sensors

MDPI

Article

Counting Crowds with Perspective Distortion Correction via Adaptive Learning

Yixuan Sun, Jian Jin *, Xingjiao Wu, Tianlong Ma and Jing Yang

School of Computer Science and Technology, East China Normal University, Shanghai 200062, China; 10175102160@stu.ecnu.edu.cn (Y.S.); 52184506007@stu.ecnu.edu.cn (X.W.); tlma@cs.ecnu.edu.cn (T.M.); jyang@cs.ecnu.edu.cn (J.Y.)

* Correspondence: jjin@cs.ecnu.edu.cn

Received: 8 June 2020; Accepted: 3 July 2020; Published: 6 July 2020

Abstract: The goal of crowd counting is to estimate the number of people in the image. Presently, use regression to count people number became a mainstream method. It is worth noting that, with the development of convolutional neural networks (CNN), methods that are based on CNN have become a research hotspot. It is a more interesting topic that how to locate the site of the person in the image than simply predicting the number of people in the image. The perspective transformation present is still a challenge, because perspective distortion will cause differences in the size of the crowd in the image. To devote perspective distortion and locate the site of the person more accuracy, we design a novel framework named Adaptive Learning Network (CAL). We use the VGG as the backbone. After each pooling layer is output, we collect the 1/2, 1/4, 1/8, and 1/16 features of the original image and combine them with the weights learned by an adaptive learning branch. The object of our adaptive learning branch is each image in the datasets. By combining the output features of different sizes of each image, the challenge of drastic changes in the size of the image crowd due to perspective transformation is reduced. We conducted experiments on four population counting data sets (i.e., ShanghaiTech Part A, ShanghaiTech Part B, UCF_CC_50 and UCF-QNRF), and the results show that our model has a good performance.

Keywords: crowd counting; localization; adaptive learning; convolutional neural network

1. Introduction

The goal of the crowd counting task is to count the number of people in an image. The crowd counting task plays an important role in the production, life, disaster management, security monitoring, and public space design [1–3]. With the improvement of people's safety awareness, crowd counting has been paid increasing attention. Recently the crowd counting task has utilized convolutional neural network (CNN) to address the scale variation issue and has achieved good improvements in crowd density estimation [4,5].

However, the perspective distortion of the image is still an important challenge for crowd counting, more specifically, the model is not particularly accurate in predicting avatars with large differences in size in the same image. Hence, how to better handle objects of different sizes is a key to improve the crowd counting model. Recently, the demand for crowd counting is no longer simply counting the total number of people in the image, but also want to locate a specific personal location, so that accurate counting can be performed more accurately. Most of the current work uses Visual Geometry Group (VGG) [6] as a backbone. Subsequently, separately extract different sizes of features after each max pooling operation, and decoded these features. We will obtain the features that size is 1/2, 1/4, 1/8, and 1/16 of the original image size after each max pooling.

The current method is to simply superimpose features of different sizes without considering the combination of different sizes brought by different image inputs and different scene inputs. The degree of perspective transformation in each image is not the same, that is, if our branch information is merged according to the same pattern, then the learned knowledge cannot cover all samples. On this basis, we envisage using a dynamic mechanism to combine branch information according to different image features to achieve the goal of dynamic evolution. Inspired by the adaptive scenario discovery framework (ASD) [7] model, we also propose a dynamic learning branch combination method. Different from ASD, our model is not only concerned with simple counting tasks, but we also add the positioning of specific objects to the model. At the same time, ASD distinguishes between sparse and dense scenes, and our model is to explore the degree of perspective change in the image. In this paper, we propose an adaptive learning framework (CAL) with perspective distortion correction for crowd counting and localization. We employ several of VGG-16 convolution layers for crowd feature extraction before the multiple receptive fields instead of utilizing them directly. Additionally, for exploring the degree of perspective change in the image, four parallel pathways with the counting and localization network named main, scale, middle, and lowest are proposed. The four pathways are designed for the people with different scale, respectively. Besides, we also designed a branch to learn the degree of perspective change. Afterwards, combine the perspective into the output branch of the model.

Our contributions are listed, as follows.

- We propose a novel adaptive framework with perspective distortion correction for crowd counting and localization. Different from the former proposed multiple columns frameworks, we use a branch to dynamically characterize the degree of perspective change of the images. We further verify the effect of our CAL network and compare with the No-CAL methods in order to explain the improvement of our architecture.
- We design a novel size characterization branch to realize both the crowd counting and the localization task.
- We use VGG [6] for the feature extraction structure and the network constructed by four branches (including the main path), which select output features of different sizes. The perspective change in the image is considered to be a linear combination of our four branches and discrete weights, while the adaptation branch aims to portray a continuous perspective change trend and make corresponding corrections.
- We apply our framework to four congested multi-scene crowd counting datasets (i.e., ShanghaiTech Part A, ShanghaiTech Part B, UCF_CC_50, and UCF-QNRF) and prove that our method outperforms the state-of-the-art methods.

In the remaining part of the paper, we discuss related works of crowd counting and localization in Section 2, describe the backbone, the CAL network architecture and training process in Section 3, verify the proposed framework in both qualitative and quantitative extent in Section 4 and finally conclude our work in Section 5.

2. Related Work

We present a survey about the recent works of crowd counting and localization in three parts: (1) traditional crowd counting methods; (2) CNNs for counting; and, (3) CNNs for localization. The earliest researches mostly based on detection frameworks, which were used to detect people and to count the number of pedestrians. However, the occlusion, the extremely dense crowds, and high background clutter limited its development, even though some improvement,such as parts-based or shape-based detectors, were proposed. To devote these issues, some researcher proposes the regression-based methods (mapping the features extracted from local images and their counts) took the place of detection-based methods. Since the CNN was proposed, it has been successfully used in various computer vision tasks, which inspired the use of CNN based methods in crowd counting tasks. Though the approaches of crowd counting scenes gained satisfying performance, in several scenes some more detailed information, such as the distribution and the location of the objects, were needed. As a result, researchers improved the CNN frameworks and

developed a series of CNN based localization models. Nowadays, the localization tasks with CNN based framework are still the hotspots for researchers.

