sensors

MDPI

Article

Congested Crowd Counting via Adaptive Multi-Scale Context Learning †

Yani Zhang [1], Huailin Zhao [2], Zuodong Duan [3], Liangjun Huang [1,*], Jiahao Deng [3] and Qing Zhang [1]

[1] School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai 201418, China; freezhangyani@gmail.com (Y.Z.); zhangqing0329@gmail.com (Q.Z.)
[2] School of Electrical and Electronic Engineering, Shanghai Institute of Technology, Shanghai 201418, China; zhao_huailin@yahoo.com
[3] School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China; zduan@bit.edu.cn (Z.D.); bitdjh@bit.edu.cn (J.D.)
* Correspondence: gene_huang@sit.edu.cn
† This paper is an extended version of the conference paper. Zhang, Y.; Zhao, H.; Zhou, F.; Zhang, Q.; Shi, Y.; Liang, L. MSCANet: Adaptive multi-scale context aggregation network for congested crowd counting. In *Proceedings of the 27th International Conference on Multimedia Modeling*; Springer: Prague, Czech Republic, 2021; pp. 1–12.

Abstract: In this paper, we propose a novel congested crowd counting network for crowd density estimation, i.e., the Adaptive Multi-scale Context Aggregation Network (MSCANet). MSCANet efficiently leverages the spatial context information to accomplish crowd density estimation in a complicated crowd scene. To achieve this, a multi-scale context learning block, called the Multi-scale Context Aggregation module (MSCA), is proposed to first extract different scale information and then adaptively aggregate it to capture the full scale of the crowd. Employing multiple MSCAs in a cascaded manner, the MSCANet can deeply utilize the spatial context information and modulate preliminary features into more distinguishing and scale-sensitive features, which are finally applied to a 1×1 convolution operation to obtain the crowd density results. Extensive experiments on three challenging crowd counting benchmarks showed that our model yielded compelling performance against the other state-of-the-art methods. To thoroughly prove the generality of MSCANet, we extend our method to two relevant tasks: crowd localization and remote sensing object counting. The extension experiment results also confirmed the effectiveness of MSCANet.

Keywords: crowd counting; crowd density estimation; multi-scale context learning; crowd localization; remote sensing object counting

Citation: Zhang, Y.; Zhao, H.; Duan, Z.; Huang, L.; Deng, J.; Zhang, Q. Congested Crowd Counting via Adaptive Multi-Scale Context Learning. *Sensors* **2021**, *21*, 3777. https://doi.org/10.3390/s21113777

Received: 11 April 2021
Accepted: 27 May 2021
Published: 29 May 2021

1. Introduction

Crowd counting is an indispensable component for smart crowd analysis, to count the number of people and describe the crowd distribution. It plays a critical role in many areas, such as video surveillance [1], public security [2], human behavior analysis [3,4], and smart cities [5–7]. However, due to the frequent occurrence of scale variations and severe occlusions, in addition to the diverse crowd distributions, the task often faces great difficulties to accurately describe the crowd, especially in scenes of overcrowding.

Deep-learning-based methods have been the main method for solving this problem and have achieved quite a few significant improvements. However, challenges remain to be settled. For one thing, the results of crowd counting are not sufficiently accurate in severe occlusions, scale variations, and diverse crowd distribution scenes, especially under the circumstances of crowds that visually share a high similarity with their surroundings, as illustrated in the first column of Figure 1.

One of the major causes is that few studies have focused on the leveraging of spatial context representation. For instance, single-scale crowd counting networks [8,9] only

employ convolution operations with a fixed kernel size, which may hurt the performance when the scale of the crowd changes. Multi-scale crowd counting networks [10–14] are carefully elaborate in order to portray different scales of people. They are still limited by the local receptive field of convolutional operation, and the features of the global spatial context cannot be fully utilized. Other studies [15,16] applied various modules to model scale-aware spatial context information; however, they merely aggregate different context features without any auxiliary processing, which cannot access the discriminative features and vastly harm the performance of the counting network.

Multi-scale context aggregation still has some space for improvement since only the typical features from a specific scale contribute to final crowd counting. We argue that the spatial context information of different scales should be aggregated in an adaptive way. For another, the estimated density maps are not reliable when considering the exact position even though the final reported count is precise. Unfortunately, in a majority of existing methods, precise crowd localization is rarely involved. Although, it is as significant as crowd counting since they are all fundamental tasks for crowd analysis.

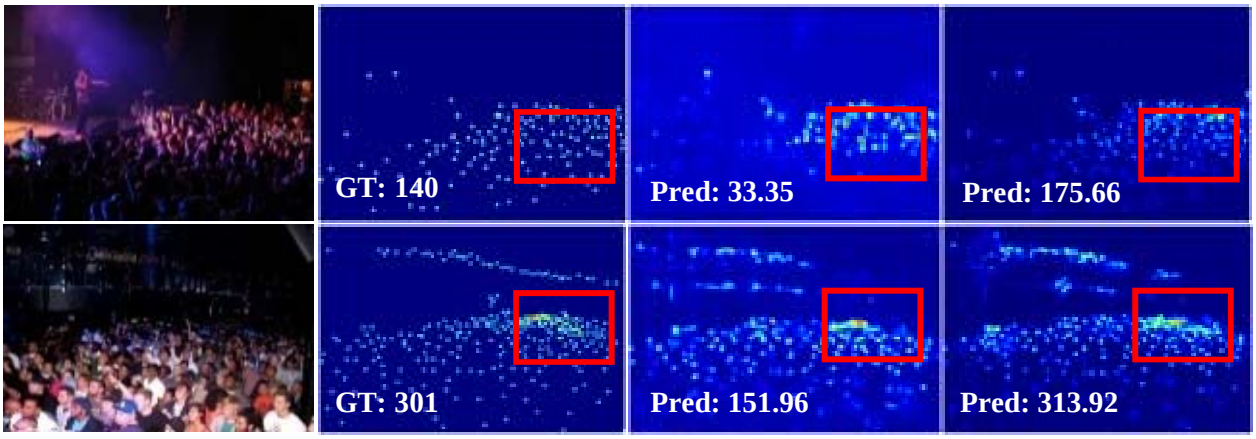

Figure 1. Representative examples in the UCF-QNRF dataset [17]. From left to right: input images, ground-truth, results of CSRNet [8], and the results of MSCANet. Compared to CSRNet, MSCANet can effectively handle the ambiguity of appearance between crowd and background objects.

Therefore, in this work, we propose a novel Adaptive Multi-scale Context learning mechanism for congested crowd counting and localization simultaneously, namely the Adaptive Multi-scale Context Aggregation Network (MSCANet). The kernel of the network is a Multi-scale Context Aggregation module (MSCA), which learns a multi-scale context representation in an adaptive way. MSCA introduces a multi-branch structure applying atrous convolution layers with different dilation rates aiming to encode multi-scale context features.

Then, the encoded features of the whole branches are aggregated layer by layer via a channel attention mechanism [18] to obtain a richer global scene representation. Multiple MSCAs concatenated in a cascaded manner are embedded in the MSCANet, where the subsequent up-sampling layer transforms the multi-scale features at each MSCA into higher-resolution representations. The high-level features from the last MSCA are further learned by a 1×1 convolution layer to output the two-channel results, including the crowd density map and crowd localization map.

MSCANet can be easily applied for various network backbones and learned in an end-to-end manner. Extensive experiments on three challenging public benchmarks (i.e., ShanghaiTech_Part_A, UCF_CC_50, and UCF-QNRF) showed that our model achieved compelling performance against the state-of-the-art methods. Additionally, to evaluate the generalization ability of our method, we extend MSCANet to two relevant tasks, i.e., crowd localization and remote sensing object counting. Our model was proven to generalize

well and achieved superior localization results on the UCF-QNRF dataset and promising counting results on the RSOC dataset.

In summary, the main contributions of this paper are two-fold:

- We propose a MSCA to adaptively aggregate small-scale context representation with large-scale context representation in a cascade manner, which encodes more compact global context features for crowds at various scales.
- Employing multiple MSCAs, we introduce the MSCANet to obtain multi-scale context features with different resolutions. This can efficiently address the ambiguous appearance challenge, especially under crowded scenes with complex backgrounds.

The remainder of this paper is organized as follows. Section 2 reviews related work regarding crowd counting and crowd localization. Section 3 presents the proposed method for crowd counting and localization. Section 4 introduces the experiment settings and presents extensive experiment results. In Section 5, we conclude this paper and with some future directions.

This paper is built on our conference paper [19], and the content is extended from three aspects: First, we give a comprehensive review about crowd counting, crowd localization, and remote sensing object counting. Secondly, to evaluate the effectiveness of our MSCANet, we also conduct a crowd localization experiment on the UCF-QNRF dataset. Our qualitative and quantitative results demonstrate the superiority of our method. Thirdly, we extend our MSCANet to remote sensing object counting tasks and conduct extensive experiments on RSOC. Our method achieves promising results compared with other state-of-the-art methods.

