electronics

MDPI

Article

Crowd Counting Using End-to-End Semantic Image Segmentation

Khalil Khan [1,*], Rehan Ullah Khan [2], Waleed Albattah [2], Durre Nayab [3], Ali Mustafa Qamar [4,5], Shabana Habib [2] and Muhammad Islam [6]

[1] Department of Information Technology and Computer Science, Pak-Austria Fachhochschule: Institute of Applied Sciences and Technology, Khyber Pakhtunkhwa 22620, Pakistan
[2] Department of Information Technology, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia; re.khan@qu.edu.sa (R.U.K.); w.albattah@qu.edu.sa (W.A.); s.habibullah@qu.edu.sa (S.H.)
[3] Faculty of Electrical and Computer Engineering, University of Engineering and Technology, Peshawar 25000, Pakistan; nayab.khan@uetpeshawar.edu.pk
[4] Department of Computer Science, College of Computer, Qassim University, Buraydah 51452, Saudi Arabia; al.khan@qu.edu.pk
[5] Department of Computing, School of Electrical Engineering and Computer Science, National University of Sciences and Technology (NUST), Islamabad 44100, Pakistan
[6] Department of Electrical Engineering, College of Engineering and Information Technology, Onaizah Colleges, Al-Qassim 51911, Saudi Arabia; m.islam@oc.edu.sa
* Correspondence: khalil.khan@fecid.paf-iast.edu.pk

check for updates

Citation: Khan, K.; Khan, R.U.; Albattah, W.; Nayab, D.; Qamar, A.M.; Habib, S.; Islam, M. Crowd Counting through End-to-End Semantic Scene Segmentation. *Electronics* **2021**, *10*, 1293. https://doi.org/10.3390/electronics10111293

Academic Editors: Jungong Han and Guiguang Ding

Received: 23 April 2021
Accepted: 26 May 2021
Published: 28 May 2021

Abstract: Crowd counting is an active research area within scene analysis. Over the last 20 years, researchers proposed various algorithms for crowd counting in real-time scenarios due to many applications in disaster management systems, public events, safety monitoring, and so on. In our paper, we proposed an end-to-end semantic segmentation framework for crowd counting in a dense crowded image. Our proposed framework was based on semantic scene segmentation using an optimized convolutional neural network. The framework successfully highlighted the foreground and suppressed the background part. The framework encoded the high-density maps through a guided attention mechanism system. We obtained crowd counting through integrating the density maps. Our proposed algorithm classified the crowd counting in each image into groups to adapt the variations occurring in crowd counting. Our algorithm overcame the scale variations of a crowded image through multi-scale features extracted from the images. We conducted experiments with four standard crowd-counting datasets, reporting better results as compared to previous results.

Keywords: artificial intelligence; crowd counting; crowd analysis; classification; deep learning; semantic scene segmentation

1. Introduction

The challenging and meaningful task of precisely estimating the number of objects and persons in an image has several applications in the Computer Vision (CV) domain. Among many applications, crowd counting is widely used, and one of the most practical usages of image object counting is that it can be exploited both for security and development purposes. Similarly, crowd counting and image object counting also help in areas such as surveys and traffic management. An accurate crowd count helps in emergency situations such as stampedes and fire events. Hence, considering these factors, many researchers are inclined to explore image-based object counting and its applications in various fields. Furthermore, much of the literature covers enormous contributions in the mentioned fields to analyze these contributions.

Whereas many data are present for crowd counting, the major bottleneck lies in the annotation process [1]. This bottleneck can be removed using crowd-sourcing, such as Amazon MTurk, or image-level annotations rather than bounding-box-focused ones. However, errors are possible in the case of relying on crowd-sourced annotations. This necessitates models that can deal with noisy labels.

