪比）和SSIM(图像[43]中的结构相似性来评估我们的设计生成的密度图的质量。来自这5个数据集的样本如图所示。7图。12，它代表了不同的密度水平。

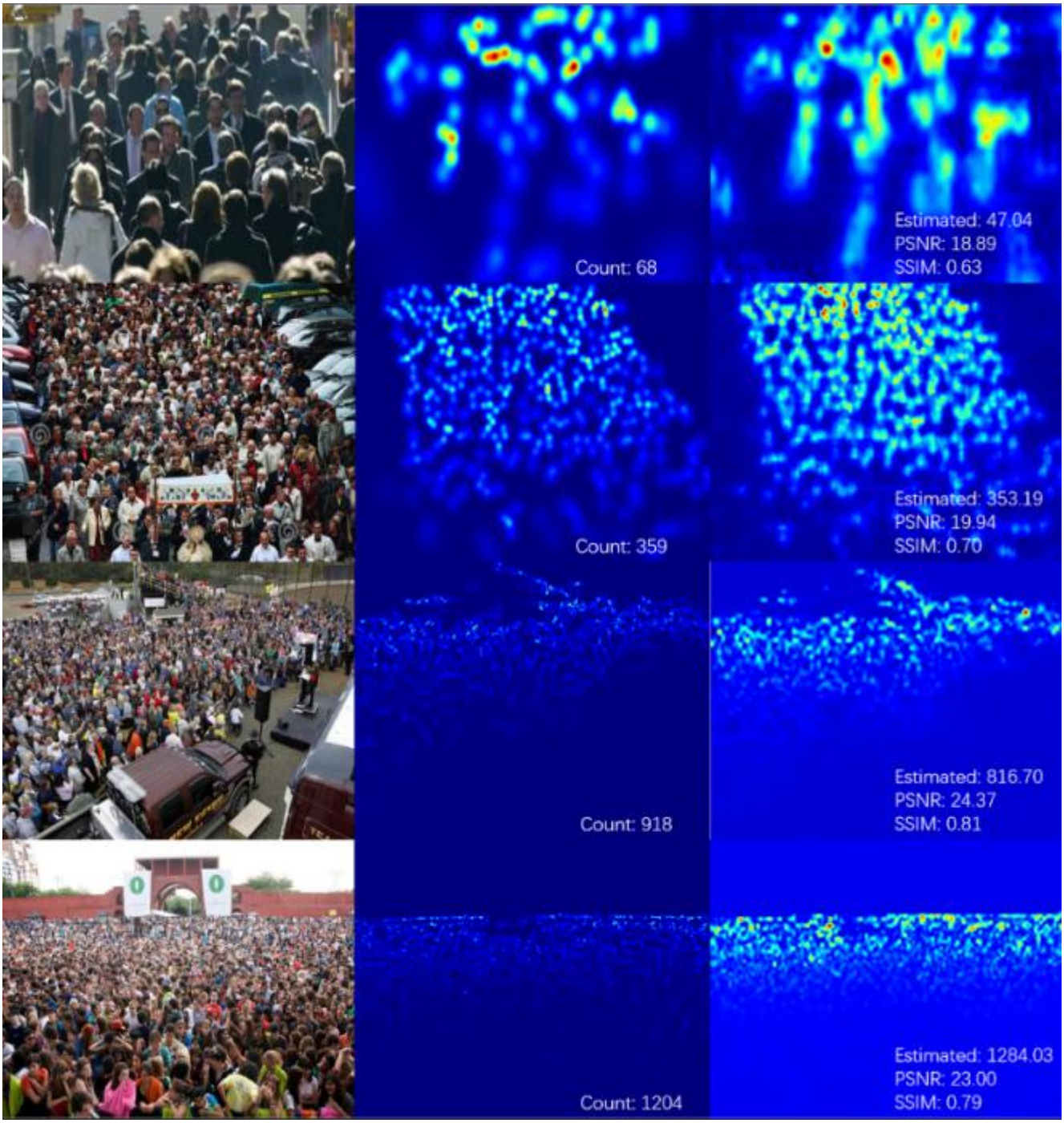