2. Related Works

In this section, we will review some related works regarding crowd counting, crowd localization, and remote sensing object counting.

2.1. Crowd Counting

The task of crowd counting has been studied for many years. Research of crowd counting can be categorized as either detection-based methods or regression-based methods. Detection-based methods usually employ pedestrian or face detectors to recognize and localize crowds. However, the performance of the detectors deteriorates in congested crowd scenes due to occlusions and large-scale variations of the crowd. Regression-based methods establish the correspondences between the input image and the number of people. Conventional methods [20–24] use carefully designed handcrafted features and apply different regression methods to regress the final count number. Although they achieved progress, their performances are constrained due to the handcrafted features of their methods, which heavily rely on the specific crowd scenes.

Recently, with the renaissance of deep learning, many CNN-based crowd counting networks have been proposed, which cast the crowd counting problem as a crowd density estimation task. The research of CNN-based crowd counting methods is primarily threefold: the design of the network architecture, the generation of the crowd density map, and the network optimization function. We will review the related work from the above three aspects as follows.

Network structure. Scale variation of the crowd head is a classical challenging problem of accurate crowd counting. Many counting networks [25–33] have been carefully designed to extract multi-scale features for handling this challenge. Early crowd counting networks typically employed multi-column structures [10,11,14,16,34] to model different scales of crowds. More recently, a graph network [35] was introduced to enhance scale-aware features. Perspective information of crowd scenes was also employed for networks [36,37] for improving the final counting performance. Later, research efforts were devoted to utilize context information efficiently.

For example, [38–40] proposed a crowd density classifier to provide each input image with a density-level label. The authors in [8] employed dilated convolutional layers

to enlarge the receptive field of the network for extracting context information. Other researchers [15] used spatial pyramid pooling [41] to enhance the different scales of context features for crowd counting. Benefiting from the efficiency of visual attention mechanisms for context information extracting [42–45], many attention-based counting networks [46–49] were designed and perform well on complicated crowd scenes in which the background objects have a similar appearance with foreground crowd.

Furthermore, to alleviate the effects of background objects for foreground crowd counting, foreground mask-based crowd counting networks [50–53] have been designed. Although the above methods achieved promising results, they rely on training data, and therefore their generalization ability is limited to new scenarios. Thus, some unsupervised domain-adaption methods [54,55] were developed for crowd counting and achieved satisfactory results.

Crowd density map generation. The density functions are considered as real-valued functions over pixel grids [56], whose integrals over image regions should match the object counts. Most CNN-based counting networks [9] applied a normalized 2-D Gaussian kernel to convolve with the head location for generating the crowd density map. Although they have achieved great performance, the density map generated by the normalized Gaussian kernel does not consider perspective changes, and thus cannot correctly model the crowd distribution, which hampers the performance of counting networks.

To solve this problem, Zhang et al. [10] employed geometry-adaptive kernels to solve the effects of perspective. Wan et al. [57] proposed a generation network to output the crowd density maps, which the counting network aims to optimize, and the counting network and generation network were trained end to end together. A. Sindagi et al. applied residual learning in a progressive fashion [58] to generate high-quality crowd density maps, and employed the MRF framework [59] to generate scale-aware density maps.

Optimization function. L2 loss was commonly used as the loss function in the CNN-based crowd counting method. However, its average effect led to blurry estimation and reduced the quality of the density map. Wan et al. [60] argued that the point annotations in the available crowd counting datasets could be considered as weak labels for density map estimation and proposed the Bayesian Loss, which constructs a density contribution probability model from the point annotations.

Adversarial Loss [61,62] was involved a Generator G and Discriminator D playing a two-player minimax game: G was trained to generate images to fool D while D was trained to distinguish synthetic images from the ground truth. It could avoid blur as well as incentivize sharp images since blurry outputs appear as unrealistic. Composition loss [17] was used for training and estimation of the three interrelated problems of counting, density map estimation, and localization, simultaneously. As a result, density maps can be "sharpened" until they approximate the localization map, whose integral should equal the true count. Cheng et al. [63] proposed a Maximum Excess over Pixels loss to learn spatial-aware crowd features.

2.2. Crowd Localization

Different from crowd counting, the task of crowd localization aims to acquire the exact locations of people in the image. It is also very challenging because people are very close to each other in the congested crowd scene. The methods of crowd localization can be divided into three categories: anchor-based localization methods, point-based localization methods, and heuristic-based localization methods.

Anchor-based Localization Methods. The anchor-based crowd localization methods draws on object detection, which designs a model to regress to the anchor box laid out by each person in advance. For instance, Liu et al. [64] proposed a DetNet based on Faster R-CNN [65] to detect sparse crowds. Lin et al. [66] employed the crowd density maps and scene depth maps to improve the detection performance of RetinaNet [67] for crowds. He et al. [68] utilized YoloV3 to detect crowds in the nearby region.

Point-based localization methods. Most crowd counting datasets only provide point annotations rather than anchor annotations. Therefore, it is more convenient to use point annotations as the supervision information for crowd localization. Specifically, they [69–72] formulated the crowd localization problem as a foreground/background segmentation problem and used the cross-entropy loss to optimize the network.

Heuristic-based localization methods. The heuristic-based localization methods [17,73,74] were proposed to obtain the crowd locations from the crowd density map. In particular, they usually adopt the non-maxima suppression to obtain the maximum local value, which presents each head location in the crowd. Then, the extracted locations are matched with true head locations by 1–1 matching. The feasible solutions are obtained via the Hungary algorithm for evaluating the performance of crowd locations.

2.3. Remote Sensing Object Counting

Remote sensing object counting, which aims to estimate the number of ground objects from remote sensing images, is a challenging and important computer vision task. Comparing with traditional object counting in natural scenes, the task of remote sensing object counting is more challenging in several aspects: large-scale variation, extremely complex backgrounds, and orientation arbitrariness [75]. It is an important way to obtain counting information by combining classification, detection, or segmentation results in remote sensing images.

For example, Bazi et al. [76] proposed an automatic method that contained a classification step using a Gaussian process classifier (GPC) and a counting step for counting olive trees in very high spatial remote sensing images. Santoro et al. [77] proposed a four-step algorithm that consisted of an asymmetrical smoothing filter, local minimum filter, mask layer, and spatial aggregation operator for tree counting. Xue et al. [78] applied a semi-supervised method for counting mammals in the open savanna. A parallel architecture was proposed by [79] to count olive trees in a crop field, which mainly uses color-based or stereo vision-based segmentation.

In recent years, deep learning methods have dominated the remote sensing object counting task. Mubin et al. [80] proposed a deep learning framework based on LeNet to detect and count oil palm trees in remote sensing images. Shao et al. [81] proposed a detection and counting system based on Yolo V2 [82] for cattle counting. A neural network named ResCeption was proposed by [83] to count cars by regression, which combined residual learning with inception layers. Context sensing is helpful for many applications (e.g., behaviour recognition [84]), and is also important for remote sensing.

Layout Proposal Networks (LPNs) with spatial kernels were proposed to count and locate cars in drone videos, which can leverage spatial context information effectively [85]. For congested remote sensing object counting scenes, the density map-based methods are more effective than detection-based methods.

Gao et al. [86] proposed an ASPD-Net for remote sensing object counting in an encoder–decoder framework. To deal with the shortcomings of hand-crafted methods used for generating density maps, an adaptive density map generator [87] was proposed for learning a density map representation for the counter, which adopted the annotation dot information as the input. The generator and counter were trained jointly in an end-to-end manner and had good performance in remote sensing object counting.

3. Proposed Method

In this section, we will first introduce the problem formulation of crowd counting in this paper. Then, we describe the details of our proposed MSCA module. After that, our MSCANet and the comparisons of different context modules from the crowd counting network are presented. Finally, the details of MSCANet for extension tasks (i.e., crowd localization and remote sensing object counting) are illustrated in detail.

3.1. Problem Formulation

We formulate crowd counting compliance with [8,10], which considers the problem as a pixel-wise regression problem. To be specific, the density map F_i is formed as follows:

$$F_i(x) = \sum_{j=1}^{M} \delta(\boldsymbol{x} - \boldsymbol{a}_j) \times G_\sigma(\boldsymbol{x}), \tag{1}$$

where $\delta(\cdot)$ stands for the Dirac delta function, G_σ represents the 2-D normalized Gaussian kernel, σ denotes the standard deviation, a_j is the head location, and M is the total crowd number of I_i. The crowd counting network learns the non-linear mapping between the input image I_i and its corresponding crowd density map F_i. The L_2 loss is defined as the network loss function:

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \|F(I_i;\Theta) - F_i\|_2^2, \tag{2}$$

where Θ represents the learning parameters of MSCANet, and N and $F(I_i;\Theta)$ denote the image number and the output of crowd counting network, respectively. More technically, in this paper, we introduce a new multi-scale contextual feature aggregation method, i.e., MSCA. The details are described in the next subsection.