Due to vast urbanization and an abrupt increase in the world population, substantial crowd gatherings such as religious and political events, parades, marathons, and concerts make crowd counting an indispensable service for managing and securing the crowd virtually and physically. Furthermore, crowd counting also helps in assessing the political significance of protests. It is not uncommon for different political parties to come up with different numbers for crowd gathering. Nevertheless, monitoring crowds from the surveillance videos is quite challenging because of the occlusion among people in the crowd. With the advent of effective deep learning algorithms and the techniques of Convolutional Neural Network (CNNs) in the computer vision field, the applications of objects and crowd counting have overwhelmingly improved. The structural and distribution patterns of all such applications are in some ways similar to each other, hence the improvement in one application implies the improvement in other related applications. This also implies that crowd counting methods can be extended to crowd analysis applications including flow analysis, density estimation, crowd monitoring, and son on.

To build a high level of cognitive ability in crowd-related applications, crowd counting is of vital value. Many research contributions have been made in object-counting and crowd-counting applications since many research communities are jointly working in this particular field. Among these research contributions, many are applied on images and videos of crowds in various domains [2–14], counting penguins in Antarctica using crowd-sourced annotated images [1], vehicle counting [3], leaf counting in rosette plants [15,16], cell microscopy [17–19], crowd analysis [20,21], pedestrian video surveillance [22], and surveys of the environment [23,24]. Similarly, crowd counting has major applications in population census [16,25–27], public event management [28–30], religious event management [31–36], and Closed-Circuit Television (CCTV) monitoring systems installed for public activity monitoring [17,37]. Finally, density management services are used in military activities for monitoring the number of soldiers, jets, drones, and vehicles to estimate the strength of the military, its positions, and areas of deployment [38–40].

In this paper, we proposed an end-to-end Semantic Scene Segmentation (SSS) framework, which uses the concept of semantic segmentation for crowd counting. To the best of our knowledge, our proposed framework is the first to use the idea of SSS for the task of crowd counting. Our proposed method highlighted the head region by suppressing the non-head part through a novel optimized loss function. This guided sort of mechanism pays comparatively more attention to the head part and encodes the specific refined density map. We also utilized the classification function, which automatically adapts the changes occurring in crowd counting. We performed extensive experiments on four standard datasets, reporting better results as compared to previous results.

2. Related Work

Crowd analysis, in general, and counting, in particular, are very mature areas of CV due to their diverse applications. Many excellent works have been reported by researchers to address these fields. Some recent survey papers [41,42] can be explored to learn about crowd analysis and counting.

There are generally four major classifications for the crowd-counting implementations. These are regression-based approaches, detection-based approaches, density-based approaches, and CNN-based approaches. These four methods are discussed in the following paragraphs;

- Detection-based approaches: Initially, most of the work on crowd counting was performed with detection-based approaches [43–47]. These approaches apply the head detector through a sliding window on an image. Recent methods such as R-CNN [48–50], You Only Look Once (YOLO) [51], and Single-Shot multibox Detector (SSD) [52] have been proposed and exploited, which attain high accuracy in sparse scenes, but these methods do not perform well in highly dense environments.
- Regression-based approaches: To target issues in detection-based methods, regression-based approaches [22,53,54] are proposed that can learn a mapping from the image

patch by extracting the global features [55] or local features [56]. The global features include the texture, edge, and gradient features, and the local features include Scale-Invariant Feature Transform (SIFT) [57], Local Binary Patterns (LBPs) [58], Histograms of Oriented Gradients (HOGs) [59], and Gray-Level Co-occurrence Matrices (GLCMs) [60]. To learn the mapping function for crowd counting, regression techniques [61] and Gaussian regression [62] are exploited. These algorithms solve occlusion and background clutter issues with detection-based approaches, but the spatial information is compromised. The regression-based techniques may overestimate the crowd in the presence of a sparse crowd.