图7。由CSRNet从上海科技部分生成的样品_[18]数据集。左列显示原始图像；中间列显示地面真实密度图，而右列表示我们生成的密度图。

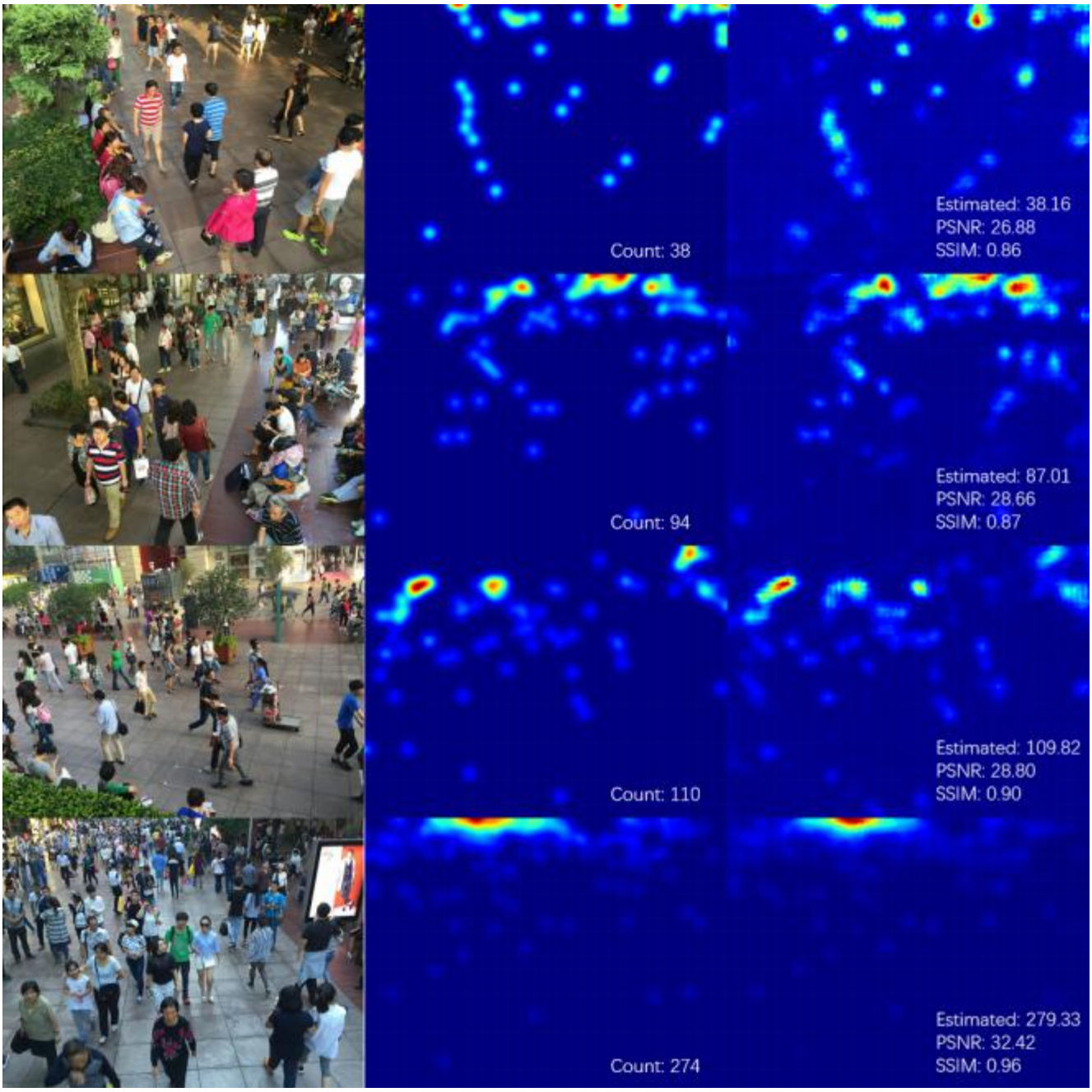

图8。由CSRNet从上海科技部分生成的样品_B [18]数据集。