3.2. Multi-Scale Context Aggregation Module

Making full use of contextual features at different scales is an effective way to address the scale variation of people. However, small-scale context features can only represent partial cues due to the limitations of receptive fields. It is ineffective to directly aggregate the small-scale context features with large-scale context features, which will introduce irrelevant and useless cues and hinder the counting performance. Thus, we resort to a selection mechanism to adaptively select and transform typical small-scale context features for aggregating them with large-scale context features. According to this consideration, we propose a MSCA module, and its specific structure is shown in Figure 2.

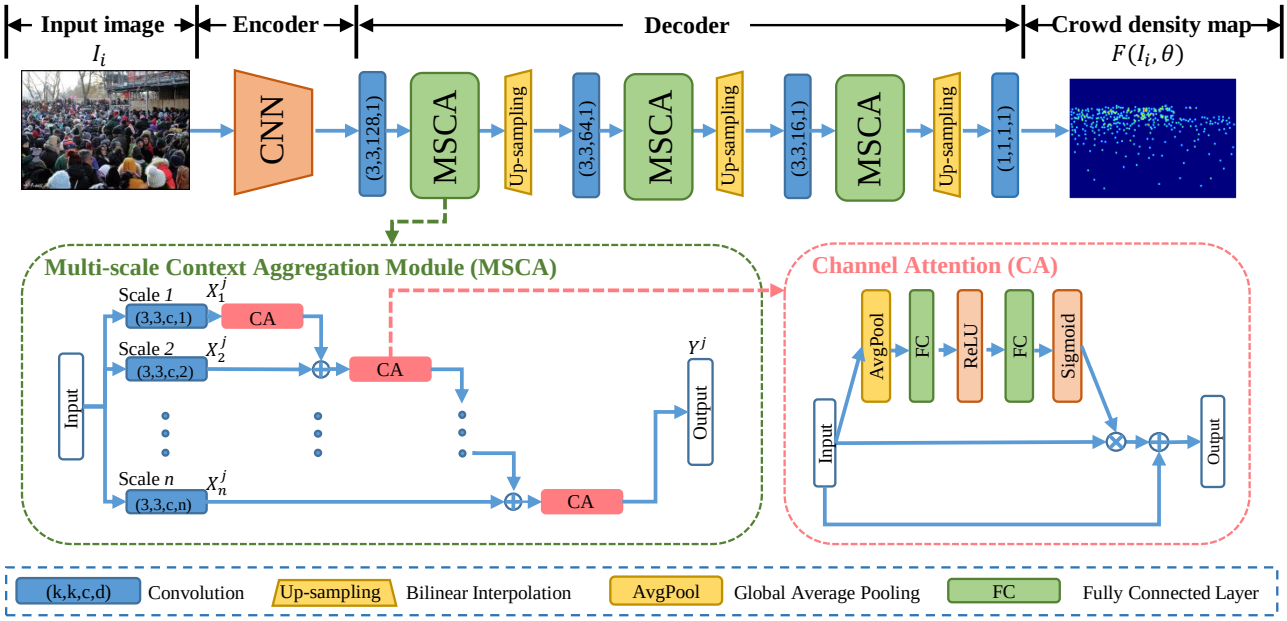

Figure 2. Detailedillustration of our Adaptive Multi-scale Context Aggregation Network for crowd counting.

The MSCA module was designed as a unified multi-branch atrous convolution layer, where each layer has a different dilated rate. Concretely, we denote i, r, and $j \in \{\frac{1}{2^{r-1}} \cdots \frac{1}{4}, \frac{1}{2}, 1\}$ as the dilated rate, reduction ratio, and resolution of the feature map, respectively. The context feature is represented by $X_i^j \in \mathbb{R}^{jW \times jH \times C}$. we adopt a function

f being responsible for selecting informative features from X_i^j. The context features are aggregated as follows:

$$Y^j = f(\cdots f(f(X_1^j) \oplus X_2^j) \oplus X_3^j) \oplus \cdots \oplus X_n^j), \tag{3}$$

where $\oplus$ represents the element-wise summation and $Y^j \in \mathbb{R}^{jW \times jH \times C}$ denotes the output feature of MSCA module. Specifically, We employed a channel attention(CA) [18] to instantiate the selecting function f without extra supervision information. As illustrated in Figure 2, the context feature is first sent to a global average pooling (F_{avg}) layer and subsequently processed by a bottleneck structure consisting of two fully connected layers. Finally, a sigmoid function is applied to normalize the output feature. The selecting operation not only highlights the typical features but also suppresses possible noise existing in the redundant features. The detailed process is as follows:

$$\alpha_i = W_2^{fc}(W_1^{fc}(F_{avg}(X_i^j))), \tag{4}$$

where $\alpha_i \in \mathbb{R}^{jW \times jH \times C}$ denotes the adaptive coefficient. W_1^{fc} and W_2^{fc} represent the weights of the two fully connected layers, and the first fully connected layer is followed by a ReLU function. For better optimization, a residual connection is adopted between the input and output of CA. The residual equation is as follows:

$$f(X_i^j) = X_i^j + \alpha_i X_i^j, \quad i = 1 \cdots n. \tag{5}$$

We summarize the computation process of MSCA and give its pseudocode as shown in Algorithm 1.

Algorithm 1 Pseudocode of Multi-scale Context Aggregation Module with three branches in a PyTorch-like style.

```
############################ initialization ############################
branch1 = nn.Conv2d(in_channels, out_channels, kernel = 3, padding = 1, dilation = 1)
branch2 = nn.Conv2d(in_channels, out_channels, kernel = 3, padding = 2, dilation = 2)
branch3 = nn.Conv2d(in_channels, out_channels, kernel = 3, padding = 3, dilation = 3)
avg_pool = nn.AdaptiveAvgPool2d(1)
CA1 = nn.Sequential( nn.Linear(out_channels, out_channels // 4, bias = False), nn.ReLU(inplace = True),
nn.Linear(out_channels // 4, out_channels, bias = False), nn.Sigmoid() )
CA2 = nn.Sequential( nn.Linear(out_channels, out_channels // 4, bias = False), nn.ReLU(inplace = True),
nn.Linear(out_channels // 4, out_channels, bias = False), nn.Sigmoid() )
CA3 = nn.Sequential( nn.Linear(out_channels, out_channels // 4, bias = False), nn.ReLU(inplace = True),
nn.Linear(out_channels // 4, out_channels, bias = False), nn.Sigmoid() )

############################ forward pass ############################
feature1 = branch1(x), feature2 = branch2(x) , feature3 = branch3(x)
b, c, _, _ = feature1.size()
y = avg_pool(feature1).view(b, c)
y = CA1(y).view(b, c, 1, 1)## Channel attention, Equation (4)
channel_attention_map1 = y.expand_as(feature1)
feature1 = feature1 * (1 + channel_attention_map1) ## Residual learning, Equation (5)
feature2 = feature2 + feature1## Context feature aggregation, Equation (3)
b, c, _, _ = feature2.size()
y = avg_pool(feature2).view(b, c)
y = CA2(y).view(b, c, 1, 1)## Channel attention, Equation (4)
channel_attention_map2 = y.expand_as(feature2)
feature2 = feature2 * (1 + channel_attention_map2)## Residual learning, Equation (5)
feature3 = feature3 + feature2## Context feature aggregation, Equation (3)
b, c, _, _ = feature3.size()
y = avg_pool(feature3).view(b, c)
y = CA3(y).view(b, c, 1, 1)## Channel attention, Equation (4)
channel_attention_map3 = y.expand_as(feature3)
feature3 = feature3 * (1 + channel_attention_map3)## Residual learning, Equation (5)
return feature3
```

3.3. Multi-Scale Context Aggregation Network

Based on MSCA, we propose an end-to-end deep neural network, i.e., MSCANet, for congested crowd counting, which leverages context cues to effectively bootstrap the task of crowd counting and localization. The pipeline is shown in Figure 2. Given an input image I_i, we first use a CNN to encode features. Then, the encoding features are fed into multiple MSCA modules aimed to obtaining ample scale information. Specifically, we employ an up-sampling layer following each MSCA to gradually transform the multi-scale context feature map into higher-resolution representations. Finally, a convolution operation is performed on the learned multi-scale context features with a 1×1 convolution kernel for predicting the crowd density map.