- Density-based approaches: Similarly, the density-based methods make use of features such as pixels or regions. This helps to maintain the location information while avoiding the disadvantages of regression-based approaches. Lemptisky et al. [19] exploited a density-based approach with a linear mapping between local features and density maps. A nonlinear method, namely Random Forest Regression (RFR), was proposed to tackle the linear approach's issues by introducing the crowdedness before and training two different forests with it [63]. The method outperforms the linear method and also requires small memory for storing the forests. The issue with this approach is that the standard features are used to extract low-level information that cannot accurately be counted with a high-quality density map.
- CNN-based approaches: More research work is currently carried out with CNN algorithms because of their robust feature representation and improved density estimation. The CNN outperformed the traditional models to predict the density of crowds with improved performance in [18,64–66]. Recently, improved versions of CNNs, such as the Fully Convolution Network (FCN), have been proposed with an enhanced architecture, density estimation performance, and crowd counting. Besides FCN, many other CNN approaches have been proposed recently in the domain of density estimation and crowd counting [67].

  Sang et al. [11] developed an improved crowd counting approach based on the Scale-adaptive CNN (SaCNN). The CNN was used to obtain the crowd density map, which was further processed to find the approximate headcount. The proposed approach was tested on the Shanghai Tech dataset and worked well on sparse and dense scenes. More recently, Zhang et al. [68] used the CNN to count people on metro platforms. A dataset consisting of 627 images and 9243 annotated heads was also developed. The images were captured during peak and off-peak times during the weekdays and weekends. The authors used the first 13 layers of VGG-16. The results on standard datasets such as ShanghaiTech and UCF-QNRF showed a smaller MAE and MSE as compared to the state-of-the-art methods.

Accurate annotation of the ground truth is critical for crowd counting. Dot annotations, sometimes called land marking, put dots in the image to mark the objects of interest. This technique is used in crowd counting, face recognition, and posture alignment. However, it is not only time-consuming, but prone to errors as well. While a single annotator usually achieves the dot annotation, Arteta et al. [1] proposed an approach whereby crowd-sourcing was used to accomplish the annotation. Thirty-five-thousand volunteers were available for annotation, and as soon as an image received 20 annotations, it was removed from the system. As opposed to crowd-sourcing, no manual annotation is required in simulated data since we are fully aware of every object and its location. Lei et al. [69] developed a weak supervision model for crowd counting. Weaker annotations only require the total count of objects. They employ the multiple density map estimation technique and are able to obtain superior performance over already existing approaches.

Tong et al. [70] developed a simple deep learning-based model for crowd counting using a smart camera. The proposed approach was based on multi-task learning to perform density-level classification. Furthermore, the potential loss of detail was overcome using transposed convolutional layers. The proposed method was used to estimate the crowd density if the number of people was more than a threshold.

Songchenchen's work [71] aimed to find the head features using texture feature analysis and crowd image edge detection. The researcher also used a multi-column multi-feature CNN for crowd counting. The proposed methods outperformed the state-of-the-art methods on datasets such as Shanghai Tech, USCD, WorldExpo'10, and GCC. Later, he discussed the hardware implementation of the neural network architecture using an FPGA for crowd counting.

Zhang et al. [12] proposed a multi-column CNN to overcome large-scale changes in crowd images. A new dataset of 1198 images having more than 300,000 annotated heads was also developed. Another key benefit of this approach is that once trained on one dataset, their model can easily work with a new dataset. Nevertheless, this approach is severely limited by the number of columns, i.e., only three branches. Cao et al. [12] developed the Scale Aggression Network (SANet) for crowd counting based on the encoder–decoder model. Kang and Chan [5] used the image pyramid CNN for crowd counting while handling scale variations. Each scale of the image pyramid was fed to the FCN, which predicted a density map. Lastly, a 1 × 1 convolution combined the density maps at various scales.

A near-real-time crowd counting approach for both images and videos using a deep CNN was developed by Bhangale et al. [72]. The proposed model required only five seconds to perform a headcount from the provided video. The researchers concluded that the optimal resolution was 300 × 450 pixels. The experiments were conducted on Google Colab using a Tesla K80 GPU while employing the Shanghai Tech dataset. The results on both the dense, as well as sparse datasets were better than the multi-column CNN of Zhang et al. [2], the SANet of Cao et al. [12], and the image pyramid by Kang and Chan [5].