2.1. Traditional Crowd Counting Methods

The early detection-based methods [8–11] rely on the detection style framework that used the slide window to detect people in images. These methods estimating the number in the low-density crowd scenes by detecting the whole body of the pedestrians. However, in high-density situations, heads are ususally the only visible part due to the occlusions. As a further development, the detectors of some body parts (such as head or head-shoulder [12]) detection methods were proposed. On the other hand, regression-based approaches [13–18] regressed the density map of crowds and the integration of which is the crowd counting result. These earlier methods [15–18] mapped global image features or combined local patch features to do counting, which produces approximately counts. When comparing these two methods, regression-based approaches perform well in high-density situations. Additionally, the detection-based methods can usually handle the counting and localization problems simultaneously.

2.2. CNNs for Crowd Counting

Recently, CNN based approaches [19–22] have shown their advantages in learning the crowd image feature mapping and the people/head detection for both crowd counting [23–28] and localization [26–30]. The Multi-column Convolutional Neural Network (MCNN) method is evaluated in [19] which contains three columns of different filters to extract feature of heads in different scales. Sam et al. [21] proposed the Switching-CNN and trained each of three columns with a subset of the patches, while a density selector is designed for extracting the structural and functional differences. Li et al. [31] introduce the CSRNet as an approach to concentrate on encoding the deeper features in congested scenes. Besides, Idrees [32] introduced a deep CNN with composition loss method to satisfy counting, density map estimation, and localization. To handle the small/tiny objects that often appear in crowd counting scenes, Basalamah et al. [27] used the scale-aware object proposal generated by perspective information which handled scale variations and makes the model (SD-CNN) able to detect human heads in both low density and high-density crowd images. Onoro et al. [33] using the Hydra-CNN fuses the multi-scale information provided by heads to handle the crowd counting problems with significant variations in the scene. Additionally, Reference [26] introduced the depth information by leveraging RGB-D data to improve the performance of small object detection. In some occasions, like the wild scenes or the congress scenes, the annotation can be costly. As a result, recent researches [34–36] aimed at dealing with the lack of labeled data by self-supervised learning [34,36] or the unsupervised learning [35]. In [34], Wang et al. used the GCC dataset to fine-tune a pre-trained crowd counter and proposed a crowd counting method via domain adaptation, which freed the researchers from data annotations. For the unsupervised ways, Reference [35] presented an unsupervised learning method using Grid Winner-Take-all (GWTA) Counting CNN to learn features from unlabeled crowd images.

2.3. CNNs for Localization

As the regression-based crowd counting methods are widely used in the counting scenes, the most direct idea is to handle the localization task by sharpening crowd density maps. However, the low accuracy of density map that was argued in most of the prior studies [28] is still an unignorable drawback. An early anomaly detection and localization method [30] introduced normalcy models jointly show the appearance and dynamics of complex congested scenes in which MDTs are learned at multiple scales to handle the problems of empirical evaluation of anomaly detectors on crowded scenes. The further anomaly detection research [37] proposed an unsupervised approach for crowd scene anomaly detection and localization while using the social network model, which outperformed the former ones. To handle the localization and detection task in the noisy foreground, Chen et al. [38] extracted noisy foreground using the person detector and foreground segmentation. Chen et al. [38]

also introduced the new framework of EGR and introduced a new metric for both errors in localization and counting. Nowadays, most researches [26–28,33] using both the density map and neural network detector for the localization task. Idrees et al. [32] introduced the composition loss to do the counting, density map estimation and localization in congested scenes simultaneously. Reference [27] devises a scale-aware head detector and using the response map to optimize the detector to make the test results more consistent with population distribution. Instead of the scale-awareness, Reference [26] approached the depth information by designed a depth-aware anchor to initialize the anchor and estimated the bounding box sizes of all heads that were utilized as the ground truth to train the RDNet. Additionally, to satisfy the demanding of large-scale RGB-D dataset, Reference [26] also introduced an RGB-D dataset contains 2193 images and 144,512 headcounts named ShanghaiTechRGBD. And Liu et al. [28] proposed the recurrent attentive zooming network to zoom the detected ambiguous image region into high resolution and using the RAZ Net for re-inspection.

The Fully Convolutional Network (FCN) is proposed for the pixel-level classification of images. Matan et al. [39] extended the classic LeNet [40] to recognize strings of digits. Additionally, in the segmentation of *C. elegans* tissues scene, Ning et al. [41] used the fully convolutional inference to design a convent. In recent years, multi-layered nets have also exploited the fully convolutional computation (such as sliding window detection [42] and image restoration [43]). Moreover, He et al. [44] generated a localized and fixed-length feature with proposals and spatial pyramid pooling. However, the drawback of this hybrid model is that it cannot be learned end-to-end. Taking these case studies into account, Long et al. [45] proposed a fully convolutional network trained end to end and pixel to pixels. The research is known as the first work trained the FCNs end to end for pixel wised segmentation and used the supervised pre-training. Additionally, [46] designed a U-net model, which used a fully-convolutional neural network as the core of the model. The model required the full per-pixel instance segmentation labels for training. Extending from [45], Issam et al. [47] designed a novel loss function for the FCN model, called localization-based counting loss (LC), and named this new FCN detection-based model with LC function the LC-FCN. Nowadays, Sam et al. [48] proposed the LSC-CNN model while using the multi-column architecture to fulfill the reliably head detection, automatic head size estimation and high precision crowd counting features simultaneously. As a multi-column FCN-liked architecture, the model performs ideally across sparse to dense crowds and only requires the point annotation.