3.4. Compared to Other Context Modules

We compare MSCA with another three context modules from [15,16,88], as shown in Figure 3. To obtain a compact context feature, the Cascade Context Pyramid Module (CCPM) [88] progressively aggregates large-scale contextual representation with small-scale contextual representation, as shown in Figure 3b. The CCPM block enhances the context features as follows:

$$Y^j = g(\cdots g(g(X_n^j \oplus X_{n-1}^j) \oplus X_{n-2}^j) \oplus \cdots \oplus X_1^j), \tag{6}$$

where $g(\cdot)$ denotes the residual block (res) from [89]. In contrast to CCPM, we fuse contextual features from small to large in an adaptive way.

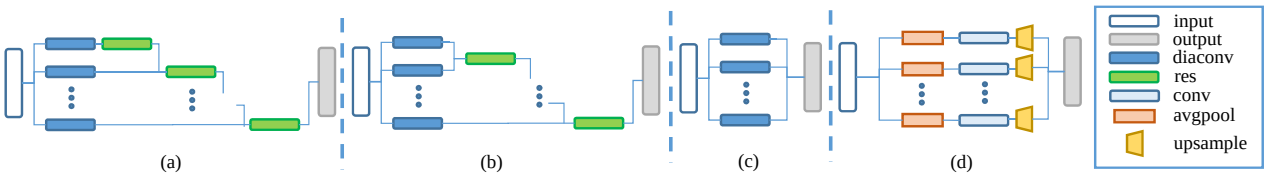

Figure 3. Different structures of multi-scale context modules. (**a**) Multi-scale context aggregation module (MSCA) w/o channel attention (CA); (**b**) cascade context pyramid module (CCPM); (**c**) scale pyramid module (SPM); and (**d**) scale-aware context module (SACM).

A Spatial Pyramid Module (SPM) [16] first adopts a multi-branch atrous convolution layer to encode context information. Then, the output feature of each branch is equally summated by an element-wise sum operation, as shown in Figure 3c. The learning process of SPM is as follows:

$$Y^j = \sum_{i=1}^{n} X_i^j = \sum_{i=1}^{n} W_i^{diaconv}(U), \tag{7}$$

where $U \in \mathbb{R}^{W \times H \times C}$ and $W_i^{diaconv}$ denote the input features and weights of the dilated convolution layers, respectively. Differently from SPM, rgw MSCA module adaptively selects reliable information from different scales of context information.

Liu et al. [15] employed spatial pyramid pooling [90] to capture multi-scale context features from local features, and then the contrast features were extracted from the differences between local features and multi-scale context features to enhance the representation of people at different scales. Referring to the above method, we introduce a Scale-Aware Context Module (SACM) for crowd counting as shown in Figure 3d. The SACM outputs context features as follows:

$$Y^j = \sum_{i=1}^{n} X_i^j = \sum_{i=1}^{n} U_p(W_i^{conv}(P_{ave_i}(U))), \tag{8}$$

where $P_{ave_i}(\cdot)$, W_i^{conv}, and U_p represent the adaptive average pooling layer that averages the input feature U into $i \times i$ blocks and the weights of the convolution layers and bilinear interpolation operation for upsampling, respectively. Compared to SACM, we apply a different way to encode scale-aware context features. The experiments in the next section verify the superiority of our MSCA module.

3.5. Extension of MSCANet

We extend our MSCANet to two relevant tasks: crowd localization and remote sensing object counting. The former aims to obtain the exact locations of the crowd, and the latter aims to obtain the accurate number of remote sensing objects from remote sensing images.

Crowd Localization. Following [17,73,74], we also obtain the crowd localization results from the crowd density map. Specifically, we first apply our MSCANet to generate the original density map. Then, we utilize the non-maximum suppression to process the extracted crowd density map to obtain the local maximum response map, which is our final crowd localization results. The comparisons are illustrated in Section 4.3.

Remote Sensing Object Counting. Given that remote sensing object counting has more similarities with crowd counting, we also formulate remote sensing object counting tasks as a density estimation problem. Thus, we use the annotations from the remote sensing object counting dataset to generate a density map following Section 3.1, and directly train our MSCANet on it. The detailed comparison results will be presented in Section 4.3.

4. Experiments

In this section, we first introduce the datasets and implementation details. Then, we describe the evaluation metrics for crowd/remote sensing object counting and crowd localization. After that, the comparison results on test sets of different benchmarks between our MSCANet and other state-of-the-art methods for crowd counting, crowd localization, and remote sensing object counting are presented. Finally, comprehensive ablation studies were performed to evaluate the effectiveness of each component of MSCANet.

4.1. Datasets

We conducted comprehensive experiments on four popular datasets, i.e., ShanghaiTech_Part_A [10], UCF_CC_50 [91], UCF-QNRF [17], and RSOC [86]:

ShanghaiTech_Part_A [10] consists of 482 images in total (300 images for training and 182 images for testing). The crowd density varies significantly between different crowd images. Specifically, the minimum number of people is 33 while the maximum is 3139, which poses a difficult challenge for accurate estimation.

UCF_CC_50 [91] contains 50 images, which are randomly crawled from the internet, and the maximum number of people is equal to 4543. Limited training images, and different perspectives and resolutions are challenging factors for crowd counting methods. We follow the standard setting in [91] to conduct a five-fold cross-validation.

UCF-QNRF [17] is a new proposed dataset, which has great improvement in the quantity and quality of crowd images. The total number of images is 1535, including 1201 training images and 334 testing images. The number of people in the UCF-QNRF dataset varies from 49 to 12,865.

RSOC [86] is the largest remote sensing object counting dataset, which contains 3057 images with 286,539 instances. It consists of four types of remote sensing objects, i.e., Building, Small-Vehicle, Large-Vehicle, and Ship, and the number of remote sensing object varies significantly.

We used the first ten layers of VGG-16 pre-trained on ImageNet as the backbone. The initial learning rate was 1×10^{-5}, and the optimizer was SGD with momentum. All experiments were performed on a C^3 Framework [92,93] with a single RTX 2080 Ti GPU card and an Intel(R) Core(TM) i7-8700 CPU with 16 GB RAM and 512 GB ROM. The experiment software environments were the Pytorch 1.1 framework, Python 3.6, CUDA 10.1, and Ubuntu 18.04 LTS operation system. The data pre-processing and augmentation strategies of the above three datasets all follow the C^3 Framework. The training batch size was set to 4 and 1 on *UCF_CC_50* and the other datasets, respectively.

4.2. Evaluation Metrics

Counting Metrics. The mean absolute error (MAE) and mean squared error (MSE) were applied to evaluate the counting performance:

$$\text{MAE} = \frac{1}{N}\sum_{1}^{N}|z_i - \hat{z}_i|, \tag{9}$$

$$\text{MSE} = \sqrt{\frac{1}{N}\sum_{1}^{N}(z_i - \hat{z}_i)^2}, \tag{10}$$

where z_i and $\hat{z}_i$ denote the truth number and the predicted number of people in image I_i respectively.

Localization Metrics. For the crowd localization task, we adopted the precision (P), recall (R), and F1-measure (F1) to evaluate the localization performance:

$$P = \frac{TP}{TP + FP}, \tag{11}$$

$$R = \frac{TP}{TP + FN}, \tag{12}$$

$$\text{F1} = \frac{2 \cdot P \cdot R}{P + R}, \tag{13}$$

where *TP*, *FP*, and *FN* denote the number of true positive samples, false positive samples, and false negative samples, respectively. Specifically, the extracted crowd localization points were matched with ground-truth points by 1–1 matching, and the *TP*, *FP*, and *FN* were calculated under the pixel distance threshold value from 1 to 100 pixels. If the distance between the extracted point and the ground truth point was less than the pixel distance value, the localization result was marked as *TP*; if the distance between the extracted point and the ground truth point was larger than the pixel distance value, the localization result was marked as *FP*; if there existed no matched extracted point with the ground truth point, the localization result was marked as *FN*.

4.3. Comparison with State-of-the-Arts

4.3.1. Crowd Counting

We compare our MSCANet with the top performing methods [8–10,17,38,39,56,70,91,94,95] on four datasets, and the comparison results are reported in Table 1.

Table 1. Comparison of the different state-of-the-art methods on the *ShanghaiTech_Part_A*, *UCF_CC_50,* and *UCF-QNRF* datasets.

Method	SHA		UCF_CC_50		UCF-QNRF	
	MAE	**MSE**	**MAE**	**MSE**	**MAE**	**MSE**
Lempitsky et al. [56]	-	-	493.4	487.1	-	-
Zhang et al. [9,10]	181.8	277.7	467.0	498.5	-	-
Idrees et al. [17,91]	-	-	419.5	541.6	315	508
MCNN, [10,17]	110.2	173.2	377.6	509.1	277	-
Switching CNN, [17,38]	90.4	135.0	318.1	439.2	228	445
CL, [17]	-	-	-	-	132	191
CP-CNN, [39]	73.6	106.4	298.8	**320.9**	-	-
CSRNet(baseline), [8]	68.2	115.0	266.1	397.5	-	-
ic-CNN(one stage), [94]	69.8	117.3	-	-	-	-
ic-CNN(two stage), [94]	68.5	116.2	-	-	-	-
CFF, [70]	65.2	109.4	-	-	-	-
TEDNet, [95]	**64.2**	109.1	249.4	354.5	113	188
MSCANet (Ours)	66.5	**102.1**	**242.8**	329.8	**104.1**	**183.8**

Performance on ShanghaiTech_Part_A. We observed that our MSCANet achieved the best performance on MSE and competitive results on MAE compared to the other methods, which verifies the effectiveness of MSCANet. Specifically, it outperformed CSRNet by −1.7 and −12.9 in terms of MAE and MSE.

