big data and cognitive computing

MDPI

Review

Advances in Convolution Neural Networks Based Crowd Counting and Density Estimation

Rafik Gouiaa [1], Moulay A. Akhloufi [1,*] and Mozhdeh Shahbazi [2,3]

1 Perception, Robotics, and Intelligent Machines Research Group (PRIME), Department of Computer Science, Université de Moncton, Moncton, NB E1A 3E9, Canada; rafik.gouiaa@primeai.ca
2 Department of Geomatics Engineering, University of Calgary, Calgary, AB T2N 1N4, Canada; mshahbazi@cgq.qc.ca
3 Centre de Géomatique du Québec, Chicoutimi, QC G7H 1Z6, Canada
* Correspondence: moulay.akhloufi@umoncton.ca

Abstract: Automatically estimating the number of people in unconstrained scenes is a crucial yet challenging task in different real-world applications, including video surveillance, public safety, urban planning, and traffic monitoring. In addition, methods developed to estimate the number of people can be adapted and applied to related tasks in various fields, such as plant counting, vehicle counting, and cell microscopy. Many challenges and problems face crowd counting, including cluttered scenes, extreme occlusions, scale variation, and changes in camera perspective. Therefore, in the past few years, tremendous research efforts have been devoted to crowd counting, and numerous excellent techniques have been proposed. The significant progress in crowd counting methods in recent years is mostly attributed to advances in deep convolution neural networks (CNNs) as well as to public crowd counting datasets. In this work, we review the papers that have been published in the last decade and provide a comprehensive survey of the recent CNNs based crowd counting techniques. We briefly review detection-based, regression-based, and traditional density estimation based approaches. Then, we delve into detail regarding the deep learning based density estimation approaches and recently published datasets. In addition, we discuss the potential applications of crowd counting and in particular its applications using unmanned aerial vehicle (UAV) images.

Keywords: density estimation; crowd counting; deep learning; CNN; UAV

Citation: Gouiaa, R.; Akhloufi, M.A.; Shahbazi, M. Advances in Convolution Neural Networks Based Crowd Counting and Density Estimation. *Big Data Cogn. Comput.* **2021**, *5*, 50. https://doi.org/10.3390/bdcc5040050

Academic Editor: Min Chen

Received: 22 August 2021
Accepted: 23 September 2021
Published: 28 September 2021

1. Introduction

In recent years, great efforts have been devoted to counting people in crowd unconstrained scenes due to its importance in applications, such as video surveillance [1], traffic monitoring [2], etc. The increasing growth of the world population and the development of urbanization has resulted in frequent crowd gatherings in numerous activities, such as stadium events, political events, and festivals (See Figure 1). In this context, crowd counting and density estimation are crucial for a better control & management and to ensure the security and the safety of the public.

Crowd counting remains a challenging task due to different difficulties related to the unconstrained scenes, such as extreme occlusions, variation in light conditions, changes in scale and camera perspective, and non-uniform density of people (See Figure 2). The aforementioned issues motivated many research communities to consider crowd counting as their main research direction, and attempted to develop more sophisticated techniques to deal with limitations in crowd counting.

In particular, with the recent progress in deep learning, convolution neural networks have been widely used to address crowd counting and made significant progress owing to their capacity of effectively modeling the scale changes of people/heads and the variation in regions' crowd density. The developed people counting techniques can be extended and applied to related tasks. Therefore, a section in this paper is dedicated to reviewing the

most important techniques that can be extended to develop potential applications using unmanned aerial vehicle (UAV) images.

Figure 1. Illustration of various unconstrained crowded scenes (**a**) Politics, (**b**) Public, (**c**) Concert, (**d**) Stadium.

Figure 2. Examples of unconstrained crowd scenes limitations: changes in perspective, scale and rotation variation of people/heads.

This work reviews papers that were published in the last decade and is organized as follows: Section 2 is dedicated to reviewing the traditional crowd counting and density estimation methods. In Section 3, we review the previous related surveys. In Section 4, we review, in detail, the CNN-based density estimation methods. Section 5, we discuss the most important public datasets along with the results of the state-of-the-art methods. We also present crowd counting applications that are based on Unmanned Aerial Vehicles (UAV) images in Section 6. Finally, we conclude our survey in Section 7.

2. Related Work and Motivation

Different approaches have been proposed to tackle the problem of crowd counting in images and videos. These approaches can be mainly divided into four categories: detection-based, regression-based, traditional density estimation, and recent CNN-based density estimation.

The scope of this survey is to focus on modern CNN-based density estimation and crowd counting approaches and review the most important techniques that can be extended to develop real-world applications using UAV images. However, first, we briefly review the detection and regression approaches using hand-crafted features.

2.1. Detection-Based Approaches

Early work on crowd counting adopted a detection framework [3–6]. Given a crowded situation, these approaches used a sliding window to detect the most visible parts of the body, which are mainly the head and the shoulders. Recently, various CNN-based object detectors have been proposed, which lead to a higher object detection performance as compared to systems based on simpler hand-crafted features [7]. In this context, we note the two-stage detectors, such as RCNN [7], Faster-RCNN [8] and Mask-RCNN [9], and the one-stage detectors, such as YOLO [10] and SDD [11]. Despite their high accuracy detection recorded in a sparse scene, these approaches do not perform well in the presence of the visual occlusions and ambiguities in crowded scenes.

2.2. Regression-Based Approaches

To overcome the limitations of the detection-based approaches, researchers attempt to formulate the crowd counting as a regression problem where they learn directly how to map the appearance of the image patches to their corresponding object density maps [12–14]. These approaches operate mainly on two steps: feature extraction and regression modeling. A variety of local features , such as SIFT [15], HOG [16], LBP [17,18], and global features, such as texture [19] and gradient [18] have been used to encode the object information. Learning a mapping from low-level features to the crowd count has been carried out using Gaussian process regression [20], linear regression [21], and ridge regression [14]

2.3. Traditional Density Estimation Based Approaches

While earlier approaches were successfully dealing with occlusion and scene cluttering, most of them regressed from global features directly to the number of objects and discard any available spatial information. In contrast, Lemptisky et al. [22] first involved spatial information in the learning process by adopting a linear mapping between local patch features and corresponding density maps. Thereby, they avoided the complex task of learning to detect and localize individual object instances and introduced a new approach to estimate an image density whose integral over any image region gives the count of objects within that region.

The learning process to estimate such density is formulated as a minimization of a regularized risk quadratic cost function, where a new appropriate loss function is introduced. Thus, the entire learning process is posed as a convex quadratic program solvable with cutting-plane optimization. To alleviate the difficulty of linear mapping, Pham et al. [23], proposed a non-linear mapping between local patch features and density maps through a random forest regressor. They obtained satisfactory results by introducing a crowdedness prior to tackle the large variation in appearance and shape between crowded image patches and non-crowded ones.

In addition, they proposed an effective forest reduction method to speed up estimation and met the real-time requirement. This method requires relatively less memory to build and store the forest. These methods incorporate the spatial information in the learning process, which improves the counting accuracy compared to the regression and detection-based approaches. However, they only used traditional hand-crafted features to extract

low-level information from local patches, which can lead to estimating a low-quality density map.

2.4. CNN-Based Density Estimation

CNN-based approaches have demonstrated a good performance in numerous computer vision problems, thus, motivating more researchers to use their ability to estimate a non-linear function mapping from crowd images to their corresponding density maps. In this context, numerous techniques have been proposed, which can be categorized into five groups according to the network architecture and the inference algorithm process, as depicted in Figure 3.