As we said at the beginning of this section, the crowd counting technology has been well developed in recent years. At the initial stage of crowd counting, researchers used sliding windows and regression methods to obtain the number of people in the image. However, with the deepening of research problems and the increase in the number of people in the image, the traditional crowd counting technology has been unable to meet those issues. As such, researchers began to explore CNN-based methods. Presently, the CNN-based method can achieve good results, but the traditional CNN-based method can only get the prediction number, and cannot calibrate the localization of the crowd in the image, so the localization method is proposed. At present, most of the localization task base on FCN structure and, at the same time, the image features of different sizes are stitched and decoded. However, this method is not ideal for scenes with a severe perspective change. Therefore, based on FCN, we propose an adaptive learning framework with perspective distortion correction for crowd counting and localization. We achieve an end-to-end regression method using CNNs, which takes the entire image as input and obtain greater accuracy when compared to previous approaches.

3. Framework

We propose a novel model that contains three parts: the backbone, the pathways, and the adaptive branch. We design a novel framework named Adaptive Learning Network (CAL) and the architecture is shown in Figure 1. In the following parts, we introduce the structure and implementation in detail.

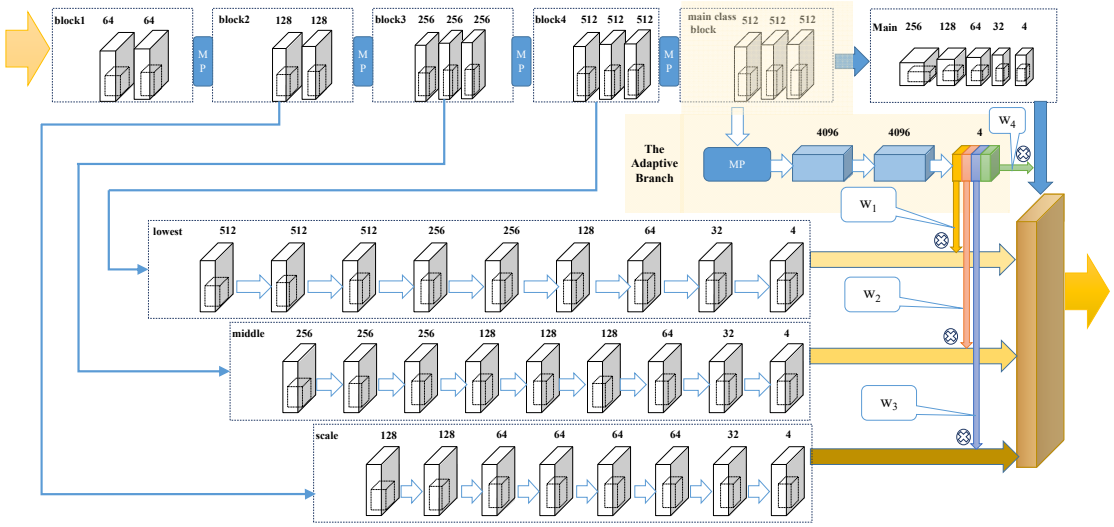

Figure 1. The architecture and weight of CAL.

3.1. Backbone

Presently, the mainstream method for extracting features from the crowd counting task is to use the VGG [6] network as a backbone. The backbone network utilizing can be separated into two ways: starting from scratch to designing a new network (e.g., [19]) or migrating a pre-trained subnet from an existing network (e.g., [31,47,49]). Between these two categories, the second way have more advantages in both time-saving and efficiency. Our network design also follows this principle. We first designed a feature extraction structure with VGG16 as the backbone. However, we duplicated and fine-tuned some blocks to adapt the feature extraction task with multiple resolutions. More specifically, our backbone removes the fully connected layer of VGG16, as shown in Table 1. Besides, our VGG model first uses the ImageNet dataset [50] for pre-training.

Table 1. The struct of backbone.

Input (224 × 224 RGB Image)				
	Channels Number	**Kernel_Size**	**Stride**	**Size**
Conv1_1	64	3	1	224 × 224
Conv1_2	64	3	1	224 × 224
Max Pooling	-	2	2	112 × 112
Conv2_1	128	3	1	112 × 112
Conv2_2	128	3	1	112 × 112
Max Pooling	-	2	2	56 × 56
Conv3_1	256	3	1	56 × 56
Conv3_2	256	3	1	56 × 56
Conv3_3	256	3	1	56 × 56
Max Pooling	-	2	2	28 × 28
Conv4_1	512	3	1	28 × 28
Conv4_2	512	3	1	28 × 28
Conv4_3	512	3	1	28 × 28
Max Pooling	-	2	2	14 × 14
Conv5_1	512	3	1	14 × 14
Conv5_2	512	3	1	14 × 14
Conv5_3	512	3	1	14 × 14

3.2. The Pathways

Following the general principles of localization network design, our network design also uses the FCN structure. Similar to many networks, we also set up four different branches to decode 1/2, 1/4, 1/8, and 1/16 of the original image size, four parallel pathways with the counting, and localization network named main, scale, middle, and lowest are proposed. Table 2 shows the pathways configuration.

Table 2. The config of the pathways.