Performance on UCF_CC_50. Our model achieved the best performance on MAE and promising performance on MSE. More remarkably, MSCANet surpasses the performance of TEDNet [95] −6.56 and −24.68 on MAE and MSE, respectively.

Performance on UCF-QNRF. Our method produced the best results on both MAE and MSE and outperformed the second-best result, i.e., TEDNet [95], by −8.9 and −4.2 on the MAE and MSE metrics, respectively. The above improvements are due to the effect of MSCA, which can learn more multi-scale context features used for crowd counting.

4.3.2. Crowd Localization

We conducted a crowd localization task on the UCF-QNRF dataset. The quantitative results are presented in Table 2. The performance of MSCANet outperformed the other state-of-the-art crowd localization methods in terms of the F1-measure, which demonstrates that our model can efficiently obtain the crowd localization in different crowd scenes. Figure 4 presents the crowd localization results of MSCANet. We can see that our model performed well on different crowd scenes with different crowd distributions, which further proves the effectiveness of our MSCANet.

Figure 4. Visualizations of MSCANet for crowd localization on the UCF-QNRF dataset. Red points denote the ground-truth, and green points denote the estimated location results of MSCANet.

Table 2. Comparison of the localization results on the UCF-QNRF dataset.

Method	Av. Precision	Av. Recall	F1-Measure
MCNN [10]	59.93%	**63.50**%	61.66%
DenseNet63 [96]	70.19%	58.10%	63.87%
CL [17]	75.80%	59.75%	66.82%
SCLNet [74]	**83.99**%	57.62%	67.36%
MSCANet (Ours)	83.65%	61.07%	**69.64**%

4.3.3. Remote Sensing Object Counting

We perform our model on RSOC for remote sensing object counting. Table 3 displays the comparison results. We can see that our method achieves comparable results against other state-of-the-art methods. Specifically, MSCANet sets a new state-of-the-art result on *Small vehicle* and *Ship* and surpasses other state-of-the-art methods by a significant margin, which proves the effectiveness of our method. Figure 5 presents the qualitative results of our model. We find that the density map generated by MSCANet are very close to the ground truth density maps, which further prove the superiority of our model.

Table 3. Comparison of the different state-of-the-art methods on *RSOC*.

Method	Building		Small Vehicle		Large Vehicle		Ship	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [10,17]	13.65	16.56	488.65	1317.44	36.56	55.55	263.91	412.30
CMTL [97]	12.78	15.99	490.53	1321.11	61.02	78.25	251.17	403.07
CSRNet [8]	8.00	11.78	443.72	1252.22	34.10	46.42	240.01	394.81
SANet [34]	29.01	32.96	497.22	1276.66	62.78	79.65	302.37	436.91
SFCN [93]	8.94	12.87	440.70	1248.27	33.93	49.74	240.16	394.81
SPN [16]	7.74	11.48	445.16	1252.92	36.21	50.65	241.43	392.88
SCAR [98]	26.90	31.35	497.22	1276.65	62.78	79.64	302.37	436.92
CAN [15]	9.12	13.38	457.36	1260.39	34.56	49.63	282.69	423.44
SFANet [99]	8.18	11.75	435.29	1284.15	29.04	47.01	201.61	332.87
ASPDNet [86]	**7.59**	**10.66**	433.23	1238.61	**18.76**	**31.06**	193.83	318.95
MSCANet (Ours)	11.13	16.02	**221.16**	**430.90**	60.92	78.20	**41.93**	**60.73**

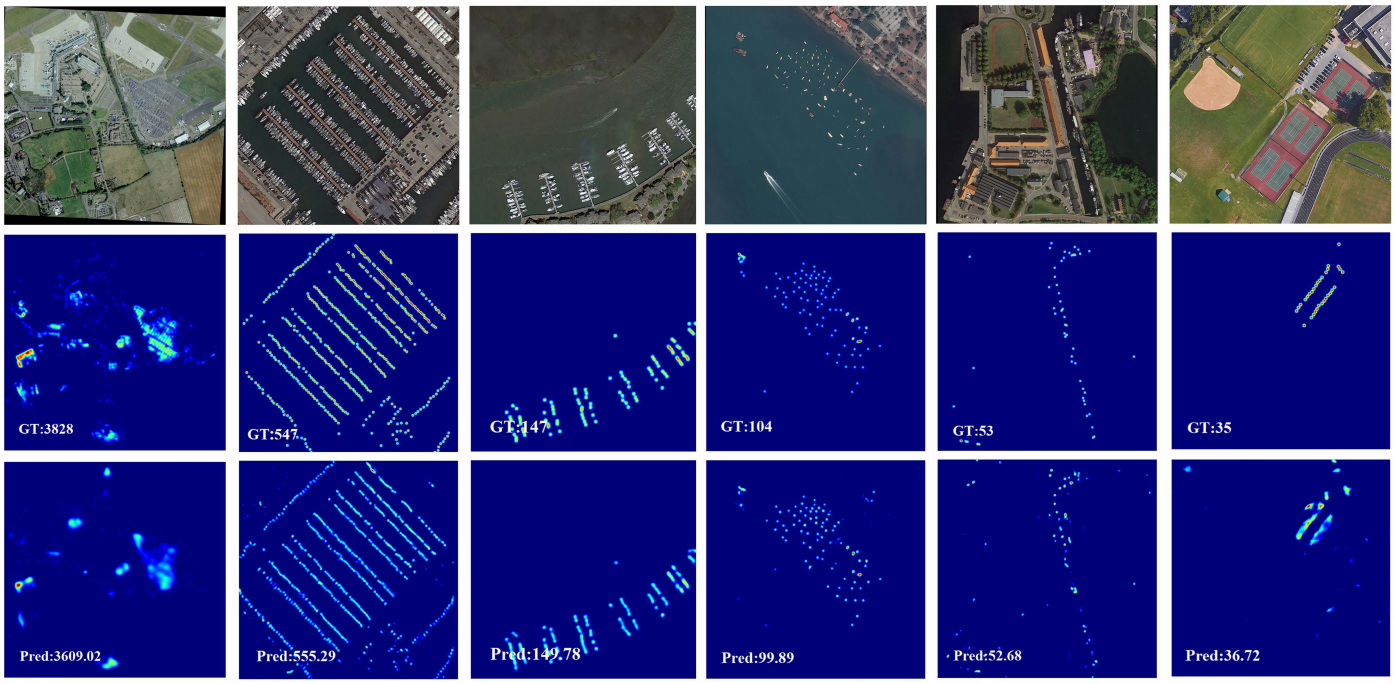

Figure 5. Visualization results of MSCANet for remote sensing object counting on RSOC dataset.

4.4. Ablation Study

4.4.1. Multi-Scale Context Aggregation Module

We first evaluated the performance of MSCANet with different pyramid scale settings. The pyramid scale setting (PS) denotes what dilated convolution branches are used in MSCA module, and the value of the PS represents the dilated rate of each branch. We investigated different PS settings to determine a suitable combination. As shown in Table 4, the performance of MSCANet gradually improved as the parameter of PS increased, reaching saturation at PS $= \{1, 2, 3\}$.

Continually increasing the parameter of PS did not improve the performance of the network. This is mainly because a larger receptive field results in redundant information, which hinders the learning of multi-scale context representation. As shown in Figure 6, we visualized the output of MSCANet with different pyramid scale settings. The predicted results of PS $= \{1, 2, 3\}$ were very close to the ground truth. Based on this analysis, we set PS = {1,2,3} in the following experiments.

Table 4. Comparisons of our proposed method with different pyramid scale settings (PS) on the UCF-QNRF dataset. The value of PS is the dilated rate of each dilation convolution branch from MSCA.