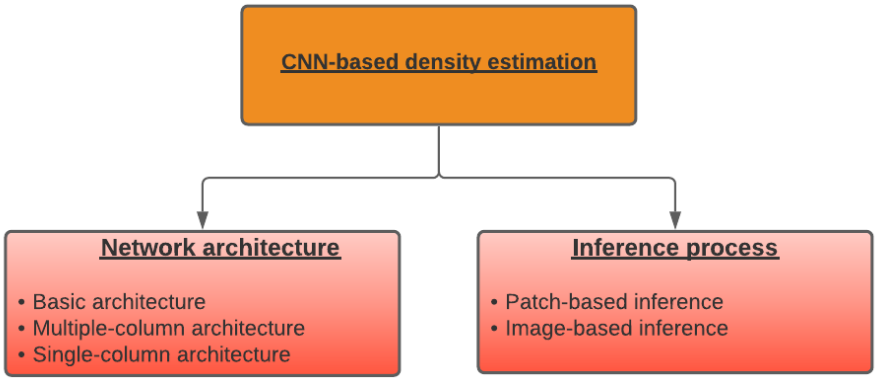


Figure 3. Taxonomy of CNN-Based density estimation.

3. Related Previous Surveys

Researchers have attempted to review the techniques of density estimation and crowd counting. Notably, Junior et al. [24] were among the first to provide a comprehensive study of the existing techniques for crowd counting. Li et al. [25] reviewed various methods for the crowded scene analysis, which covered different tasks, such as crowd motion, pattern learning, and anomaly detection in crowds. In [26], Zitouni et al. reviewed the existing visual crowd analysis techniques based on different key statistical evidence, which was inferred from the literature, and provided recommendations toward the general aspects of techniques instead of focusing on a specific algorithm.

In [27], Loy et al. provided a study that evaluates and compares the state-of-the-art techniques of visual crowd counting using the same protocol. Saleh et al. [28] presented a survey on crowd counting methods used in video surveillance and categorized the existing algorithms into two main approaches: direct and indirect. While these surveys provide detailed and comprehensive studies of the existing density estimation and crowd counting techniques, they all reviewed the traditional methods based on the hand-crafted features. Recently, Sindagi et al. [29] provided a survey of advances in CNN-based density estimation and crowd counting from a single image, published up to the year 2017. Guangshuai et al. [30] put forward a survey on CNN-based density estimation and crowd counting where over 220 papers have been reviewed up to the year 2020.

Though the last survey [30] covers the most recent CNN-based crowd counting approaches, as with the previous surveys, it focuses only on the statistical evaluation and comparison between different approaches without analyzing the importance of extending the approaches used for counting people in crowds in order to develop real-world applications in various areas. In this paper, we survey various papers that adapt crowd counting approaches to counting different objects from UAV images.

4. Taxonomy for CNN-Based Density Estimation

In this section, we review different CNN-based density estimation and crowd counting methods in view of the network architectures and the training and inference paradigm of the methods. Table 1 summarizes a categorization of different CNN-based crowd

counting and density estimation methods according to the network architecture and the inference process.

4.1. Typical CNN Architecture for Density Estimation and Crowd Counting

Based on the type of the network architecture, we classify the approaches into three major categories (see Figure 3):

4.1.1. Basic Network Architecture

These architectures are among the first deep learning approaches applied to density estimation and crowd counting. They are basically composed of convolution layers, pooling layers, and fully connected layers.

Fu et al. [31] and Wang et al. [32] were among the first researchers that attempted to use a convolution neural network in the context of crowd density estimation. Wang et al. [32] proposed the first end-to-end deep convolution neural network regression model for counting people in images of extremely dense crowds. The proposed architecture is composed of five Conv-layers and two fully connected layers, where its output is the estimated people counts in the input image. In addition, to reduce the false positive errors, which are mainly caused by the existence of trees and buildings in the background, training data are augmented by adding negative samples whose ground truth count is set as zero.

In a different approach, Fu et al. [31] used the multi-stage ConvNet model proposed in [33] to ensure better shift, scale and distortion invariance. To reduce the computation time at both training and detection stages, they optimized the model by discarding all similar features maps. In addition, two optimized multi-stage ConvNets are cascaded as a strong classifier to achieve boosting in which the first classifier is trained to pick out the hard samples, whereas the second one is trained to give them a final determination. Yao et al. [34] fine-tuned several architectures of deep residual network (ResNet [35]) to develop a cell counting framework.

Elad et al. [36] used a basic CNN and incorporated layered boosting and selective sampling to enhance the accuracy and the training computation. The training process is done in stages, where CNNs are iteratively added so that each new CNN is trained on the difference between the estimation of its predecessor and the ground truth. The selective sampling approach is used to speed up the training process by reducing the effect of low-quality samples, such as trivial samples and outlier samples.

As stated by the authors, trivial samples are those that are correctly classified early on. Feeding again these samples to the network tends to introduce a bias toward them, thereby, affecting its generalization performance. On the other hand, the presence of outliers, such as mislabeled samples, can affect the generalization of the model and, in particular, increase the computation of the training time (i.e., due to the boosting technique).

These basic CNNs approaches are simple and easy to implement. However, their performance is often limited due to the quality of the extracted features, which are usually not invariant to perspective effect or image resolution.

Table 1. Categorization of existing CNN-based methods.

	Category	
Methods	**Network Architecture**	**Inference Paradigm**
Wang et al [32]	Basic	Patch-based
Fu et al [31]	Basic	Patch-based
Yao et al. [34]	Basic	Pacth-based
Elad et al. [36]	Basic	Patch-based
Zhang et al. [37]	Multiple-column	Patch-based
Boominathan et al. [38]	Multiple-column	Patch-based

Table 1. *Cont.*

	Category	
Methods	**Network Architecture**	**Inference Paradigm**
Oñoro-Rubio et al. [12]	Multiple-column	Patch-based
Deepak et al. [39]	Multiple-column	Patch-based
Deepak et al. [40]	Multiple-column	Patch-based
Liu et al. [41]	Multiple-column	Patch-based
Zhang et al [42]	Multiple-column	Patch-based
Hossain et al. [43]	Multiple-column	Patch-based
Guo et al. [44]	Multiple-column	Patch-based
Jiang et al. [45]	Multiple-column	Patch-based
Li et al. [46]	Single-column	Whole-image
Zhang et al. [47]	Signle-column	Patch-based
Wang et al. [48]	Single-column	Patch-based
Cao et al. [49]	Single-column	Patch-based
Varun et al. [50]	Single-column	Patch-based
Xiaolong et al. [51]	Single-column	Patch-based
Mohammed et al. [52]	Single-column	Patch-based
Liu et al. [53]	Single-column	Patch-based
Zhang et al. [54]	Multiple-column	Patch-based
Tian et al. [55]	Multiple-column	Patch-based
Sajid et al. [56]	Multiple-column	Patch-based
Chong et al. [57]	Multiple-column	Patch-based

4.1.2. Multiple-Column Architecture

These network architectures incorporate multiple columns to extract multi-scale features that allow generating high-quality crowd density maps.

Zhang et al. [37] were among the first to introduce the idea of using multiple-column architecture for crowd counting. They proposed Multi-column Convolutional Neural Network (MCNN) architecture to map the image to its crowd density map. As depicted in Figure 4, MCNN incorporates different columns where each one adopts filters with receptive fields of different sizes, so that the extracted features are adaptive to scene variations (i.e., people/head size). In addition, they proposed a new large dataset with around 330,000 head annotations and showed that MCNN is easily transferred for cross-scene crowd counting.

In [38], the authors introduced CrowdNet, which combines deep and shallow networks at two different columns, so that the shallow network is used to extract low-level features, whereas the deep network is used to extract high-level features. Both extracted features are crucial for detecting people under large-scale variations and severe occlusion. Hydra-CNN is introduced in [12]. It uses a pyramid of input patches so that each level has a different scale. By doing that, Hydra-CNN extracts multi-scale features, which are combined to generate the crowd density map.

Deepak et al. [39] developed a switching-CNN for crowd counting by training various CNN crowd density regressors on patches from a crowd scene. The regressors were designed to incorporate different receptive fields. In addition, a switch classifier was

trained to select the best regressor that estimates the density map corresponding to the crowd scene patch.

In another work, Deepak et al. [40] proposed a top-down feedback architecture that carries high-level features to correct false predictions. This architecture has a bottom-up CNN, which is directly connected to a top-down CNN so that the top-down generated feedback to the bottom-up to help deliver a better crowd density map. In [41], Liu et al. proposed a CNN framework to address the issues of scale variation and rotation variation using the Spatial Transform Network [58].