Orgin Image Size: 224 × 224				
Input Size	**14 × 14**	**112 × 112**	**56 × 56**	**28 × 28**
	Main	**Scale**	**Middle**	**Lowest**
The config of the Pathways	conv(3, 512) relu conv(3, 512) relu conv(3, 512) relu conv(3, 256) relu conv(3, 128) relu conv(3, 64) relu conv(3, 32) relu conv(3, 4) relu	conv(3, 128) relu conv(3, 128) relu conv(3, 64) relu conv(3, 64) relu conv(3, 64) relu conv(3, 32) relu conv(3, 4) relu	conv(3, 256) relu conv(3, 256) relu conv(3, 256) relu conv(3, 128) relu conv(3, 128) relu conv(3, 128) relu conv(3, 64) relu conv(3, 32) relu conv(3, 4) relu	conv(3, 512) relu conv(3, 512) relu conv(3, 512) relu conv(3, 256) relu conv(3, 256) relu conv(3, 128) relu conv(3, 64) relu conv(3, 32) relu conv(3, 4) relu

3.3. The Adaptive Branch

However, unlike other networks connected directly in series, we propose using different weights to combine the output of each branch. We constructed a self-learning classification branch, named the adaptive branch. The input of this branch is the feature parameter extracted by VGG16. The branch structure is as follows: conv (3, 512, 1)–conv (3, 512, 1)–conv (3, 512, 1)–pool(2)–FC (25088, 4096)–RELU–FC(4096, 4096)–RELU–FC (4096, 4). Where 'conv' represents a convolutional layer, and 'pool' represents a max-pooling layer, 'FC' represents the fully connected layer, 'RELU' represents Rectified Linear Unit. The numbers in the parentheses are respectively kernel size, the number of channels and dilation rate. Finally, we can obtain four channels (CH) and then normalize each channel separately after summing. The weight coefficient is obtained, as shown in Equation (1).

We provide different weights through the adaptive branch, in order to determine the proportion of the original image that scales different sizes in the result. Using different weight values determines which size image details we will pay more attention to. If we pay attention to more than 1/16 of the image, then it is bound to be ignored for those particularly small heads. Conversely, if we pay attention to 1/2 of the image, then we are bound to pay more attention to those larger heads. Through dynamic learning, we can allocate the proportion of images with different degrees of attention according to the specific scene. That helps eliminate the effects of perspective changes.

$$w_i = \frac{|\sum CH_i|}{\sum_{i=1}^{4} |\sum CH_i| + 10^{-9}} \ (i = 1, 2, 3, 4) \tag{1}$$

3.4. Implementation Details

Our perspective distortion correction model is implemented using PyTorch [51]. To train the model, we first initialize the batch size as typically four, while the momentum parameter is set as 0.9. We then set the learning rate of $1e^{-3}$ for all the datasets as initial, and use SGD [52] for training. For the training of UCF_CC_50, we especially use the five-fold cross-validation to make full use of the datasets to test the effectiveness of the algorithm.

3.4.1. Loss Function

Following the design of loss function in [4,5], we proposed the loss function as Equation (2). In which N, X_i, and θ represent for the batch size, the ith input image, and a set of trainable parameters, respectively. Besides, γ_i is the ground truth of X_i. Additionally, $Y(X_i;\theta)$ stands for the estimated density map generated by our proposed model with parameters θ. $L(\theta)$ denotes the loss function between the estimated results and the ground truth.

$$L(\theta) = \frac{1}{2N}\sum_{n=1}^{N}\left(\left\|\gamma(X_i;\theta) - \gamma_i^{GT}\right\|_2^2\right) \tag{2}$$

3.4.2. Density Map Generation

CNN needs to process continuous data for crowd counting tasks. As a result, we have to convert the discrete point annotated data (including the annotation of ground truth and the result of prediction) into the density map. The conversion is pixel level and the idea is to convert the point annotation information into images that probably contain density information. The details of the operation are shown in Algorithm 1.

Algorithm 1: Ground-truth generation

Input: I : Image matrix, $label$: label list
Output: DM : density map matrix
create matrix DM, which width and height are the same as the input image: I;
for $i = 1; i \leq length(label)$ **do**
    Find the three nearest neighbors' distance: l_1, l_2, l_3;
    Calculate Gaussian smoothing parameters: $\sigma_i = (l_1 + l_2 + l_3)/3 * \beta$;
    Calculate the density of $DM(i)$.
return DM;

4. Experiments

In this section, we introduce three popular crowd counting datasets that are frequently used in crowd counting and localization tasks. Besides, several ways to evaluate the performance of the architectures are introduced. Afterwards, we compare the previous experimental results and evaluate our method on these datasets.

4.1. Evaluation Metrics

Several ways are used to evaluate both the person detection and counting performance. For the counting evaluation, the commonly used mean absolute error (MAE) and mean square error (MSE) is used by us to measure the deviation of the prediction and the ground truth. The MAE and the MSE are defined as:

$$MAE = \frac{1}{T}\sum_{t=1}^{T}|\mu_t - G_t| \tag{3}$$

$$MSE = \frac{1}{T}\sum_{t=1}^{T}(\mu_t - G_t)^2 \tag{4}$$

where T is the sum amount of testing frames. While μ_t and G_t are the frame t prediction count and the ground-truth count of pedestrians, respectively.

4.2. Datasets

Currently, various of public datasets for crowd counting task is available, such as MALL [16], UCSD [53], ShanghaiTech [19], UCF_CC_50 [17], UCF-QNRF [32], etc. The comparison of the images in the listed datasets is shown in Figure 2. In our experiment, we evaluate the proposed model on three crowd counting datasets, including ShanghaiTech [19], UCF_CC_50 [17], and UCF-QNRF [32]. In the latter parts, we present the chosen datasets and explain why these datasets are chosen.

ShanghaiTech. Shanghai Tech [19] is one of the largest large-scale datasets in recent years which consists of total 1198 crowd images with 330,165 annotations. The dataset is divided into two sets, named Part A (SHT A) and Part B (SHT B), respectively. Part A is composed of images randomly selected from the Internet, in which the density fluctuates between 33 and 3139 people per image and with an average count of 501.4. In contrast, images in Part B are taken from a busy street of Shanghai and the crowd distribution of which is less diverse and sparser (123.6 in average).