PS	MAE	MSE
{1}	110.9	197.2
{1,2}	105.2	184.6
{1,2,3}	**104.1**	**183.8**
{1,2,3,4}	104.8	186.1

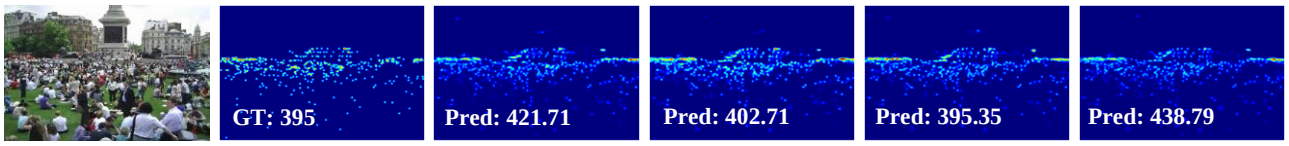

Figure 6. Impacts of different pyramid scale settings on UCF-QNRF. From left to right: input image, ground truth, result of PS $= \{1\}$, result of PS$= \{1,2\}$, result of PS $= \{1,2,3\}$, and result of PS$= \{1,2,3,4\}$.

We studied the effects of the MSCA block by comparing our full model to one of the same architecture without MSCA, denoted as MSCANet w/o MSCA (Decoder). Moreover, to measure the effectiveness of CA for feature aggregation, we designed another network variant, MSCA w/o CA, by replacing the CA block with a simple residual block. Table 5 reports the comparison results of the above changes. MSCA outperformed MSCA w/o CA and Decoder in terms of MAE. The visual results in Figure 7 show the impacts of CA. We can see that MSCA w/o CA performed worse than MSCA, which further verifies the importance of CA in MSCANet.

Table 5. Comparisons of our proposed method with different architecture changes on the UCF-QNRF dataset.

Configuration	MAE	MSE
Decoder (baseline)	111.3	**182.0**
MSCA w/o CA	105.7	186.9
MSCA	**104.1**	183.8
CSRNet (our reimplementation)	118.8	204.4
CAN [15]	107.0	183.0
CCPM	111.9	182.3
SPM	108.1	187.2
SACM	116.2	211.2

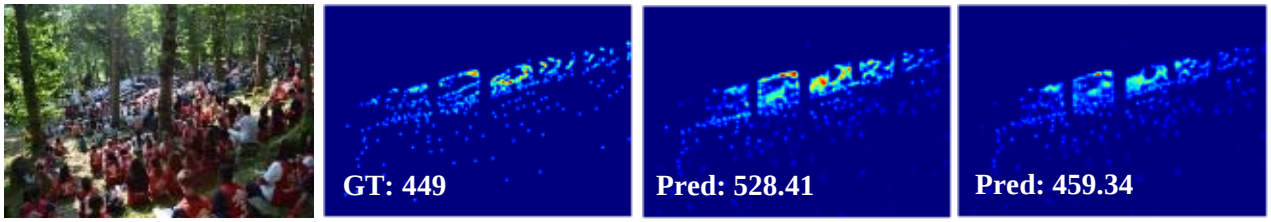

Figure 7. Impacts of CA on UCF-QNRF. From left to right: input image, ground-truth, result of MSCA w/o CA, and result of MSCA.

4.4.2. Multi-Scale Context Modules

We compared our MSCANet with the other prominent context-based crowd counting networks, i.e., *Congested Scene Recognition Network* (CSRNet) [8] and *Context-aware Network* (CAN) [15], which also employ the first 10 layers of VGG-16 pre-trained on ImageNet as a

backbone. The detailed results are reported in Table 5. Our MSCANet achieved the top performance on both the MAE and MSE metrics.

Then, we studied the influence of using MSCA, CCPM, SPM, and SACM. For a fair comparison, all of them had three branch structures, and the feature extractor was the same as MSCANet. The comparison results are shown in Table 5. MSCA achieved the lowest MAE on the UCF-QNRF dataset. Figure 8 displays the predicted results of typical images with different crowd density levels. The qualitative and quantitative results demonstrate that the MSCA block was critical for our model to improve performance, especially in congested scenery.

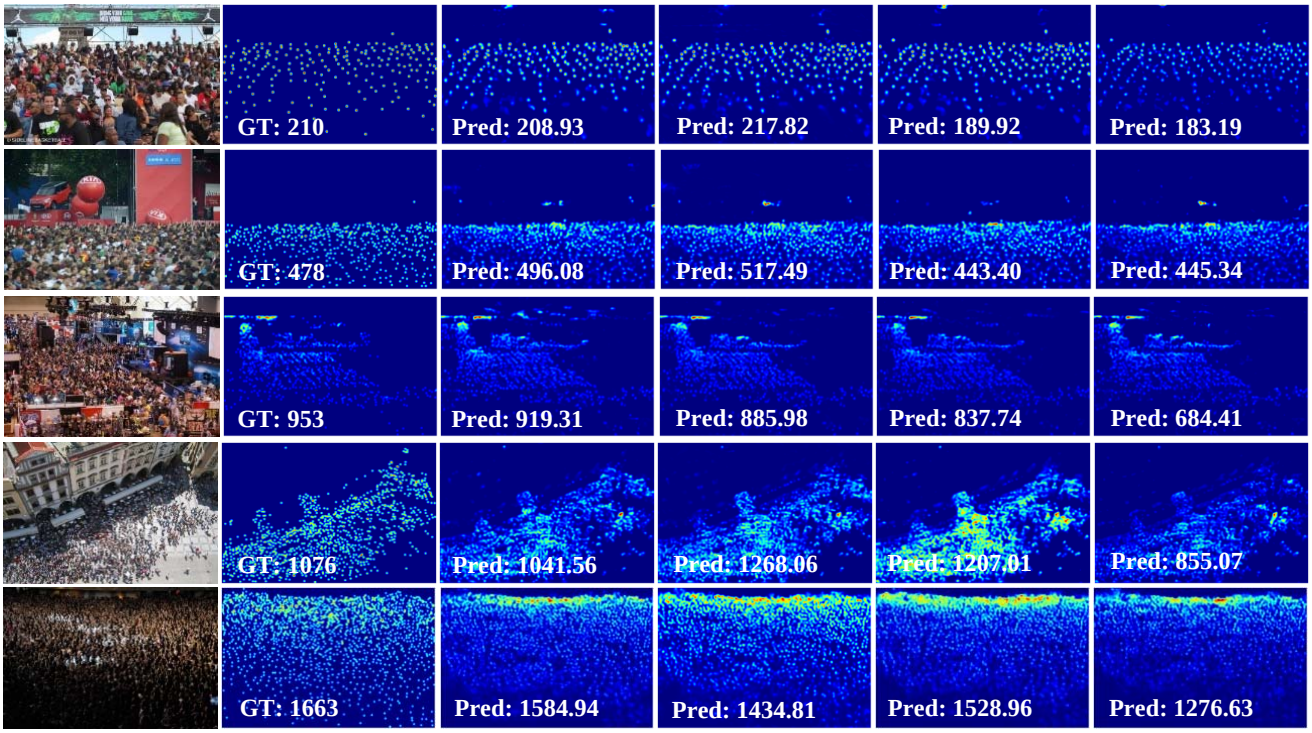

Figure 8. Visual comparision of different multi-scale context modules on UCF-QNRF. From left to right: input images, ground-truth, results of our method, results of CCPM, results of SPM, and results of SACM.

5. Conclusions and Future Work

In this paper, we proposed a novel MSCANet for congested crowd counting. The core of MSCANet is the MSCA block, which consists of multi-branch atrous convolution layers and channel attention modules. The atrous convolution layers aim to extract multi-scale contextual features while channel attention modules contribute to filter the redundancy features and highlight the features that are beneficial for crowd counting. Extensive experiments were performed on three congested crowd datasets, and our MSCANet achieved favorable results against the other prominent methods. Moreover, we extended our model to two relevant tasks, i.e., crowd localization and remote sensing object counting. The experimental results on UCF-QNRF and RSOC demonstrated the generalization ability of MSCANet.

However, our model only utilizes the spatial context information of a single image, and the performance of MSCANet is limited for video object counting. In future work, we will extend our model with temporal context information for the task of video object counting. Specifically, we can first count each frame using our proposed MSCANet to obtain the count result of each frame. We can obtain the global information of the video sequence from the count result of each frame. Then, with the help of the global information, we can apply the rescore method to modify the unsatisfied count result of those frames. Finally, we obtain the counting number of the video from the estimated and refined count results.