3. Proposed Method

As compared to Traditional Machine Learning (TML), recent Deep Learning Methods (DLMs) have shown better performance for various visual recognition tasks. We in the proposed work also employed a DLM for crowd counting. In this section of the paper, we discuss our proposed crowd-counting method using the concept of semantic image segmentation and the DLM.

The performance of a DLM relies on many factors, for example the kernel used, the number of convolutional layers, and the specific filters used in each layer. We used various combinations of Convolutional Layers (ConLs), and each layer was followed by Maximum Pooling layers (MaxPs). We also performed experiments regarding the size of the ConL to be used. Details of these parameters are presented in Tables 1 and 2.

Table 1. CNN layer information.

Layer	Stride	Size of Kernel	Feature Maps	Output Size
Input	–	–	–	250 × 250
ConL1	2	5 × 5	96	124 × 124
MaxP1	2	3 × 3	96	62 × 62
ConL2	2	5 × 5	256	30 × 30
MaxP2	2	3 × 3	256	15 × 15
ConL3	2	5 × 5	316	12 × 12
MaxP3	2	3 × 3	316	6 × 6
ConL4	2	5 × 5	512	4 × 4
MaxP4	2	3 × 3	512	2 × 2

We used ReLU as the function activator. As usual, a Deep Convolutional Neural Network (DCNN) has three layers, ConL, MaxP, and FCL. We also used the same setting. $N \times M \times C$ was the kernel with N representing the height and M the width of a specific filter C. Similarly, we represented the MaxP filters with $P \times Q$, where P is the height and Q the width of each filter. Lastly, FCL was the final layer, which performed the classification.

Table 2. Parameter setting for CNN training.

Parameters	Vales
Epochs	30
Batch size	125
Momentum	0.9
Base learning rate	10^{-4}

3.1. Model Learning

Our proposed network consisted of three parts, i.e., classification, SSS, and Density Estimation (DE). Our proposed crowd counting model is presented in Figure 1. To extract features from images, we used the Feature Extractor Framework (FEF). The stages in Figure 2 represent the main blocks of the deep feature learning architecture. Stage 1 handles the initial feature variations. There are many scaling variations in images due to different environmental circumstances. To overcome all these variation problems, we used four receptive fields. Each of these fields had sixteen filters. The output of Stage 1 is fed into the Stage 2 FCL layer. In Stage 2 and onwards, to extract multi-scale features, we used 2 × 2 pooling layers (maximum). Each ConL was followed by Rectified Linear Unit (ReLU). We placed Spatial Pyramid (SP) pooling layers between the ConL and Fully Connected Layer (FCL). We then fed the feature map, which was extracted from the input images to the SP pooling layers. The SP pooling layers produced output, which was given to the FCL of Stage 3. The shared module block in Stage 2 represents the SP module. The different stages, which are Stage 3 and Stage 4, take care of the feature extraction at different scales of the pyramid. Finally, in Stage 4, the FC layer (3) is used to extract the final features, which are then fed into the modules for the classification, SSS, and DE modules. The details of the different parameters can be seen in Tables 1 and 2.

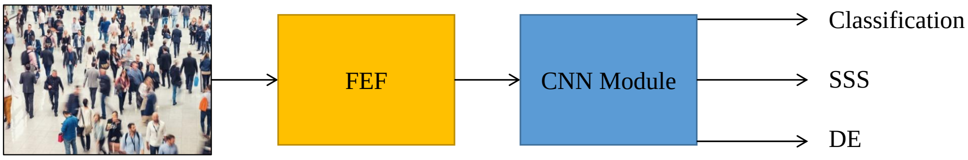

Figure 1. Proposed SSS-based crowd-counting model.