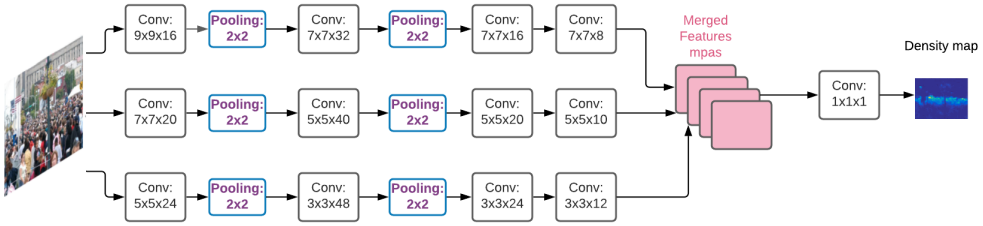


Figure 4. The MCNN architecture proposed in [37].

Zhang et al. [42] exploited the attention mechanism to address the limitations of the pixel-wise regression technique, which is very popular for estimating the crowd density map. This technique assumes the interdependence of pixels, which leads to noisy and inconsistent predictions. Thus, the proposed Relational Attention Network (RANet) incorporates local self-attention (LSA) and global self-attention (GSA) to capture the interdependence of pixels. In addition, a relational model is used to combine LSA and GSA to obtain a more informative aggregated feature representation.

In the same context, Hossain et al. [43] proposed a CNN model based on the attention mechanism to extract global and local scale features appropriate for the image. By combining both local and global features, the model outputs an improved crowd density map. Guo et al. [44] introduced a deep model called Dilated-Attention-Deformable ConvNet (DADNet), which incorporates two modules: multi-scale dilated attention and deformable convolutional DME (Density Map Estimation). The multi-scale dilated attention is based on using various kernel dilation levels to extract different visual features of crowd regions of interest, whereas the deformable convolution is used to generate a high-quality density map.

Observing that most of the existing techniques are susceptible to overestimate or underestimate people counts of regions with different patterns, Jiang et al. [45], introduced a new CNN model based on the attention mechanism, which consists of two components Density Attention Network (DANet) and Attention Scaling Network (ASNet). DANet is used to extract attention masks from regions of different density levels. ASNet, on the other hand, outputs density maps and scaling factors and multiplies them by the attention masks yielding separate attention-based density maps. These maps are then summed to form that final density map.

Liu et al. introduced DecideNet [59], which consists of detection and regression based density maps. These two count modes are used to deal with the crowd density variation in the image regions. An attention module is used to adaptively assess the reliability of the two count modes. Liu et al. [60] proposed a self-supervised method to improve the training of models for crowd counting. This was based on the fact that crops sampled from a crowd image contain the same or fewer persons than the original image.

Thus, crops can be ranked and used to train a model to estimate whether one image contains more persons than another image. Fine-tuning the resulting model on a small labeled dataset achieved state-of-the-art results. In [61], PACNN a perspective-aware CNN was introduced. It is specifically designed to predict multi-scale perspective maps and to deal with the perspective distortion problem.

Dingkang Liang et al. [62] introduced a weakly-supervised crowd counting in images by introducing TransCrowd, which is a sequence-to-count framework based on a

Transformer-encoder. In the same context, Sun et al. [63] used transformers to encode features with global receptive fields and proposed two modules: a token-attention module and regression-token module. In [64], Gao et al. proposed a crowd localization network called the Dilated Convolutional Swin Transformer (DCST). It provides the location information of each instance in addition to counting the numbers for a scene.

Despite the significant improvement and the great performance recorded by the multi-column architecture, they are still suffering from various limitations as detailed in [46]. The multi-columns CNNs are difficult to train since they require huge computation times and memory. Such architecture can generate redundant features since the columns implement almost the same network architecture. Moreover, multi-column architecture often requires a density level classifier before feeding the image to the network. However, since the number of objects is varying widely in a congested scene, it is very difficult to estimate the granularity of the crowd density maps. In addition, using crowd density level classifiers leads to the implementation of more columns, which increases the complexity of the architecture and yields more feature redundancy.

These limitations motivated some researchers to adopt single-column CNNs, which are a much simpler yet efficient architecture to overcome these disadvantages and deal with different challenging scenarios in crowd counting.

4.1.3. Single-Column Architecture

The single-column network architectures are based on deeper end-to-end CNNs. This implements more specific features to deal with critical problems in crowd counting and generate high-quality crowd density maps.

In [46], Li et al. proposed CRSNet, a CNN model for crowd counting in highly congested scenes. This model consists mainly of a frond-end CNN used to extract 2D features and a back-end dilated CNN, which uses dilated kernels to provide larger receptive fields and to replace pooling operations. The dilated convolution layers expand the receptive field without losing resolution and, thereby, aggregate the multi-scale contextual information. CSRNet is an easy-trained end-to-end approach that generates high-quality density maps and achieves state-of-the-art performance on four datasets.

Zhang et al. [47] proposed a scale-adaptive convolution neural network for crowd counting. The proposed architecture is composed of a backbone, which is a fully convolution neural network (FCN) with fixed small receptive fields. It extracts feature maps with different dimensions from multiple layers. The extracted feature maps are resized to have the same output dimension and combined to calculate the final crowd density map. Two loss functions have been introduced in the training process, density map loss and count loss, to jointly optimize the model. The loss count is used to reduce the variance of the prediction error and enhance the generalization performance of the model on a very sparse scene. This architecture is illustrated in Figure 5.

Wang et al. [48] proposed SCNet, which is a compact single-column architecture for crowd counting. It consists of three modules: residual fusion modules (RFM) to extract multi-scale features, a pyramid pooling module (PPM) to combine features at different stages, and a sub-pixel convolutional module (SPCM) followed by an upsampling layer to recover the resolution. These three modules allow SCNet to generate multi-scale features, which leads to generating a high-quality density map. In [65], Shi et al. proposed a learning strategy called deep negative correlation learning (NCL), based on learning a pool of decorrelated regressors.

In a different approach, Cao et al. [49] proposed an encoder–decoder based on the inception network [66] called SANet for crowd counting. The encoder is used to extract multi-scale features, whereas the decoder used transposed convolution layers to upscale the extracted features and generate the final crowd density map. In addition, unlike most of the existing approaches, which use only Euclidean loss that ignores the correlation between pixels of the density, they introduced a new training loss that combines the Euclidean loss and local pattern consistency loss.

In [50], Varun et al. extended the U-Net [67] by adding a decoding reinforcement branch to accelerate the training of the network and using Structural Similarity Index to maintain the local correlation of the density map in order to generate a good crowd density map. Xiaolong et al. [51] proposed a new trellis encoder–decoder architecture, which consists of multiple decoding paths and a multi-scale encoder. The multiple decoding paths are used to hierarchically aggregate features at different decoding stages, whereas the multi-scale encoder is incorporated to preserve the localization precision in the encoded feature maps.

In recent years, models based on attention mechanisms have demonstrated significant performance in different computer vision tasks [52]. Instead of extracting features from the entire image, the attention mechanism allows models to focus only on the most relevant regions. In this context, Mnih et al. [52] used the attention mechanism to introduce a scale-aware attention network to address the scale variation in crowd counting images. Due to the attention mechanism, the model can automatically focus on the most important global and local features and combine them to generate a high-quality crowd density map.

Liu et al. also proposed the ADCrowdNet [53], which is an attention-based network for crowd counting. ADCrowdNet consists of two concatenated networks: an Attention Map Generator (AMG), which first estimates crowd regions in images as well as their congestion degree, and a Density Map Estimator (DME), which is a multi-scale deformable network that uses the output of AMG to generate a crowd density map.