Author Contributions: Conceptualization, Y.Z. and H.Z.; methodology, Y.Z., H.Z., and Z.D.; software, Y.Z. and Z.D.; validation, H.Z., L.H., J.D., and Q.Z.; formal analysis, Q.Z.; investigation, L.H.; resources, H.Z. and L.H.; data curation, H.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, H.Z., Z.D., L.H., J.D., and Q.Z.; visualization, Y.Z. and Z.D.; supervision, H.Z. and L.H.; project administration, H.Z.; funding acquisition, Q.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by Natural Science Foundation of Shanghai under Grant No. 19ZR1455300, and National Natural Science Foundation of China under Grant No. 61806126.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Reference

1. Yu, Y.; Huang, J.; Du, W.; Xiong, N. Design and analysis of a lightweight context fusion CNN scheme for crowd counting. *Sensors* **2019**, *19*, 2013. [CrossRef]
2. Tong, M.; Fan, L.; Nan, H.; Zhao, Y. Smart camera aware crowd counting via multiple task fractional stride deep learning. *Sensors* **2019**, *19*, 1346. [CrossRef] [PubMed]
3. Csönde, G.; Sekimoto, Y.; Kashiyama, T. Crowd counting with semantic scene segmentation in helicopter footage. *Sensors* **2020**, *20*, 4855. [CrossRef] [PubMed]
4. Ilyas, N.; Shahzad, A.; Kim, K. Convolutional-neural network-based image crowd counting: Review, categorization, analysis, and performance evaluation. *Sensors* **2020**, *20*, 43. [CrossRef]
5. Fortino, G.; Savaglio, C.; Spezzano, G.; Zhou, M. Internet of Things as System of Systems: A Review of Methodologies, Frameworks, Platforms, and Tools. *IEEE Trans. Syst. Man Cybern. Syst.* **2020**. [CrossRef]
6. Abualsaud, K.; Elfouly, T.M.; Khattab, T.; Yaacoub, E.; Ismail, L.S.; Ahmed, M.H.; Guizani, M. A survey on mobile crowd-sensing and its applications in the IoT era. *IEEE Access* **2018**, *7*, 3855–3881. [CrossRef]
7. Solmaz, G.; Wu, F.J.; Cirillo, F.; Kovacs, E.; Santana, J.R.; Sánchez, L.; Sotres, P.; Munoz, L. Toward understanding crowd mobility in smart cities through the internet of things. *IEEE Commun. Mag.* **2019**, *57*, 40–46. [CrossRef]
8. Li, Y.; Zhang, X.; Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1091–1100.
9. Zhang, C.; Li, H.; Wang, X.; Yang, X. Cross-scene crowd counting via deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 833–841.
10. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 589–597.
11. Onoro-Rubio, D.; López-Sastre, R.J. Towards perspective-free object counting with deep learning. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 615–629.
12. Deb, D.; Ventura, J. An aggregated multicolumn dilated convolution network for perspective-free counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018; pp. 195–204.
13. Wang, Z.; Xiao, Z.; Xie, K.; Qiu, Q.; Zhen, X.; Cao, X. In Defense of Single-column Networks for Crowd Counting. *arXiv* **2018**, arXiv:1808.06133.
14. Liu, N.; Long, Y.; Zou, C.; Niu, Q.; Pan, L.; Wu, H. ADCrowdNet: An Attention-injective Deformable Convolutional Network for Crowd Understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 3225–3234.
15. Liu, W.; Salzmann, M.; Fua, P. Context-Aware Crowd Counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 5099–5108.
16. Chen, X.; Bin, Y.; Sang, N.; Gao, C. Scale pyramid network for crowd counting. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 7–11 January 2019; pp. 1941–1950.
17. Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Maadeed, S.; Rajpoot, N.; Shah, M. Composition loss for counting, density map estimation and localization in dense crowds. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 532–546.
18. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.

19. Zhang, Y.; Zhao, H.; Zhou, F.; Zhang, Q.; Shi, Y.; Liang, L. MSCANet: Adaptive Multi-scale Context Aggregation Network for Congested Crowd Counting. In Proceedings of the 27th International Conference on Multimedia Modeling, Prague, Czech Republic, 22–24 June 2021; pp. 1–12.
20. Chan, A.B.; Liang, Z.S.J.; Vasconcelos, N. Privacy preserving crowd monitoring: Counting people without people models or tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 24–26 June 2008; pp. 1–7.
21. Zhang, J.; Tan, B.; Sha, F.; He, L. Predicting pedestrian counts in crowded scenes with rich and high-dimensional features. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 1037–1046. [CrossRef]
22. Ma, Z.; Chan, A.B. Counting people crossing a line using integer programming and local features. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *26*, 1955–1969. [CrossRef]
23. Zheng, H.; Lin, Z.; Cen, J.; Wu, Z.; Zhao, Y. Cross-line pedestrian counting based on spatially-consistent two-stage local crowd density estimation and accumulation. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 787–799. [CrossRef]
24. Sheng, B.; Shen, C.; Lin, G.; Li, J.; Yang, W.; Sun, C. Crowd counting via weighted VLAD on a dense attribute feature map. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *28*, 1788–1797. [CrossRef]
25. Li, J.; Xue, Y.; Wang, W.; Ouyang, G. Cross-Level Parallel Network for Crowd Counting. *IEEE Trans. Ind. Inf.* **2020**, *16*, 566–576. [CrossRef]
26. Liu, L.; Qiu, Z.; Li, G.; Liu, S.; Ouyang, W.; Lin, L. Crowd counting with deep structured scale integration network. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 1774–1783.
27. Cheng, Z.Q.; Li, J.X.; Dai, Q.; Wu, X.; He, J.Y.; Hauptmann, A.G. Improving the learning of multi-column convolutional neural network for crowd counting. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1897–1906.
28. Qiu, Z.; Liu, L.; Li, G.; Wang, Q.; Xiao, N.; Lin, L. Crowd counting via multi-view scale aggregation networks. In Proceedings of the IEEE International Conference on Multimedia and Expo, Shanghai, China, 8–12 July 2019; pp. 1498–1503.
29. Kang, D.; Chan, A. Crowd Counting by Adaptively Fusing Predictions from an Image Pyramid. In Proceedings of the British Machine Vision Conference, Newcastle, UK, 2–6 September 2018.
30. Liu, L.; Wang, H.; Li, G.; Ouyang, W.; Lin, L. Crowd Counting Using Deep Recurrent Spatial-aware Network. In Proceedings of the 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 849–855.
31. Amirgholipour, S.; He, X.; Jia, W.; Wang, D.; Zeibots, M. A-CCNN: Adaptive CCNN for Density Estimation and Crowd Counting. In Proceedings of the IEEE International Conference on Image Processing, Athens, Greece, 7–10 October 2018; pp. 948–952.
32. Ding, X.; Lin, Z.; He, F.; Wang, Y.; Huang, Y. A Deeply-Recursive Convolutional Network For Crowd Counting. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Seoul, South Korea, 22–27 April 2018; pp. 1942–1946.
33. Zhang, L.; Shi, M.; Chen, Q. Crowd counting via scale-adaptive convolutional neural network. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1113–1121.
34. Cao, X.; Wang, Z.; Zhao, Y.; Su, F. Scale aggregation network for accurate and efficient crowd counting. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 734–750.
35. Luo, A.; Yang, F.; Li, X.; Nie, D.; Jiao, Z.; Zhou, S.; Cheng, H. Hybrid Graph Neural Networks for Crowd Counting. *arXiv* **2020**, arXiv:2002.00092.
36. Shi, M.; Yang, Z.; Xu, C.; Chen, Q. Revisiting perspective information for efficient crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 7279–7288.
37. Yan, Z.; Yuan, Y.; Zuo, W.; Tan, X.; Wang, Y.; Wen, S.; Ding, E. Perspective-guided convolution networks for crowd counting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 952–961.
38. Sam, D.B.; Surya, S.; Babu, R.V. Switching convolutional neural network for crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 24–30 June 2017; pp. 4031–4039.
39. Sindagi, V.A.; Patel, V.M. Generating high-quality crowd density maps using contextual pyramid cnns. In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1861–1870.
40. Gao, J.; Wang, Q.; Li, X. PCC Net: Perspective Crowd Counting via Spatial Convolutional Network. *IEEE Trans. Circuits Syst. Video Technol.* **2019**. [CrossRef]
41. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
42. Wang, W.; Zhou, T.; Yu, F.; Dai, J.; Konukoglu, E.; Van Gool, L. Exploring cross-image pixel contrast for semantic segmentation. *arXiv* **2021**, arXiv:2101.11939.
43. Li, X.; Zhou, T.; Li, J.; Zhou, Y.; Zhang, Z. Group-Wise Semantic Mining for Weakly Supervised Semantic Segmentation. *arXiv* **2020**, arXiv:2012.05007.
44. Zhou, T.; Qi, S.; Wang, W.; Shen, J.; Zhu, S.C. Cascaded parsing of human-object interaction recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**. [CrossRef]
45. Zhou, T.; Li, J.; Wang, S.; Tao, R.; Shen, J. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 8326–8338. [CrossRef]