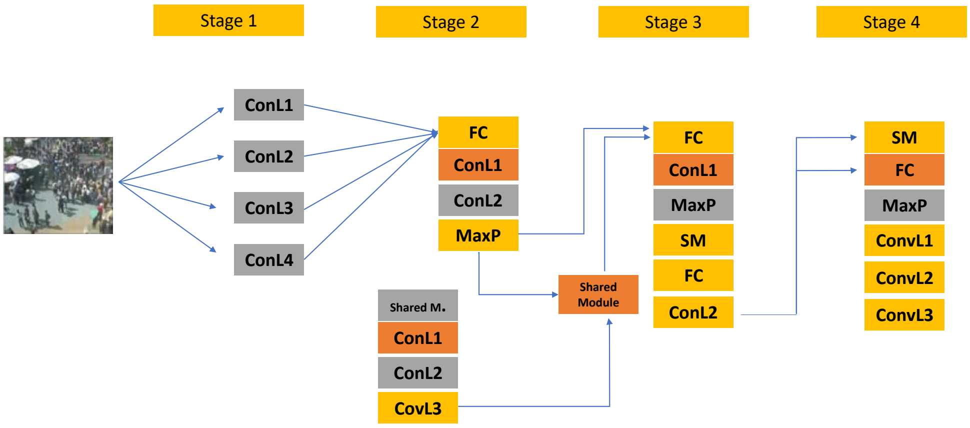

Figure 2. The proposed architecture of the feature extractor framework.

Our proposed classification part automatically learns the crowd's count distribution to adapt to the changes occurring in the crowd. We quantized the crowd count in each crowded image into several groups. We connected the FCL to the backbone at the end side. Both FCLs were followed by ReLU individually and had 64 and 6 neurons, respectively. In our case, six neurons showed the count groups. We did not change the input image's size to keep the distribution of the original crowd, as in the original image. We placed the SPP layers between the ConL and FCL. The feature map extracted from the crowd images was fed to the SPP layers, producing outputs provided to the FCL. In the classification phase, the counts from each database were classified into six groups, which adapted to changes in crowd counts.

In the SSS part, the training data along with the Ground Truth (GT) annotations are given to the framework. In the DE, we predicted the final density map with a kind of supervision from the GT density maps. We added a segmentation map and estimated the density maps, then fed the results to the ConL. The ReLU layers encoded the final density maps. We fixed the head regions' weight higher. Therefore, more attention was given to the head in the density estimation. We also introduced a loss based on the Dice coefficient in the segmentation part. Similarly, we introduced the Euclidean distance loss in the density estimation, which optimized the estimated density map more.

3.2. CNN Optimization

Our proposed framework included classification, segmentation, and crowd density estimation. To overcome and alleviate the overfitting problem, we used the methodology as followed in [73]. We optimized our framework by minimizing four loss functions, which also included supervision loss. In the DE, we utilized the Euclidean distance, which optimized the ED map in a better way. As a result, the obtained ED map can be given as shown in Equations (1) and (2).

$$Loss_{int} = \frac{1}{2M}\sum_{j=1}^{M} ||\hat{d}_j - D_j||_2^2 \tag{1}$$

$$Loss_{den} = \frac{1}{2M}\sum_{j=1}^{M} ||\hat{D}_j - D_j||_2^2 \tag{2}$$

In Equation (1), $\hat{d}$ represents the predicted density in the intermediate supervision process. Similarly, $\hat{D}_j$ shows the final ED, and D_j shows the GT density. M represents the pixel numbers in the GT density map.