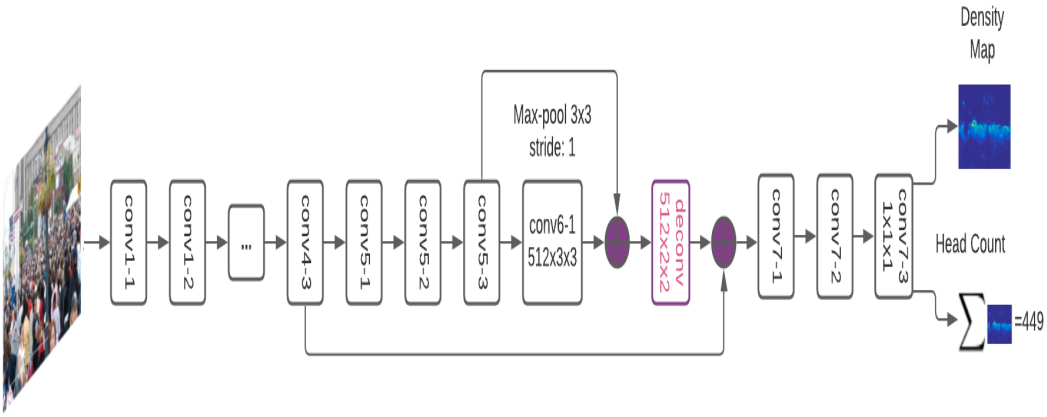


Figure 5. Illustration of scale-adaptive CNN for crowd counting proposed by Zhang et al. [47].

Due to the simplicity of their architectures and their effective training process, single-column network approaches have received more attention in recent years.

4.2. Typical Inference Paradigm

Based on the inference methodology, we can categorize the crowd density estimation techniques into the following two categories:

4.2.1. Patch-Based Inference

The patch-based model is trained on random crops from the original image. During the inference, a sliding window is applied to the test image, and the prediction is obtained for each crop. The total count is obtained by summing the counts over all the crops.

In [54], the model is trained on random patches extracted from the input images so that every crop cover 3 by 3 m square in the actual scene. The patches are then resized to 72 × 72 pixels (see Figure 6) and fed as input to the CNN model to obtain the corresponding crowd density map. The number of objects in every patch is obtained by integrating over the crowd density map. In [55], a new CNN model called PaDNet was proposed.

It consists of three modules as follows: (1) the Density-Aware Network (DAN) incorporates multiple CNN sub-networks, which are pre-trained on images with different crowd density levels, and used to capture the crowd density level information; (2) the Feature

Enhancement Layer (FEL) generates weighted local and global contextual features; and (3) the Feature Fusion Network (FFN) is used to combine these contextual features. The network is trained on patches so that nine crops are taken from every input image.

In [56], Sajid et al. proposed a plug-and-play-based patch rescaling module (PRM) to address the problem of crowd diversity in the scene. As shown in Figure 7, the PRM module takes a patch image as input and then rescales it using the appropriate scaler (Up-scaler or Down-scaler) according to its crowd density level, which is computed by the classifier before using PRM. In this approach, the low-crowd and high-crowd regions pass directly through the Up-scaler or Down-scaler, the Medium-crowd bypasses the PRM without rescaling, whereas the no-crowd regions are automatically discarded.

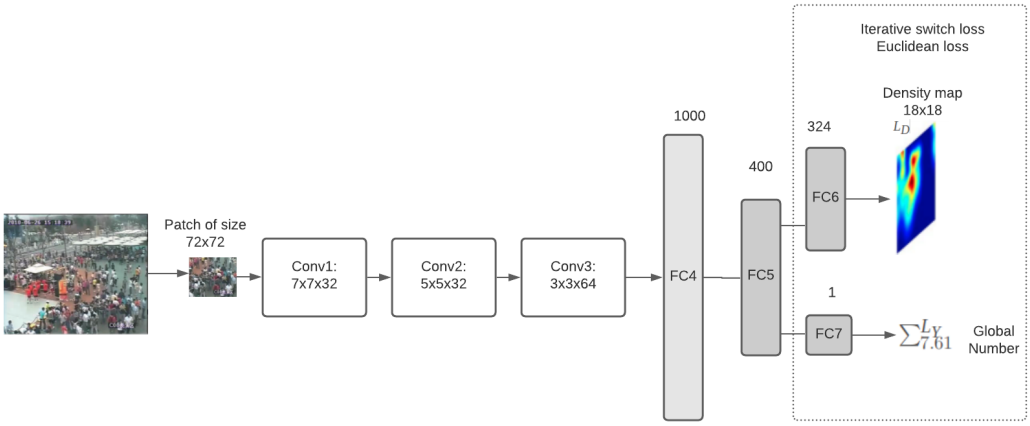
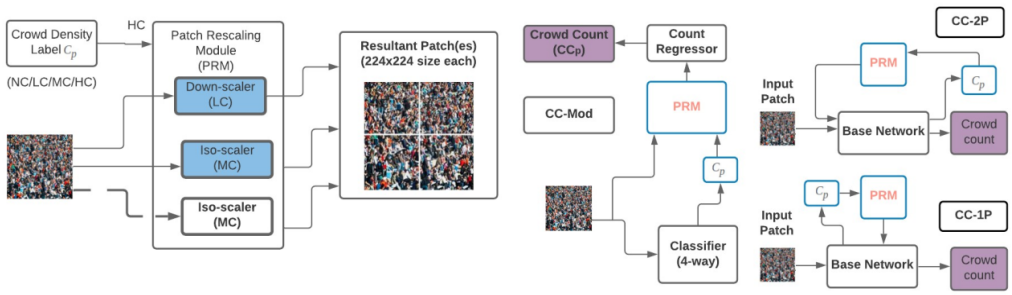


Figure 6. The patched-based inference approach proposed in [54].



(a) PRM Module (b) Modular Scheme(CC-Mod) (c) End-to-End Networks (CC-1P, CC-2P)

Figure 7. The PRM module presented in [56].

In [68], Sam et al. proposed a hierarchical CNN tree where the CNN child regressors are more accurate than any of their parents. At test time, a classifier guides the input image patches to the appropriate regressors.

4.2.2. Image-Based Inference

By taking random patches from the input image, patch-based methods ignore the global information and also require a huge computation during the inference due to the use of a sliding window. Training with the whole image help exploit the global information. However, it is still dependant on the resolution of the image, which is often very large in the context of crowd counting.

In [57], Chong et al. proposed an end-to-end CNN model that takes the whole image as input and directly produces the final count. First, the image is fed into a pre-trained CNN to obtain high-level features, which are then mapped to local counting numbers using a recurrent neural network with memory cells. In addition, sharing computation

over overlapping regions leads to a reduction in model complexity and allows the model to incorporate contextual information when predicting both local and global counts.

5. Datasets and Results

With the increasing development of crowd counting approaches, numerous datasets have been proposed over the last decade to drive research on crowd counting and develop models to deal with various limitations including changes in perspective and scale, variation in light conditions, crowd density, cluttering, and severe occlusion. Table 2 summarizes the most popular datasets, which can be categorized into three different groups according to the view type: free view, crowd surveillance view, and drone view.

Some images of these categories are depicted in Figure 8.

Figure 8. Sample images from various datasets.