左列显示原始图像；中间列显示地面真实密度图，而右列表示我们生成的密度图。

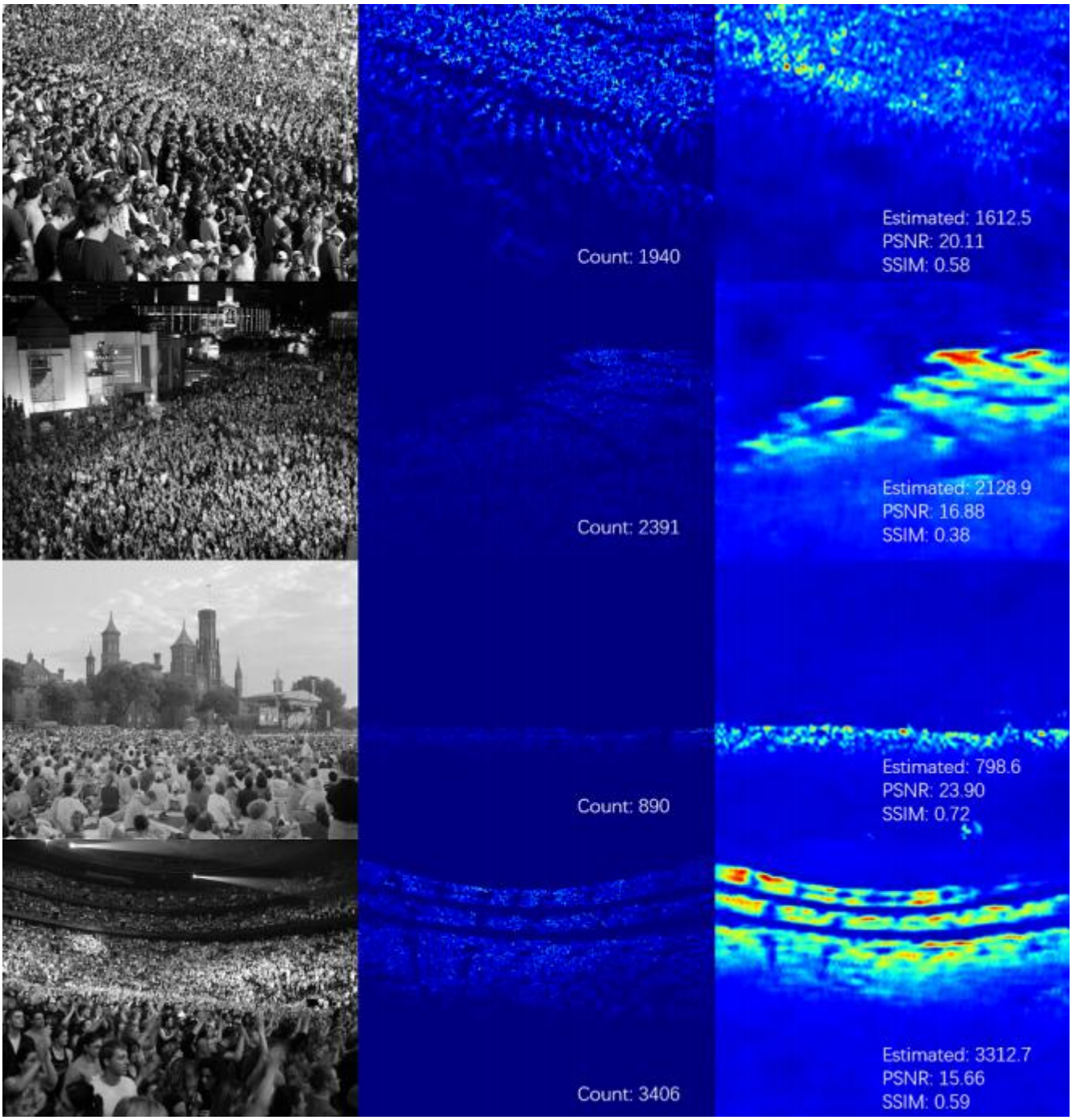

图9。由CSRNet从UCF中生成的样本_抄送_50 [22]数据集。左列显示原始图像；中间列显示地面真实密度图，而右列表示我们生成的密度图。