UCF_CC_50. UCF_CC_50 [17] is the first challenging dataset on multiple counts created from Web images. The dataset contains various densities and different perspective distortions for multiple scenes. Being a small set of 50 images with crowd counts ranging in 50 to 4543, the dataset poses a serious problem for deep neural networks.

UCF-QNRF. UCF-QNRF [32] is collected from Web Search, Flickr, and Hajj footage, which was first introduced by Idrees et al. in [32]. The dataset is consist of a 1201 images train set and a 334 images test set with 1.25 million annotations in total and the density of images varying from 49 people per image to 12,865.

Figure 2. Sample images from various datasets. In order from left to right, each column is in turn UCSD [53], Mall [16], Shanghai Tech PartA [19], Shanghai Tech Part B [19], UCF_CC_50 [17], UCF-QNRF [32]. It is obvious that in UCSD and Mall dataset, the images providing no variation in perspective across images.

Extracting from these three datasets, the ideal dataset to examine the performance can be concluded as the following list:

- **Challenging images** Some challenging images are necessary to evaluate the performance of the model in extreme conditions. As the development of the crowd counting methods, most of them perform stably in the sparse scenes. As a result, our model focus on improving the performance in congestion crowds and achieve localization tasks. For the crowd counting and localization task,

images of some exceeding congestion crowds are the ideal material to evaluate the robustness and the accuracy of our model.

- **Proper density distribution** The distribution of the images can directly affect the performance of the model in the scenes with different levels of congestion. The proper amount of sparse, middle and congested images can improve training accuracy and make verification more effective.
- **Multiple scenes** The dataset contains multiple scenes, such as the street view, the market view, the live show view, etc., can improve the robustness of our model. The multiple scenes is not only the images take from a different location but also the different condition of weather (such as rainy and foggy), light intensity etc., which can affect the performance of our model, especially in the localization task.

In conclusion, the chosen datasets can well meet these issues while the Mall [16] and the UCSD [53] are insufficient in some respects. This explains why we exclude these two datasets.

4.3. Results and Discussion

ShanghaiTech. Following the introduction of ShanghaiTech dataset above, we evaluated the proposed framework with several state-of-the-art methods, including the localization method utilizing the adaptive fusion scheme named RAZNet [28], the LSC-CNN [48] with different receptive fields and ASD [49] introducing the adaptive scenario discovery framework. Table 3 summarizes the MAE and MSE of the former approaches and ours in two parts of ShanghaiTech. On Part A of ShanghaiTech, we achieve an impressing improvement of 2.1 of absolute MAE value over ASD [49] and 1.6 of MAE over the state-of-the-art RAZNet [28]. When compared with the state-of-the-art (LSC-CNN [48]) on Part B, our CAL network also achieved the best MAE of 8.1 and MSE of 11.9. As the output of our crowd counting and localization model, Figures 3 and 4 show the localization performance of some images from part A and Part B, respectively.

Table 3. The comparison among the state-of-the-arts and our approach in ShanghaiTech (Part A & Part B). The best result is in bold.

	Methods	Part A		Part B	
		MAE	**MSE**	**MAE**	**MSE**
Counting	MCNN [19]	110.2	173.2	26.4	41.3
	CMTL [54]	101.3	152.4	20	31.1
	TDF-CNN [55]	97.5	145.1	20.7	32.8
	Switching CNN [21]	90.4	135	21.6	33.4
	SaCNN [56]	86.8	139.2	16.2	25.8
	MSCNN [57]	83.8	127.4	17.7	30.2
	ACSCP [58]	75.7	102.7	17.2	27.4
	CP-CNN [59]	73.6	106.4	20.1	30.1
	D-ConvNet-v1 [25]	73.5	112.3	18.7	26
	DRSAN [60]	69.3	**96.4**	11.1	18.2
	CSRNet [31]	68.2	115	10.6	16
	SANet [61]	67	104.5	8.4	13.6
	PACNN [62]	66.3	106.4	8.9	13.5
	ASD [49]	65.6	98	8.5	13.7
Localization	RAZNet [28]	65.1	106.7	8.4	14.1
	RDNet [26]	-	-	8.8	15.3
	LC-FCN8 [47]	-	-	13.14	-
	LSC-CNN [48]	66.4	117	8.1	12.7
	CAL	**63.5**	99.2	**8.1**	**11.9**

UCF_CC_50. As a challenging crowd counting dataset introduced above, we also evaluated the CAL in UCF_CC_50. The results are shown in Table 4 and the instance results are reported in Figure 5. The same as the results on ShanghaiTech, the proposed framework shows better results, and the

performance improves on the former state-of-the-art results by 14.2 for the MAE metric, which shows the less volatility of the model in high crowd density images.

Table 4. The comparison among the state-of-the-arts and our approach in UCF_CC_50. The best result is in bold.

	Methods	MAE	MSE
Counting	Idrees 2013 [17]	468.0	590.3
	Zhang 2015 [63]	467.0	498.5
	MCNN [19]	377.6	509.1
	MSCNN [57]	363.7	468.4
	TDF-CNN [55]	354.7	491.4
	CMTL [54]	322.8	397.9
	Switching CNN [21]	318.1	439.2
	SaCNN [56]	314.9	424.8
	CP-CNN [59]	298.8	320.9
	PACNN [62]	267.9	357.8
	CSRNet [31]	266.1	397.5
	SPN [64]	259.2	335.9
	SANet [61]	258.4	334.9
	HA-CCN [65]	256.2	348.4
Localization	LSC-CNN [48]	225.6	**302.7**
	CAL	**211.4**	306.7

Figure 3. Qualitative results on the ShanghaiTech Part A.