- **UCSD** [69]: It was among the first datasets to be collected to count people. It was acquired with a stationary camera mounted at an elevated position, overlooking pedestrian walkways. The dataset contains 2000 frames of size 158 × 512 along with annotations of pedestrians in every 1/5 frames, while the other frames are annotated using linear interpolation. It provides also the bounding box coordinates for every pedestrian. The dataset has 49,885 person instances, which are split into training and test subsets. UCSD has a low-density crowd with an average of 25 pedestrian instances, and the perspective across images does not change greatly since all images are captured from the same location.
- **Mall** [70]: This dataset is collected from a publicly accessible webcam in a shopping mall. The video sequence of the dataset contains over 2000 frames of size 640 × 480 in which 62,325 heads were annotated with an average of 25 heads per image. By comparing to UCSD, the Mall dataset was created with higher crowd densities as well as more significant changes in illumination conditions and different activity patterns (static vs. moving people). The scene has severe perspective distortion along the video sequence, which results in large variations in scale and appearance of objects. In addition, there exist severe occlusions caused by different objects in the mall.
- **UCF_CC_50** [13]: It is the first challenging dataset, which was created by directly scraping publicly web images. The dataset presents a wide range of crowd densities along with large varying perspective distortion. It contains only 50 images whose size is 2101 × 2888 pixels. These images contain a total of 241,677 head instances with an average of 1279 heads in each image. Due to its small size, the performance of recent CNN-based models is far from optimal.
- **WorldExpo'10** [54]: Zhang et al. [54] remarked that most existing crowd counting methods are scene-specific and their performance drops significantly when they are applied to unseen scenes with different layouts. To deal with this, they introduced the WorldExpo'10 dataset to perform a data-driven cross-scene crowd counting. They collected the data from Shanghai 2010 World-Expo, which contains 1132 video sequences captured by 108 cameras with an image resolution of 576 × 720 pixels. The dataset contains 3980 frames that contain a total of 200,000 annotated heads for an average of 50 heads by frame.
- **AHU-Crowd** [71]: It is composed of diverse video sequences representing dense crowds in different public places including stations, stadiums, rallies, marathons, and pilgrimage. The sequences have different perspective views, resolutions, and crowd densities and cover a large multitude of motion behaviors for both obvious and subtle instabilities. The dataset contains 107 frames whose size is 720 × 576 pixels, and 45,000 annotated heads.
- **ShanghaiTechRGBD** [72]: It is a large-scale dataset composed of 2193 for a total of 144,512 annotated head count. The images are captured by a stereo camera with a valid depth ranging from 0 to 20 m. The images are captured in very busy streets of metropolitan areas and crowded public parks, while the light conditions vary from very bright to very dark.
- **CityUHK-X** [73]: It contains 55 scenes captured using a moving camera with a tilt angle range of $[-10^{\circ}, -65^{\circ}]$ and a height range of [2.2, 16.0] meters. The dataset is split into training and test subsets. The training subset is composed of 43 scenes for a total of 2503 images and 78,592 people, while the test subset is composed of 12 scenes for a total of 688 images and 28,191 people.
- **SmartCity** [47]: It consists of 50 images, collected from 10 different cities for outdoor scenes of different places, such as shopping malls, office entrances, sidewalks, and atriums.
- **Crowd Surveillance** [74]: It is composed of 13,945 high-resolution images. It is split into 10,880 images for training and 3065 images for testing for a total of 386,513 head count.

- **DroneCrowd** [75]: It was captured using a drone-mounted camera and recorded at 25 frames per second with a resolution of 1920 × 1080 pixels. It contains 112 video clips with 33,600 frames. The annotation was performed by over 20 experts for more than two months so that more than 4.8 million heads are annotated on 20,800 people trajectories.
- **DLR-ACD** [76]: It is a collection of 33 aerial images for crowd counting and density estimation. It was captured through 16 different flights and over various urban scenes including sports events, city centers, and festivals.
- **Fudan-ShanghaiTech** [77]: It is a large-scale video crowd counting dataset, and it is the largest dataset for crowd counting and density estimation. It is composed of 100 videos captured from 13 different scenes. It contains 150,000 images for a total of 394,081 annotated head count.
- **Venice** [78]: It is a small dataset acquired in Piazza San Marco in Venice (Italy). It contains four different sequences for a total of 167 annotated images with a resolution of 1280 × 720 pixels.
- **CityStreet** [79]: It was collected from a busy city street using a multiview camera system, which is composed of five synchronized cameras. The dataset contains a total of 500 multi-view images in total.
- **DISCO** [80]: It was collected to jointly utilize ambient sounds and visual contexts for crowd counting. The dataset contains a total of 248 video clips, where each clip was recorded at 25 frames per second with a resolution of 1920 × 1080.
- **DroneVehicle** [81]: It consists of 15,532 pairs of RGB and infrared images for a total of 441,642 annotated objects. The images were acquired by a drone-mounted camera over various urban areas, including different types of urban roads, residential areas, and parking lots from day to night.
- **NWPU-Crowd** [82]: It contains 5109 images for a total of 2,133,375 annotated heads with point and box labels. Compared to existing datasets, it has negative samples and a large appearance variation.
- **JHU-CROWD++** [83]: It is composed of 4372 images and 1.51 million annotations and acquired under various scenarios and environmental conditions. Labeling is provided in different formats, including dots, approximate bounding boxes, and blur levels.

Table 2. Summary of various datasets, including free-view datasets, crowd-surveillance view, and drone-view.

Name	Year	Attributes	Avg. Resolution	No. Samples	No. Instances	Avg. Count
Free view datasets						
NWPU-Crowd [82]	2020	Congested, Localization	2191 × 3209	5109	2,133,375	418
JHU-CROWD++ [83]	2020	Congested	1430 × 910	4372	1,515,005	346
JHU-CROWD++ [84]	2018	Congested	2013 × 2902	1535	1,251,642	815
ShanghaiTech Part A [85]	2016	Congested	589 × 868	482	241,677	501
UCF_CC_50 [13]	2013	Congested	2101 × 2888	50	241,677	1279
Crowd Surveillance-view						
DISCO [80]	2020	Audiovisual, extreme conditions	1080 × 1920	1935	170,270	88
Crowd Surveillance [74]	2019	Free scenes	840 × 1342	13,945	386,513	28
ShanghaiTechRGBD [72]	2019	Depth	1080 × 1920	2193	144,512	65.9
Fudan-ShanghaiTech [77]	2019	400 Fixed Scenes, Synthetic	1080 × 1920	15,211	7,625,843	501
Venice [78]	2019	4 Fixed Scenes	720 × 1280	167	-	-
CityStreet [79]	2019	Multi-view	1520 × 2704	500	-	-
SmartCity [47]	2018	-	1080 × 1920	50	369	7
CityUHK-X [73]	2017	55 Fixed Scenes	384 × 512	3191	106,783	33
ShanghaiTech Part B [71]	2016	Free Scenes	768 × 1024	716	88,488	123

Table 2. *Cont.*

Name	Year	Attributes	Avg. Resolution	No. Samples	No. Instances	Avg. Count
AHU-Crowd [71]	2016	-	720 × 576	107	45,000	421
WorldExpo'10 [54]	2015	108 Fixed Scenes	576 × 720	3980	199,923	50
Mall [70]	2012	1 Fixed Scene	480 × 640	2000	62,325	31
UCSD [69]	2008	1 Fixed Scene	158 × 238	2000	49,885	25
			Drone-View			
DroneVehicle [81]	2020	Vehicle	840 × 712	31,064	441,642	14.2
DroneCrowd [75]	2019	Video	1080 × 1920	33,600	4,864,280	145
DLR-ACD [76]	2019	-	-	33	226,291	6857

6. Results and Discussions

We report results of recent traditional approaches along with the CNN-based methods on the most popular datasets. The count estimation performance is reported directly from the original public work. We compare different methods based on the following metrics:

$$Mean\ Absolute\ Error(MAE) = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'| \tag{1}$$

$$RootMean\ Square\ Error(RMSE) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - y_i')^2} \tag{2}$$

- N: is the number of test samples.
- y_i is the ground truth result corresponding to sample i.
- y_i' is the estimated result corresponding to sample i.

The comparison results are summarized in Table 3. In general, CNN-based methods highly outperformed the traditional approaches. CNN-based methods showed effective results in a very cluttered scene with large density crowds and under different scene conditions (lighting, scaling, etc.). While the multiple-column techniques achieved state-of-the-art results on three datasets: UCF_CC_50, ShanghaiTech Part A, and Mall, some single-column techniques also achieved a high performance, such as [53], which obtained state-of-the-art results on ShanghaiTech Part B.

In addition, CSRNet [46] and SaNet [49] showed comparable results on almost all datasets. The single-column technique in [53], which involves the attention mechanism, achieved state-of-the-art results on the WorlExpo'10 dataset. Finally, single-column techniques, like [46,48,49], not only presented great performances but are also easy to implement and were applied in a real-time scenario.