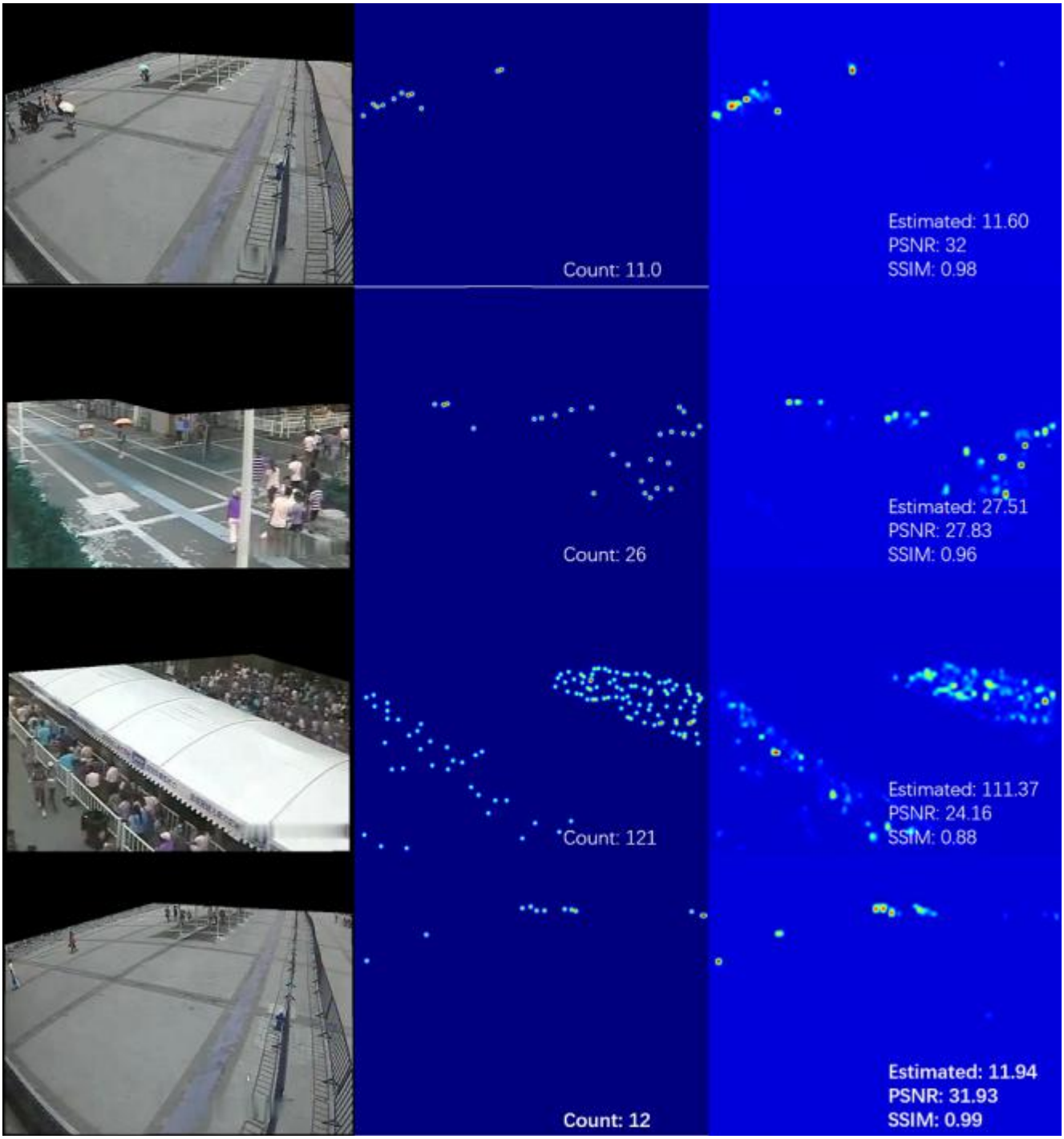

图10。由CSRNet从世博会的10个[3]数据集中生成的样本。左列显示了被ROI（感兴趣的区域）所掩盖的图像；中间列显示地面真实密度图，右列表示我们生成的密度图。