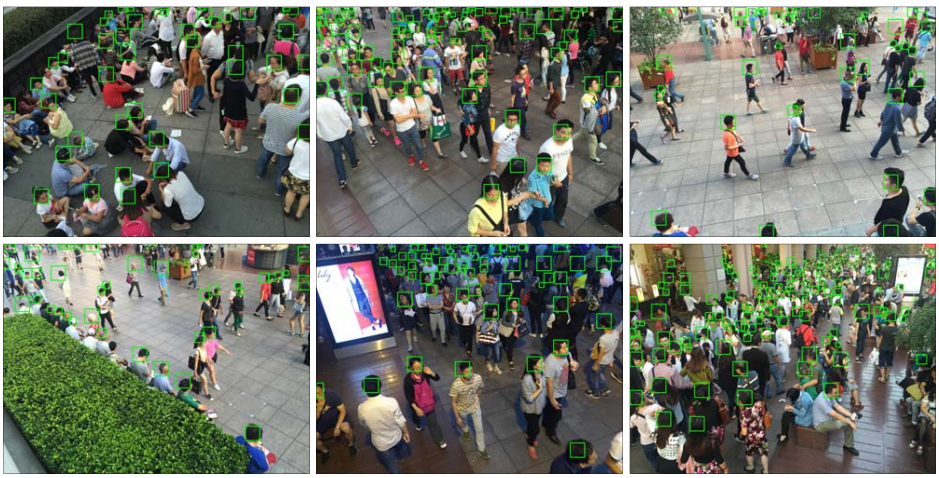

Figure 4. Qualitative results on the ShanghaiTech Part B.

Figure 5. Qualitative results on the UCF_CC_50.

UCF-QNRF. Follow the process and the idea of the other two datasets, we use MAE as the evaluation metric and keep the consistent detail for training. Table 5 compares our CAL model with state-of-the-art methods. It is obvious that our model outperforms all of the preceding models. Especially comparing with other localization methods, our network improves at least 10.2 in MAE. Additionally, we provide the performance in predicting the bounding box for localization in Figure 6, which illustrates the localization performance of some images in UCF-QNRF.

Figure 6. Qualitative results on the UCF-QNRF.

Table 5. The comparison among the state-of-the-arts and our approach in UCF-QNRF. The best result is in bold.

	Method	MAE	MSE
Counting	Idrees 2013 [17]	315	508
	MCNN [19]	277	426
	CMTL [54]	252	514
	Switching CNN [21]	228	445
	HA-CCN [65]	118.1	180.4
	TEDnet [66]	113	188
	RANet [67]	111	190
Localization	RAZNet [28]	116	195
	CL [32]	132	191
	LSC-CNN [48]	120.5	218.2
	CAL	**110.3**	**178.2**

4.4. Ablation Studies

In this part, we focus on two issues regarding the effectiveness of the structure of the multi-branch network and the performance of the adaptive branch. For this issue, we adjust our model and remove the adaptive branch to make it similar to some normal multi-column models (as shown in Figure 7). Moreover, we name the adjusted model the 'NO-CAL'. We removed the adaption branch is that we want to explore the improvement effect of the adaption branch on the model. We removed the adaption branch and create the NO-CAL structure in order to better compare the experimental results. To respond to the first issue, we compared our models (CAL & NO-CAL) with the previous multi-branch networks. Additionally, for the second issue, we make a comparison between our CAL model and the NO-CAL model. The results are shown in Tables 6 and 7.

Table 6. The comparison between other structure and our approach.

	ShanghaiTech Part A		**ShanghaiTech Part B**		**UCF_CC_50**		**UCF-QNRF**	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [19]	110.2	173.2	26.4	41.3	377.6	509.1	277	426
CMTL [54]	101.3	152.4	20	31.1	322.8	397.9	252	514
Switching CNN [21]	90.4	135	21.6	33.4	318.1	439.2	228	445
NO-CAL	70.8	119.5	14.2	18.9	258.9	369.0	163.7	200.9
CAL	63.5	99.2	8.1	11.9	211.4	306.7	110.3	178.2

Table 7. The comparison between the NO-CAL structure and our approach.

	ShanghaiTech Part A		**ShanghaiTech Part B**		**UCF_CC_50**		**UCF-QNRF**		**FPS**
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	ShanghaiTech Part B
CAL	63.5	99.2	8.1	11.9	211.4	306.7	110.3	178.2	12
NO-CAL	70.8	119.5	14.2	18.9	258.9	369.0	163.7	200.9	13

4.4.1. The Effectiveness of the Multi-Branch Structure

Table 6 shows the comparison of the former multi-branch structure with our design on ShanghaiTech, UCF_CC_50 and UCF-QNRF. It can be seen that our design outperforms the previous methods (MCNN [19], Switch-CNN [21], CMTL [54]). Additionally, the result shows that even the adaptive-branch-cutoff model (NO-CAL), the performance still at least improves on the former results (Part A: 70.8 vs. 90.4; Part B: 14.2 vs. 21.6; UCF_CC_50: 258.9 vs. 318.1; UCF-QNRF: 163.7 vs.228 on MAE). Moreover, the performance is much better than the NO-CAL structure (Part A: 7.3; Part B: 6.1; UCF_CC_50: 47.5; UCF-QNRF: 53.4 improvement on MAE). This is an illustration of the effectiveness of our multi-branch structure.

4.4.2. The Effectiveness of the Adaptive Branch

The previous method cannot handle the perspective distortion challenge properly, as discussed in Section 1. To deep-in validate if our proposed method is affected by the adaptive branch, we first conduct experiments with the CAL model and the same model cancels the adaptive branch (names NO-CAL). We validated both of the models on the three datasets, and the results are revealed in Table 7. It is shown that the CAL model visibly outperforms the NO-CAL, which is due to the good handling of the perspective distortion challenge. The experiment proves the effectiveness of the adaptive branch.