Table 3. Comparison of crowd density estimation on various datasets.

Approach Type	Dataset	Mall		UCF CC 50		WorldExpo 10		UCSD		UCF-QNRF		ShanghaiTech Part A		ShanghaiTech Part B	
	Method	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**
Traditional approach	Learning To Count Objects in Images [22]	-	-	-	-	-	-	1.59	-	-	-	-	-	-	-
	COUNT Forest [23]	2.5	10.0	-	-	-	-	1.61	4.40	-	-	-	-	-	-
	Multi-source Multi-scale Counting [13]	-	-	468.0	590.3	-	-	-	-	-	-	-	-	-	-
Multiple-column approaches	MCNN [37]	-	-	377.6	509.1	11.6	-	1.07	1.35	-	-	110.2	173.2	26.4	41.3
	Cross-scene crowd counting [54]	-	-	467.0	498.5	12.9	-	1.60	3.31	-	-	181.8	277.7	32.0	49.8
	Hydra-CNN [12]	-	-	333.7	425.2	-	-	-	-	-	-	-	-	-	-
	Switching-CNN [39]	-	-	318.1	439.2	9.4	-	1.62	2.10	228	445	90.4	135	21.6	33.4
	Crowd counting using deep recurrent net. [41]	1.72	2.1	219.2	250.2	7.76	-	-	-	-	-	69.3	96.4	11.1	18.2
	RANet [42]	-	-	239.8	319.4	-	-	-	-	111	190	59.4	102.0	7.9	12.9
	DADNet [44]	-	-	285.5	389.7	-	-	-	-	-	-	64.2	99.9	-	-
	DANet [45]	-	-	268.3	373.2	-	-	-	-	-	-	71.4	120.6	9.1	14.7
Single-column approach	CRSNet [46]	-	-	266.1	397.5	-	-	-	-	-	-	68.2	115	10.6	16
	SaCNN [47]	-	-	314.9	424.8	8.5	-	-	-	-	-	86.8	139.2	16.2	25.8
	SCNet [48]	-	-	280.5	332.8	8.4	-	-	-	-	-	71.9	117.9	9.3	14.4
	SANet [49]	-	-	258.4	334.9	-	-	-	-	-	-	67.0	104.5	8.4	13.6
	ADCrowdNet (AMG-attn-DME) [53]	-	-	273.6	362.0	7.3	-	1.09	1.35	-	-	70.9	115.2	7.7	12.9

7. Potential Application of Crowd Counting

Crowd counting techniques have been applied to count and estimate the number of persons in crowded and cluttered scenes to develop real-world applications, mostly related to video surveillance and public safety. In addition, these techniques have been adapted and applied to different related problems, such as traffic control, plant/fruits counting. Different applications use different image sources, including a fixed camera, multi-cameras, a moving camera, unmanned aerial vehicles (UAVs), etc.

In this section, we review the crowd counting applications developed using unmanned aerial vehicles (UAV). In [86,87], the authors introduced an automated vehicle detection and counting system in high-resolution aerial images. The proposed method used a convolution neural network to generate a vehicle spatial density map from the aerial image.

Jingyu et al. [88] introduced an efficient convolution neural network called Flounder-Net, which used aerial images captured by a drone (See Figure 9), to count crowd people for a security purpose. Flounder-Net architecture (See Figure 10) involves an interleaved group convolution to eliminate the redundancy of the network, and the rapid shrink of feature maps in order to tackle the high-resolution problem. The model is implemented and integrated into the embedded system of the drone. In the same context, Castellano et al. [89] introduced a light-weight and fast fully-convolutional neural network to regress a crowd density map on aerial images captured by a drone.

Recently, crowd counting techniques have been applied in the agriculture domain. In [90], Jintao et al. developed and implemented an automatic counting of in situ rice seedlings in the field using a basic fully convolution neural network to regress a crowd density map. The system takes, as input, aerial images captured by an UAV equipped with RGB cameras. Sungchan et al. [91] implemented an automatic cotton plant counting by adapting the Yolo3 [92] deep learning algorithm. In the same context, Kitano et al. [93] used a fully convolution neural network to develop an application that captured images using a UAV and returned the number of corn plants.

The technology of crowd counting is being increasingly adapted to agriculture, which facilitates the decision-making of the farmer and the management process of work-labor and products.

Figure 9. Crowd counting using a drone.

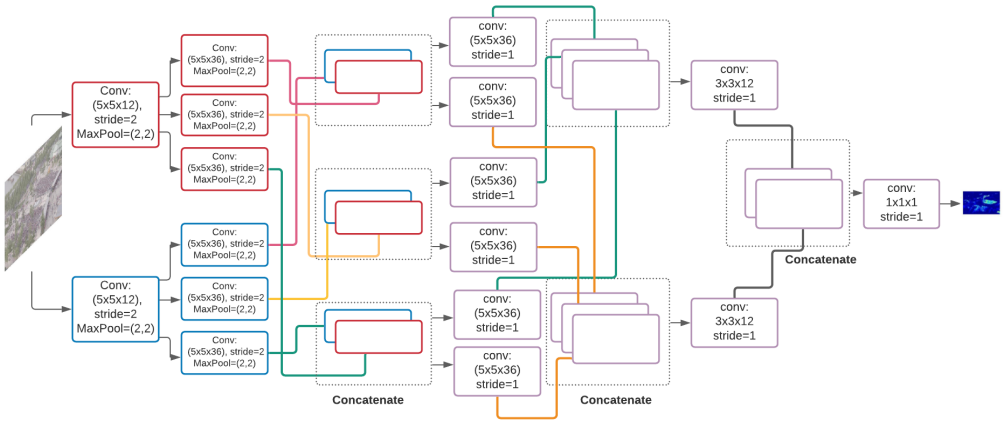


Figure 10. Flounder-Net architecture as presented in [88].

8. Conclusions

In recent years, the need for crowd counting in many areas has greatly boosted research in crowd counting and density estimation. With the development of deep learning, the performance of crowd counting models has been remarkably improved, and the real-world applications scenarios have been expanded. This paper presented a survey on the recent advances in convolution neural network (CNN)-based crowd counting and density estimation. We explored the existing approaches from different perspectives, including the network architecture and the learning paradigm. We presented a description of the most popular datasets that are used to evaluate the crowd counting models. In addition, we conducted a performance evaluation for the most representative crowd counting algorithms. Finally, we reviewed the potential applications of crowd counting in the context of unmanned aerial vehicle (UAV) images.

Author Contributions: Conceptualization, R.G. and M.A.A.; methodology, R.G. and M.A.A.; software, R.G.; validation, R.G., M.A.A. and M.S.; formal analysis, R.G., M.A.A. and M.S.; investigation, M.A.A. and M.S.; resources, R.G.; data curation, R.G.; writing—original draft preparation, R.G., M.A.A. and M.S.; writing—review and editing, R.G., M.A.A. and M.S.; visualization, R.G.; supervision, M.A.A. and M.S.; project administration, M.A.A. and M.S.; funding acquisition, M.A.A. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was enabled in part by support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC), funding reference number RGPIN-2018-06233 and the Mitacs Accelerate program.

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.

References

1. Xiong, F.; Shi, X.; Yeung, D.Y. Spatiotemporal modeling for crowd counting in videos. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5151–5159.
2. Zhang, S.; Wu, G.; Costeira, J.P.; Moura, J.M. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3667–3676.
3. Dollar, P.; Wojek, C.; Schiele, B.; Perona, P. Pedestrian Detection: An Evaluation of the State of the Art. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 743–761. [CrossRef]
4. Xu, H.; Lv, P.; Meng, L. A people counting system based on head-shoulder detection and tracking in surveillance video. In Proceedings of the 2010 International Conference on Computer Design and Applications, Qinhuangdao, China, 25–27 June 2010; Volume 1, pp. V1-394–V1-398.