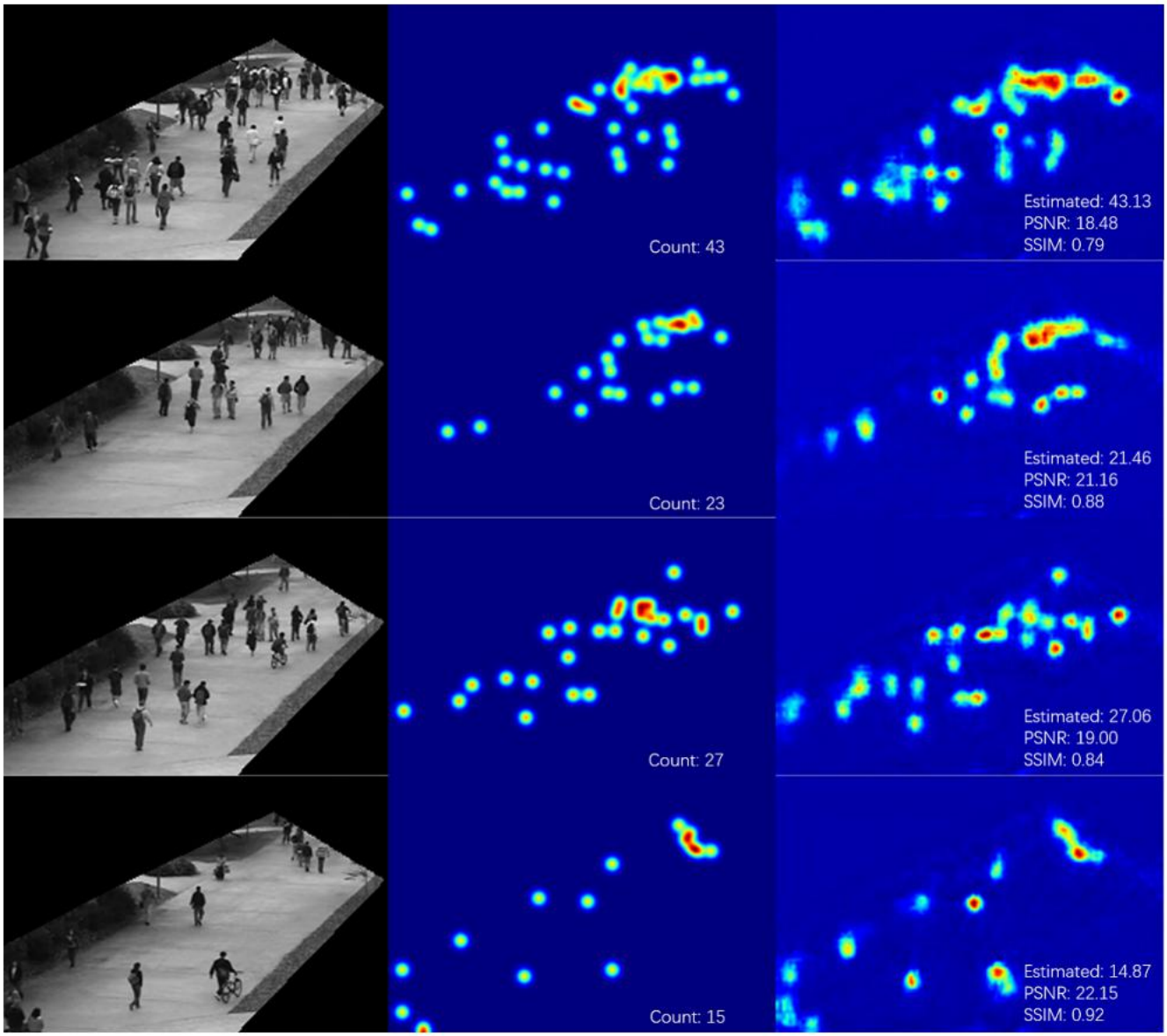

图11。由CSRNet从UCSD [23]数据集生成的样本。左列显示被ROI掩盖的图像；介质列显示地面真实密度图，右列表示我们生成的密度图。