To compare the efficiency of our adaptive branch, we evaluate its time performance. Because the size of the image of the ShanghaiTech Part B is the fix, we use ShanghaiTech Part B as a benchmark to test the time efficiency of the model. The CAL achieves 12 FPS detection speed on an Nvidia TITAN XP GPU and the NO-CAL achieves 13 FPS detection speed on an Nvidia TITAN XP GPU during inference. It may take a little time to use the adaptive branch, but the time spent is in an acceptable range as compared with the improved accuracy.

As our ablation study shows, our design of the network structure is effective and well-performed among three chosen datasets.

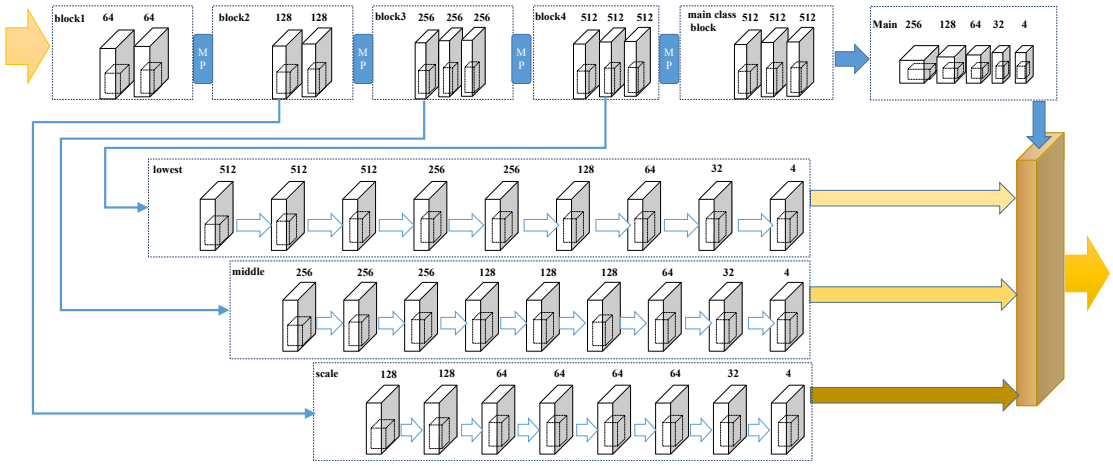

Figure 7. The architecture and weight of NO-CAL.

5. Conclusions

In this paper, we have presented a novel architecture for counting crowds with perspective distortion correction via adaptive learning. The focus of our method is to use a dynamic learning network to learn the dynamic combination relationship under different samples, and use this dynamic combination relationship to form different ratios for each image sample. Experimental comparisons with the state-of-the-art approaches (at most 15 methods) on ShanghaiTech, UCF_CC_50, and UCF-QNRF showed the effectiveness and efficiency of our proposed adaptive scenario discovery framework for the crowd counting task.

Author Contributions: Y.S. and X.W. contributed to the design and implementation of the research, analyzed the results, and wrote the manuscript. J.J. devised the project and wrote the manuscript. T.M. and J.Y. devised the main conceptual ideas, planned the experiments, and wrote the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Youth Program of National Natural Science Foundation of China No. 61907015, the Science and Technology Commission of Shanghai Municipality of China No. 18511103801, 18511103802, 18511106202.

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

References

1. Gao, G.; Gao, J.; Liu, Q.; Wang, Q.; Wang, Y. CNN-based Density Estimation and Crowd Counting: A Survey. *arXiv* **2020**, arXiv:2003.12783.
2. Kang, D.; Ma, Z.; Chan, A.B. Beyond Counting: Comparisons of Density Maps for Crowd Analysis Tasks Counting, Detection, and Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 1408–1422. [CrossRef]
3. Sindagi, V.A.; Patel, V.M. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognit. Lett.* **2018**, *107*, 3–16. [CrossRef]
4. 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]
5. 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]
6. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
7. Wu, X.; Zheng, Y.; Ye, H.; Hu, W.; Ma, T.; Yang, J.; He, L. Counting crowds with varying densities via adaptive scenario discovery framework. *Neurocomputing* **2020**, *397*, 127–138. [CrossRef]
8. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893.

9. Leibe, B.; Seemann, E.; Schiele, B. Pedestrian detection in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 878–885.
10. Tuzel, O.; Porikli, F.; Meer, P. Pedestrian detection via classification on riemannian manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *30*, 1713–1727. [CrossRef]
11. Enzweiler, M.; Gavrila, D.M. Monocular pedestrian detection: Survey and experiments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2008**, *31*, 2179–2195. [CrossRef]
12. Li, M.; Zhang, Z.; Huang, K.; Tan, T. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In Proceedings of the International Conference on Pattern Recognition (ICPR), Tampa, FL, USA, 8–11 December 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–4.
13. Chan, A.B.; Vasconcelos, N. Bayesian poisson regression for crowd counting. In Proceedings of the International Conference on Computer Vision (ICCV), Kyoto, Japan, 29 September–2 October 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 545–551.
14. Ryan, D.; Denman, S.; Fookes, C.; Sridharan, S. Crowd counting using multiple local features. In Proceedings of the Digital Image Computing: Techniques and Applications, Melbourne, Australia, 1–3 December 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 81–88.
15. Kong, D.; Gray, D.; Tao, H. A viewpoint invariant approach for crowd counting. In Proceedings of the International Conference on Pattern Recognition (ICPR), Hong Kong, China, 20–24 August 2006; IEEE: Piscataway, NJ, USA, 2006; Volume 3, pp. 1187–1190.
16. Chen, K.; Loy, C.C.; Gong, S.; Xiang, T. Feature mining for localised crowd counting. In Proceedings of the British Machine Vision Conference (BMVC), Surrey, UK, 3–7 September 2012; Volume 1, p. 3.
17. 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 (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 2547–2554.
18. Chen, K.; Gong, S.; Xiang, T.; Change Loy, C. Cumulative attribute space for age and crowd density estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 2467–2474.
19. 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 (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 589–597.
20. Walach, E.; Wolf, L. Learning to count with cnn boosting. In *European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2016; pp. 660–676.
21. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 4031–4039.
22. Stewart, R.; Andriluka, M.; Ng, A.Y. End-to-end people detection in crowded scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 2325–2333.
23. Liu, L.; Qiu, Z.; Li, G.; Liu, S.; Ouyang, W.; Lin, L. Crowd counting with deep structured scale integration network. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27–28 October 2019; pp. 1774–1783.
24. Guo, D.; Li, K.; Zha, Z.J.; Wang, M. Dadnet: Dilated-attention-deformable convnet for crowd counting. In Proceedings of the ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1823–1832.
25. Zhang, L.; Shi, Z.; Cheng, M.M.; Liu, Y.; Bian, J.W.; Zhou, J.T.; Zheng, G.; Zeng, Z. Nonlinear regression via deep negative correlation learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**. [CrossRef]
26. 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 (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1821–1830.