We introduced a novel loss in the segmentation part. This loss was based on the Dice coefficient. In simple words, the Dice coefficient is twice the area of overlap between the predicted segmentation and the ground truth divided by the total number of pixels in two images. We optimized this loss to estimate the segmentation map for the head part. The range of the Dice coefficient is between 0 and 1. We quantized the crowd counts into six groups. For example, if the crowd counts in a densely crowded image ranged from 1 to 600, the images in the range from 1 to 100 would lie in the first class, and so on. We utilized the cross-entropy loss function, which is given in Equation (3).

$$Loss_{Xentropy} = -\frac{1}{N}\sum_{b=1}^{N}\sum_{c=1}^{M}(x_c^b)log(x^b) \tag{3}$$

where N is the total samples used for training and M represents the number of classes, which in our case was six. Similarly, x_c^b represents the GT class, and x^b shows the classification output. We represented the weighted loss function by the following equation:

$$W.L = Loss_{int} + Loss_{den} + \lambda Loss_{X-entropy} \tag{4}$$

where we fixed the value of λ as 0.02.

3.3. Data Annotation

Tools: For a machine learning task, GT data are created through annotation. The original data are in the form of audio, images, text, etc. A computer recognizes patterns similar in data not provided previously through a learning process from the GT data. These annotation categories vary, such as 3D cuboids, lines, bounding box, dot, and landmark annotation. In the crowd counting case, normally, dot annotation is the first step that creates the GT and is carried out with tools such as RectLabel and Label Me.

A tool for online annotation was also developed in JAVA and Python. This specific tool creates data for head points only. Two kinds of labels are supported by this tool: a point and a bounding box. The image is first zoomed-in this method, and then, the head part is labeled with some desired scaling factor. Then, the image is divided into patches having a size of 16 × 16. This specific size allows annotators to make the GT under various scales times the original crowded image size. Annotation with this tool is comparatively easy, and the quality is also good. For details, readers are requested to explore [54].

Pointwise annotation: In this way, annotation is divided into two stages. In the first stage, labeling is performed, followed by the refinement of the previous labeling. In the first step, annotators perform the labeling process. However, this method of creating the GT is a laborious and time-consuming task. After creating the GT, additional individuals perform the preliminary annotation, which brings a kind of refinement to the whole labeling process.

Annotation at the box level: This is a more time-consuming task as annotation is performed in three steps here. Initially, ten to twenty percent of the points are typically selected in an image for drawing a bounding box. Secondly, for those points having no box, a linear regression method is adapted to obtain its nearest box along with the size. Third, a manual kind of refining of the estimated box label is performed.

In summary, GT labels are produced through a manual process. This labeling is performed without any automatic labeling tool. This labeling depends on a subjective perception of a person who is involved in the labeling task. Hence, giving an accurate GT label in this scenario is complex, and chances for error exist.

Unlike these methods, we adopted a different strategy for data annotation and creating the GT. Since most of these methods are manual works that involve laborious work and time-consuming efforts, we adopted a different approach for data annotation. For all training images, a point located at the center of each head was provided. We encoded the GT density map by employing a Gaussian kernel known as the normalized Gaussian on every point p, which is:

$$G(x,y) = \sum_{p \in S} M(p;\mu,\sigma) \tag{5}$$

where symbol (x,y) shows the location of a specific pixel in an image and S represents a series of annotated points. Similarly, $M(p;\mu,\sigma)$ is the normalized Gaussian Kernel having mean value 0 and variance 4. We used a window size of 15 × 15. We used this method to generate GT density maps. As it is impossible to label GT data for segmentation for larger datasets manually, we proposed an effective way for the GT segmentation map to have the same background and foreground as the GT density maps.

4. Results and Discussion

4.1. Experimental Setup

We performed our experiments with an Intel i7 CPU having 16 G RAM. The graphical processing unit used was the NVIDIA 840 M graphics card. All the tests were performed with Google TensorFlow and Keras in the Python environment. We trained the model for 30 epochs while keeping the batch training size as 125. We kept this setting for all four datasets and their experiments.

4.2. Databases

We evaluated the performance of our proposed crowd-counting framework with four datasets including NWPU-crowd, UCF-QNRF, Shanghai Tech, and World Expo10. A summary of the crowd counting dataset is presented in Table 3. We provide the details about these datasets in the following paragraphs.