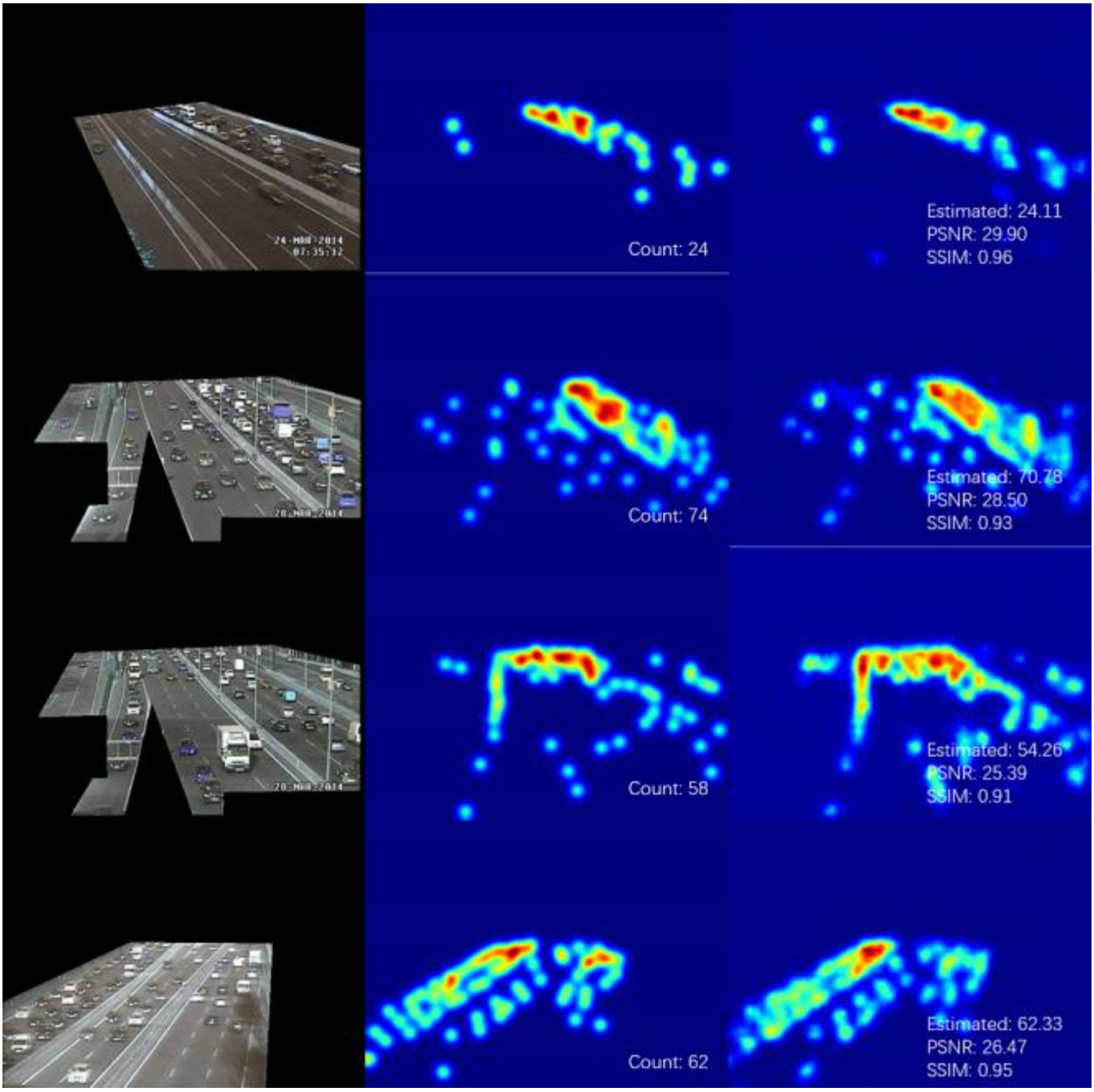

图12。由CSRNet从TRANCOS [44]数据集生成的样本。左列显示被ROI掩盖的图像；介质列显示地面真实密度图，而右列显示生成的密度图。