27. Basalamah, S.; Khan, S.D.; Ullah, H. Scale driven convolutional neural network model for people counting and localization in crowd scenes. *IEEE Access* **2019**, *7*, 71576–71584. [CrossRef]
28. 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 (CVPR), Long Beach, CA, USA, 15–20 June 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1217–1226.
29. Wen, L.; Du, D.; Zhu, P.; Hu, Q.; Wang, Q.; Bo, L.; Lyu, S. Drone-based Joint Density Map Estimation, Localization and Tracking with Space-Time Multi-Scale Attention Network. *arXiv* **2019**, arXiv:1912.01811.
30. Li, W.; Mahadevan, V.; Vasconcelos, N. Anomaly detection and localization in crowded scenes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 18–32.
31. 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 (CVPR), Salt Lake City, UT, USA, 18–21 June 2018; pp. 1091–1100.
32. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 532–546.
33. Onoro-Rubio, D.; López-Sastre, R.J. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2016; pp. 615–629.
34. 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 (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8198–8207.
35. Sam, D.B.; Sajjan, N.N.; Maurya, H.; Babu, R.V. Almost unsupervised learning for dense crowd counting. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8868–8875.
36. Liu, X.; van de Weijer, J.; Bagdanov, A.D. Exploiting unlabeled data in cnns by self-supervised learning to rank. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1862–1878. [CrossRef] [PubMed]
37. Chaker, R.; Al Aghbari, Z.; Junejo, I.N. Social network model for crowd anomaly detection and localization. *Pattern Recognit.* **2017**, *61*, 266–281. [CrossRef]
38. Chen, S.; Fern, A.; Todorovic, S. Person count localization in videos from noisy foreground and detections. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1364–1372.
39. Matan, O.; Burges, C.J.; LeCun, Y.; Denker, J.S. Multi-digit recognition using a space displacement neural network. In *Advances in Neural Information Processing Systems*; MIT: Cambridge, MA, USA, 1992; pp. 488–495.
40. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
41. Ning, F.; Delhomme, D.; LeCun, Y.; Piano, F.; Bottou, L.; Barbano, P.E. Toward automatic phenotyping of developing embryos from videos. *IEEE Trans. Image Process.* **2005**, *14*, 1360–1371. [CrossRef]
42. Sermanet, P.; Eigen, D.; Zhang, X.; Mathieu, M.; Fergus, R.; LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv* **2013**, arXiv:1312.6229.
43. Eigen, D.; Krishnan, D.; Fergus, R. Restoring an image taken through a window covered with dirt or rain. In Proceedings of the International Conference on Computer Vision (ICCV), Sydney, Australia, 1–8 December 2013; pp. 633–640.
44. 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]
45. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
46. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2015; pp. 234–241.

47. Laradji, I.H.; Rostamzadeh, N.; Pinheiro, P.O.; Vazquez, D.; Schmidt, M. Where are the blobs: Counting by localization with point supervision. In *European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 547–562.
48. Sam, D.B.; Peri, S.V.; Kamath, A.; Babu, R.V. Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection. *arXiv* **2019**, arXiv:1906.07538.
49. 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 (ICASSP), Brighton, UK, 12–17 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 2382–2386.
50. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
51. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*; MIT: Cambridge, MA, USA, 2019; pp. 8024–8035.
52. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
53. 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 (CVPR), Anchorage, AK, USA, 23–28 June 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–7.
54. 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 IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
55. Sam, D.B.; Babu, R.V. Top-down feedback for crowd counting convolutional neural network. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018.
56. 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 (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1113–1121.
57. Zeng, L.; Xu, X.; Cai, B.; Qiu, S.; Zhang, T. Multi-scale convolutional neural networks for crowd counting. In Proceedings of the International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 465–469.
58. 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 (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5245–5254.
59. 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 (ICCV), Venice, Italy, 22–29 October 2017; pp. 1861–1870.
60. Liu, L.; Wang, H.; Li, G.; Ouyang, W.; Lin, L. Crowd counting using deep recurrent spatial-aware network. *arXiv* **2018**, arXiv:1807.00601.
61. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
62. 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 (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 7279–7288.
63. 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 (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 833–841.
64. 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 (WACV), Waikoloa Village, HI, USA, 7–11 January 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1941–1950.
65. Sindagi, V.A.; Patel, V.M. Ha-ccn: Hierarchical attention-based crowd counting network. *IEEE Trans. Image Process.* **2019**, *29*, 323–335. [CrossRef] [PubMed]

66. 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 (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 6133–6142.
67. Zhang, A.; Shen, J.; Xiao, Z.; Zhu, F.; Zhen, X.; Cao, X.; Shao, L. Relational attention network for crowd counting. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 6788–6797.