5. Subburaman, V.; Descamps, A.; Carincotte, C. Counting People in the Crowd Using a Generic Head Detector. In Proceedings of the 2012 9th IEEE International Conference on Advanced Video and Signal Based Surveillance, IEEE Computer Society, Beijing, China, 18–21 September 2012; pp. 470–475. [CrossRef]
6. Topkaya, I.S.; Erdogan, H.; Porikli, F. Counting people by clustering person detector outputs. In Proceedings of the 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Seoul, Korea, 26–29 August 2014; pp. 313–318.
7. Girshick, R.B.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June, 2014; pp. 580–587.
8. Ren, S.; He, K.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef]
9. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
10. Redmon, J.; Divvala, S.; Girshick, R.B.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
11. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A. SSD: Single Shot MultiBox Detector. *arXiv* **2016**, arXiv:1512.02325.
12. Oñoro-Rubio, D.; López-Sastre, R.J. Towards Perspective-Free Object Counting with Deep Learning. In *Computer Vision—ECCV 2016;* Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 615–629.
13. Idrees, H.; Saleemi, I.; Seibert, C.; Shah, M. Multi-source Multi-scale Counting in Extremely Dense Crowd Images. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 25–27 June 2013; pp. 2547–2554.
14. Chen, K.; Loy, C.C.; Gong, S.; Xiang, T. Feature Mining for Localised Crowd Counting. Available online: http://www.bmva.org/bmvc/2012/BMVC/paper021/paper021.pdf (accessed on 16 September 2021).
15. Tota, K.; Idrees, H. Counting in Dense Crowds using Deep Features. Available online: https://www.crcv.ucf.edu/REU/2015/Tota/Karunya_finalreport.pdf (accessed on 16 September 2021).
16. Ma, Z.; Chan, A.B. Crossing the Line: Crowd Counting by Integer Programming with Local Features. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 25–27 June 2013; pp. 2539–2546.
17. Wang, Z.; Liu, H.; Qian, Y.; Xu, T. Crowd Density Estimation Based on Local Binary Pattern Co-Occurrence Matrix. In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops, Melbourne, Australia, 9–13 July 2012; pp. 372–377.
18. Balbin, J.R.; Garcia, R.G.; Fernandez, K.E.D.; Golosinda, N.P.G.; Magpayo, K.D.G.; Velasco, R.J.B. Crowd counting system by facial recognition using Histogram of Oriented Gradients, Completed Local Binary Pattern, Gray-Level Co-Occurrence Matrix and Unmanned Aerial Vehicle. In *Third International Workshop on Pattern Recognition*; Jiang, X., Chen, Z., Chen, G., Eds.; International Society for Optics and Photonics, SPIE: Bellingham,WA, USA, 2018; Volume 10828, pp. 238–242. [CrossRef]
19. Ghidoni, S.; Cielniak, G.; Menegatti, E. Texture-Based Crowd Detection and Localisation. In *Intelligent Autonomous Systems 12*; Springer: Berlin/Heidelberg, Germany, 2013; Volume 193, [CrossRef]
20. Chan, A.B.; Vasconcelos, N. Counting People With Low-Level Features and Bayesian Regression. *IEEE Trans. Image Process.* **2012**, *21*, 2160–2177. [CrossRef]
21. Huang, X.; Zou, Y.; Wang, Y. Cost-sensitive sparse linear regression for crowd counting with imbalanced training data. In Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, USA, 11–15 July 2016; pp. 1–6.
22. Lempitsky, V.; Zisserman, A. Learning To Count Objects in Images. In *Advances in Neural Information Processing Systems 23*; Lafferty, J.D., Williams, C.K.I., Shawe-Taylor, J., Zemel, R.S., Culotta, A., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2010; pp. 1324–1332.
23. Pham, V.; Kozakaya, T.; Yamaguchi, O.; Okada, R. COUNT Forest: CO-Voting Uncertain Number of Targets Using Random Forest for Crowd Density Estimation. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 13–16 December 2015; pp. 3253–3261.
24. Silveira Jacques Junior, J.C.; Musse, S.R.; Jung, C.R. Crowd Analysis Using Computer Vision Techniques. *IEEE Signal Process. Mag.* **2010**, *27*, 66–77. [CrossRef]
25. Li, T.; Chang, H.; Wang, M.; Ni, B.; Hong, R.; Yan, S. Crowded Scene Analysis: A Survey. *IEEE Trans. Circuits Syst. Video Technol.* **2015**, *25*, 367–386. [CrossRef]
26. Zitouni, M.S.; Bhaskar, H.; Dias, J.; Al-Mualla, M. Advances and trends in visual crowd analysis: A systematic survey and evaluation of crowd modelling techniques. *Neurocomputing* **2016**, *186*, 139–159. [CrossRef]
27. Loy, C.C.; Chen, K.; Gong, S.; Xiang, T. Crowd counting and profiling: Methodology and evaluation. In *Modeling, Simulation and Visual Analysis of Crowds*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 347–382.
28. Saleh, S.A.M.; Suandi, S.A.; Ibrahim, H. Recent survey on crowd density estimation and counting for visual surveillance. *Eng. Appl. Artif. Intell.* **2015**, *41*, 103–114. [CrossRef]

29. 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]
30. Gao, G.; Gao, J.; Liu, Q.; Wang, Q.; Wang, Y. CNN-based Density Estimation and Crowd Counting: A Survey. *arXiv* **2020**, arXiv:2003.12783.
31. Fu, M.; Xu, P.; Li, X.; Liu, Q.; Ye, M.; Zhu, C. Fast crowd density estimation with convolutional neural networks. *Eng. Appl. Artif. Intell.* **2015**, *43*, 81–88. [CrossRef]
32. Wang, C.; Zhang, H.; Yang, L.; Liu, S.; Cao, X. Deep People Counting in Extremely Dense Crowds. In Proceedings of the 23rd ACM International Conference on Multimedia; Brisbane, Australia, 26–30 October 2015; pp. 1299–1302. [CrossRef]
33. Sermanet, P.; Chintala, S.; LeCun, Y. Convolutional neural networks applied to house numbers digit classification. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012), Tsukuba, Japan, 11–15 November 2012; pp. 3288–3291.
34. Xue, Y.; Ray, N.; Hugh, J.; Bigras, G. Cell Counting by Regression Using Convolutional Neural Network. In *Computer Vision–ECCV 2016 Workshops*; Hua, G., Jégou, H., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 274–290.
35. 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 (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
36. Walach, E.; Wolf, L. Learning to Count with CNN Boosting. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016.
37. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Ma, Y. Single-image crowd counting via multi-column convolutional neural network. In Proceedings of the IEEEConference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 589–597.
38. Boominathan, L.; Kruthiventi, S.S.; Babu, R.V. Crowdnet: A deep convolutional network for dense crowd counting. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 640–644.
39. Babu Sam, D.; Surya, S.; Venkatesh Babu, R. Switching convolutional neural network for crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5744–5752.
40. Sam, D.B.; Babu, R.V. Top-down feedback for crowd counting convolutional neural network. *arXiv* **2018**, arXiv:1807.08881.
41. Liu, L.; Wang, H.; Li, G.; Ouyang, W.; Lin, L. Crowd counting using deep recurrent spatial-aware network. *arXiv* **2018**, arXiv:1807.00601.
42. 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–2 November 2019; pp. 6788–6797.
43. Hossain, M.; Hosseinzadeh, M.; Chanda, O.; Wang, Y. Crowd counting using scale-aware attention networks. In Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019; pp. 1280–1288.
44. 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.
45. Jiang, X.; Zhang, L.; Xu, M.; Zhang, T.; Lv, P.; Zhou, B.; Yang, X.; Pang, Y. Attention Scaling for Crowd Counting. Available online: https://openaccess.thecvf.com/content_CVPR_2020/papers/Jiang_Attention_Scaling_for_Crowd_Counting_CVPR_2020_paper.pdf (accessed on 16 September 2021).
46. 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, 18–22 June 2018; pp. 1091–1100.
47. Zhang, L.; Shi, M.; Chen, Q. Crowd Counting via Scale-Adaptive Convolutional Neural Network. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1113–1121.
48. 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.
49. Cao, X.; Wang, Z.; Zhao, Y.; Su, F. Scale Aggregation Network for Accurate and Efficient Crowd Counting. In *Computer Vision–ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 757–773.
50. Valloli, V.K.; Mehta, K. W-Net: Reinforced U-Net for Density Map Estimation. *arXiv* **2019**, arXiv:1903.11249.
51. 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, 16–20 June 2019; pp. 6133–6142.
52. Mnih, V.; Heess, N.; Graves, A.; Kavukcuoglu, K. Recurrent models of visual attention. Available online: https://proceedings.neurips.cc/paper/2014/file/09c6c3783b4a70054da74f2538ed47c6-Paper.pdf (accessed on 16 September 2021).
53. 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, 16–20 June 2019; pp. 3225–3234.
54. Zhang, C.; Li, H.; Wang, X.; Yang, X. Cross-scene crowd counting via deep convolutional neural networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 833–841.
55. Tian, Y.; Lei, Y.; Zhang, J.; Wang, J.Z. PaDNet: Pan-Density Crowd Counting. *IEEE Trans. Image Process.* **2020**, *29*, 2714–2727. [CrossRef]