46. Sindagi, V.A.; Patel, V.M. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Trans. Image Process.* **2019**, *29*, 323–335. [CrossRef]
47. Zhang, A.; Shen, J.; Xiao, Z.; Zhu, F.; Zhen, X.; Cao, X.; Shao, L. Relational attention network for crowd counting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 6788–6797.
48. Zhang, A.; Yue, L.; Shen, J.; Zhu, F.; Zhen, X.; Cao, X.; Shao, L. Attentional neural fields for crowd counting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 5714–5723.
49. Guo, D.; Li, K.; Zha, Z.J.; Wang, M. Dadnet: Dilated-attention-deformable convnet for crowd counting. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1823–1832.
50. Zhao, M.; Zhang, J.; Zhang, C.; Zhang, W. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 12736–12745.
51. Jiang, S.; Lu, X.; Lei, Y.; Liu, L. Mask-aware networks for crowd counting. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *30*, 3119–3129 [CrossRef]
52. Wu, X.; Zheng, Y.; Ye, H.; Hu, W.; Yang, J.; He, L. Adaptive scenario discovery for crowd counting. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2019; pp. 2382–2386.
53. Sajid, U.; Wang, G. Plug-and-Play Rescaling Based Crowd Counting in Static Images. *arXiv* **2020**, arXiv:2001.01786.
54. Wang, Q.; Gao, J.; Lin, W.; Yuan, Y. Learning from synthetic data for crowd counting in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 8198–8207.
55. Li, W.; Yongbo, L.; Xiangyang, X. CODA: Counting Objects via Scale-Aware Adversarial Density Adaption. In Proceedings of the IEEE International Conference on Multimedia and Expo, Shanghai, China, 8–12 July 2019; pp. 193–198.
56. Lempitsky, V.; Zisserman, A. Learning to count objects in images. *Adv. Neural Inf. Process. Syst.* **2010**, *23*, 1324–1332.
57. Wan, J.; Chan, A. Adaptive density map generation for crowd counting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 1130–1139.
58. Sindagi, V.A.; Yasarla, R.; Patel, V.M. Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 1221–1231.
59. Sindagi, V.A.; Patel, V.M. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 1002–1012.
60. Ma, Z.; Wei, X.; Hong, X.; Gong, Y. Bayesian loss for crowd count estimation with point supervision. In Proceedings of the International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 6142–6151.
61. Shen, Z.; Xu, Y.; Ni, B.; Wang, M.; Hu, J.; Yang, X. Crowd counting via adversarial cross-scale consistency pursuit. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5245–5254.
62. Zhou, Y.; Yang, J.; Li, H.; Cao, T.; Kung, S.Y. Adversarial learning for multiscale crowd counting under complex scenes. *IEEE Trans. Cybern.* **2020**. [CrossRef]
63. Cheng, Z.Q.; Li, J.X.; Dai, Q.; Wu, X.; Hauptmann, A.G. Learning spatial awareness to improve crowd counting. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 6152–6161.
64. Liu, J.; Gao, C.; Meng, D.; Hauptmann, A.G. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5197–5206.
65. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv* **2015**, arXiv:1506.01497.
66. Lian, D.; Li, J.; Zheng, J.; Luo, W.; Gao, S. Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 1821–1830.
67. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
68. He, G.; Ma, Z.; Huang, B.; Sheng, B.; Yuan, Y. Dynamic Region Division for Adaptive Learning Pedestrian Counting. In Proceedings of the IEEE International Conference on Multimedia and Expo, Shanghai, China, 8–12 July 2019; pp. 1120–1125.
69. Liu, C.; Weng, X.; Mu, Y. Recurrent Attentive Zooming for Joint Crowd Counting and Precise Localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 1217–1226.
70. Shi, Z.; Mettes, P.; Snoek, C.G. Counting with focus for free. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–3 November 2019; pp. 4200–4209.
71. Xu, C.; Qiu, K.; Fu, J.; Bai, S.; Xu, Y.; Bai, X. Learn to Scale: Generating Multipolar Normalized Density Map for Crowd Counting. *arXiv* **2019**, arXiv:1907.12428.
72. Xu, C.; Liang, D.; Xu, Y.; Bai, S.; Zhan, W.; Bai, X.; Tomizuka, M. Autoscale: Learning to scale for crowd counting. *arXiv* **2019**, arXiv:1912.09632.
73. Khan, S.D.; Ullah, H.; Uzair, M.; Ullah, M.; Ullah, R.; Cheikh, F.A. Disam: Density Independent and Scale Aware Model for Crowd Counting and Localization. In Proceedings of the IEEE International Conference on Image Processing, Taipei, Taiwan, 22–25 September 2019; pp. 4474–4478.

74. Wang, S.; Lu, Y.; Zhou, T.; Di, H.; Lu, L.; Zhang, L. SCLNet: Spatial context learning network for congested crowd counting. *Neurocomputing* **2020**, *404*, 227–239. [CrossRef]
75. Xia, G.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3974–3983.
76. Bazi, Y.; Al-Sharari, H.; Melgani, F. An automatic method for counting olive trees in very high spatial remote sensing images. In Proceedings of the IEEE International Geoscience and Remote Sensing Symposium, Cape Town, South Africa, 12–17 July 2009; Volume 2, pp. 125–128.
77. Santoro, F.; Tarantino, E.; Figorito, B.; Gualano, S.; D'Onghia, A.M. A tree counting algorithm for precision agriculture tasks. *Int. J. Digital Earth* **2013**, *6*, 94–102. [CrossRef]
78. Xue, Y.; Wang, T.; Skidmore, A.K. Automatic counting of large mammals from very high resolution panchromatic satellite imagery. *Remote Sens.* **2017**, *9*, 878. [CrossRef]
79. Salamí, E.; Gallardo, A.; Skorobogatov, G.; Barrado, C. On-the-fly olive tree counting using a UAS and cloud services. *Remote Sens.* **2019**, *11*, 316. [CrossRef]
80. Mubin, N.A.; Nadarajoo, E.; Shafri, H.Z.M.; Hamedianfar, A. Young and mature oil palm tree detection and counting using convolutional neural network deep learning method. *Int. J. Remote Sens.* **2019**, *40*, 7500–7515. [CrossRef]
81. Shao, W.; Kawakami, R.; Yoshihashi, R.; You, S.; Kawase, H.; Naemura, T. Cattle detection and counting in UAV images based on convolutional neural networks. *Int. J. Remote Sens.* **2020**, *41*, 31–52. [CrossRef]
82. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 24–30 June 2017; pp. 7263–7271.
83. Mundhenk, T.N.; Konjevod, G.; Sakla, W.A.; Boakye, K. A large contextual dataset for classification, detection and counting of cars with deep learning. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 785–800.
84. Fahim, M.; Baker, T.; Khattak, A.M.; Shah, B.; Aleem, S.; Chow, F. Context mining of sedentary behaviour for promoting self-awareness using a smartphone. *Sensors* **2018**, *18*, 874. [CrossRef] [PubMed]
85. Hsieh, M.R.; Lin, Y.L.; Hsu, W.H. Drone-based object counting by spatially regularized regional proposal network. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4145–4153.
86. Gao, G.; Liu, Q.; Wang, Y. Counting From Sky: A Large-Scale Data Set for Remote Sensing Object Counting and a Benchmark Method. *IEEE Trans. Geosci. Remote Sens.* **2020**. [CrossRef]
87. Wan, J.; Wang, Q.; Chan, A.B. Kernel-based Density Map Generation for Dense Object Counting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, 1. [CrossRef] [PubMed]
88. Zhang, P.; Liu, W.; Lei, Y.; Lu, H.; Yang, X. Cascaded Context Pyramid for Full-Resolution 3D Semantic Scene Completion. *arXiv* **2019**, arXiv:1908.00382.
89. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
90. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 24–30 June 2017; pp. 2881–2890.
91. Idrees, H.; Saleemi, I.; Seibert, C.; Shah, M. Multi-source multi-scale counting in extremely dense crowd images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2547–2554.
92. Gao, J.; Lin, W.; Zhao, B.; Wang, D.; Gao, C.; Wen, J. C^3 Framework: An Open-source PyTorch Code for Crowd Counting. *arXiv* **2019**, arXiv:1907.02724.
93. Wang, Q.; Gao, J.; Lin, W.; Yuan, Y. Pixel-Wise Crowd Understanding via Synthetic Data. *Int. J. Comput. Vision* **2021**, *129*, 225–245. [CrossRef]
94. Ranjan, V.; Le, H.; Hoai, M. Iterative crowd counting. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 270–285.
95. Jiang, X.; Xiao, Z.; Zhang, B.; Zhen, X.; Cao, X.; Doermann, D.; Shao, L. Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 6133–6142.
96. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 24–30 June 2017; pp. 2261–2269.
97. Sindagi, V.A.; Patel, V.M. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In Proceedings of the 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; pp. 1–6.
98. Gao, J.; Wang, Q.; Yuan, Y. SCAR: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing* **2019**, *363*, 1–8. [CrossRef]
99. Zhu, L.; Zhao, Z.; Lu, C.; Lin, Y.; Peng, Y.; Yao, T. Dual path multi-scale fusion networks with attention for crowd counting. *arXiv* **2019**, arXiv:1902.01115.