56. Sajid, U.; Wang, G. Plug-and-play rescaling based crowd counting in static images. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Snowmass Village, CO, USA, 1–5 March, 2020; pp. 2287–2296.
57. Shang, C.; Ai, H.; Bai, B. End-to-end crowd counting via joint learning local and global count. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 1215–1219.
58. Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. Available online: https://proceedings.neurips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf (accessed on 16 September 2021).
59. Liu, J.; Gao, C.; Meng, D.; Hauptmann, A.G. DecideNet: Counting Varying Density Crowds through Attention Guided Detection and Density Estimation. *arXiv* **2018**, arXiv:1712.06679.
60. Liu, X.; van de Weijer, J.; Bagdanov, A.D. Leveraging Unlabeled Data for Crowd Counting by Learning to Rank. *arXiv* **2018**, arXiv:1803.03095.
61. Shi, M.; Yang, Z.; Xu, C.; Chen, Q. Revisiting Perspective Information for Efficient Crowd Counting. *arXiv* **2019** arXiv:1807.01989.
62. Liang, D.; Chen, X.; Xu, W.; Zhou, Y.; Bai, X. TransCrowd: Weakly-Supervised Crowd Counting with Transformer. *arXiv* **2021**, arXiv:2104.09116.
63. Sun, G.; Liu, Y.; Probst, T.; Paudel, D.P.; Popovic, N.; Gool, L.V. Boosting Crowd Counting with Transformers. *arXiv* **2021**, arXiv:2105.10926.
64. Gao, J.; Gong, M.; Li, X. Congested Crowd Instance Localization with Dilated Convolutional Swin Transformer. *arXiv* **2021**, arXiv:2108.00584.
65. Shi, Z.; Zhang, L.; Liu, Y.; Cao, X.; Ye, Y.; Cheng, M.M.; Zheng, G. Crowd Counting with Deep Negative Correlation Learning. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5382–5390. [CrossRef]
66. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 8–10 June 2015; pp. 1–9.
67. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
68. Sam, D.B.; Sajjan, N.N.; Babu, R.V. Divide and Grow: Capturing Huge Diversity in Crowd Images with Incrementally Growing CNN. *arXiv* **2018**, arXiv:1807.09993.
69. Chan, A.; Morrow, M.; Vasconcelos, N. Analysis of Crowded Scenes using Holistic Properties. Available online: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.214.8754&rep=rep1&type=pdf (accessed on 16 September 2021).
70. Loy, C.C.; Gong, S.; Xiang, T. From Semi-supervised to Transfer Counting of Crowds. Available online: https://personal.ie.cuhk.edu.hk/~ccloy/files/iccv_2013_crowd.pdf (accessed on 16 September 2021).
71. Lim, M.K.; Kok, V.J.; Loy, C.C.; Chan, C.S. Crowd Saliency Detection via Global Similarity Structure. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August 2014; pp. 3957–3962.
72. 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, 16–20 June 2019.
73. Kang, D.; Dhar, D.; Chan, A. Incorporating Side Information by Adaptive Convolution. In *Advances in Neural Information Processing Systems 30*; Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc.: Red Hook, NJ, USA, 2017; pp. 3867–3877.
74. 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/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019.
75. Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q. Vision meets drones: A challenge. *arXiv* **2018**, arXiv:1804.07437.
76. Bahmanyar, R.; Vig, E.; Reinartz, P. MRCNet: Crowd Counting and Density Map Estimation in Aerial and Ground Imagery. *arXiv* **2019**, arXiv:1909.12743.
77. Fang, Y.; Zhan, B.; Cai, W.; Gao, S.; Hu, B. Locality-constrained Spatial Transformer Network for Video Crowd Counting. *arXiv* **2019**, arXiv:1907.07911.
78. Liu, W.; Salzmann, M.; Fua, P. Context-Aware Crowd Counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019.
79. Zhang, Q.; Chan, A.B. Wide-Area Crowd Counting via Ground-Plane Density Maps and Multi-View Fusion CNNs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 8297–8306.
80. Hu, D.; Mou, L.; Wang, Q.; Gao, J.; Hua, Y.; Dou, D.; Zhu, X. Ambient Sound Helps: Audiovisual Crowd Counting in Extreme Conditions. *arXiv* **2020**, arXiv:2005.07097.
81. Zhu, P.; Sun, Y.; Wen, L.; Feng, Y.; Hu, Q. Drone Based RGBT Vehicle Detection and Counting: A Challenge. *arXiv* **2020**, arXiv:2003.02437.
82. Wang, Q.; Gao, J.; Lin, W.; Li, X. NWPU-Crowd: A Large-Scale Benchmark for Crowd Counting and Localization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**. [CrossRef]

83. 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–2 November 2019; pp. 1221–1231.
84. Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Máadeed, S.; Rajpoot, N.; Shah, M. Composition Loss for Counting, Density Map Estimation and Localization in Dense Crowds. *arXiv* **2018**, arXiv:1808.01050.
85. Liu, W.; W. Luo, D.L.; Gao, S. Future Frame Prediction for Anomaly Detection—A New Baseline. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
86. Tayara, H.; Gil Soo, K.; Chong, K.T. Vehicle Detection and Counting in High-Resolution Aerial Images Using Convolutional Regression Neural Network. *IEEE Access* **2018**, *6*, 2220–2230. [CrossRef]
87. Amato, G.; Ciampi, L.; Falchi, F.; Gennaro, C. Counting Vehicles with Deep Learning in Onboard UAV Imagery. In Proceedings of the 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain, 29 June–3 July 2019; pp. 1–6. [CrossRef]
88. Chen, J.; Xiu, S.; Chen, X.; Guo, H.; Xie, X. Flounder-Net: An efficient CNN for crowd counting by aerial photography. *Neurocomputing* **2021**, *420*, 82–89. [CrossRef]
89. Castellano, G.; Castiello, C.; Mencar, C.; Vessio, G. Crowd Counting from Unmanned Aerial Vehicles with Fully-Convolutional Neural Networks. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8. [CrossRef]
90. Wu, J.; Yang, G.; Yang, X.; Xu, B.; Han, L.; Zhu, Y. Automatic counting of in situ rice seedlings from UAV images based on a deep fully convolutional neural network. *Remote Sens.* **2019**, *11*, 691. [CrossRef]
91. Oh, S.; Chang, A.; Ashapure, A.; Jung, J.; Dube, N.; Maeda, M.; Gonzalez, D.; Landivar, J. Plant Counting of Cotton from UAS Imagery Using Deep Learning-Based Object Detection Framework. *Remote Sens.* **2020**, *12*, 2981. [CrossRef]
92. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
93. Kitano, B.T.; Mendes, C.C.; Geus, A.R.; Oliveira, H.C.; Souza, J.R. Corn plant counting using deep learning and UAV images. *IEEE Geosci. Remote Sens. Lett.* **2019**. [CrossRef]