Table 3. Major Crowd-Monitoring System (CMS) datasets.

Database	Year	Task	Number of Images	Source Obtained
NWPU-Crowd [47]	2020	crowd counting and localization	5109	surveillance and Internet
UCF-QNRF [74]	2018	counting in a crowd	1525	surveillance
Shanghai Tech [2]	2016	cross-scene crowd counting	482	surveillance cameras and Internet
World Expo'10 [66]	2015	counting in a crowd	3980	surveillance

- NWPU-crowd [47]: Deep learning-based models need large-scale data for training and testing phases. Most of the available datasets do not meet the requirements of the deep learning-based methods. NWPU-Crowd is a large-scale dataset that can be utilized for crowd counting. The dataset consists of 5 k images in which almost 2,133,375 heads are annotated. The density range of NWPU-Crowd is large, and diverse illumination conditions were considered. Both Internet and self-shooting data were used for dataset creation. For data collection, a very diverse data collection strategy was adopted; for example, malls, stations, plazas, and resorts were used.
- UCF-QNRF [74]: This is the latest database introduced for various crowd analysis tasks including crowd counting. The dataset has 1535 images with massive variation in density. Images in UCF-QNRF have a higher resolution of × 300 to 9000 × 6000). Images in UCF-QNRF were collected from Hajj footage, web search, and Flicker. Annotation for the data is also provided. Lighting variations, diverse density conditions, and changes in view points are the main characteristics of the dataset. Images were collected from unconstrained conditions with sources such as buildings, roads, sky, and vegetation. Due all the mentioned conditions, the dataset is challenging and fit for deep learning-based models.
- Shanghai Tech [2]: The dataset has the particular feature of large-scale counting. It has 1198 images and 330,165 annotated heads are. The dataset consists of two parts, where Part A has 482 and B 716 images. Part A images were collected from web sources and B images from the streets of Shanghai. The authors defined various combination sets for experiments. Most of the literature uses 300 images for training and the remaining 182 for testing for Part A. Similarly, four-hundred images of Part B are used for training and 316 for the testing phase. Diverse scenes and highly varying density conditions were included in the data collection to make the database challenging.
- World Expo [66]: All data for World Expo were collected from 108 surveillance cameras installed at various places. The dataset is suited for cross-scene management scenarios and is efficient. It has 3980 frames, which have a size of 576 × 720 each. There are 199,923 labeled pedestrians in the dataset. To ensure diversity in the collected data, disjoint bird views were used by the creators of the dataset. The reported literature divides the training data as one-thousand one-hundred twenty-seven videos with a length of one minute. Due to having fewer data as compared to the State-Of-The-Art (SOTA) dataset, the database is not suitable for performing experiments with deep learning-based models.

  Some sample images from these dataset are shown in Figures 3–6. The datasets are summarized in Table 3.

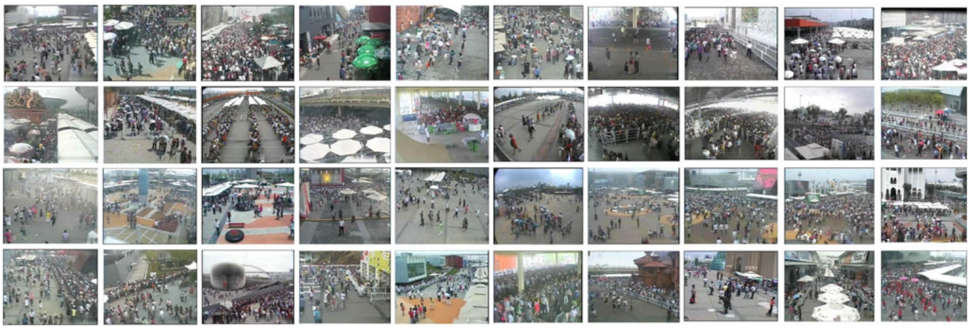

Figure 3. Images from the World Expo [66] dataset. Adapted from ref. [66].

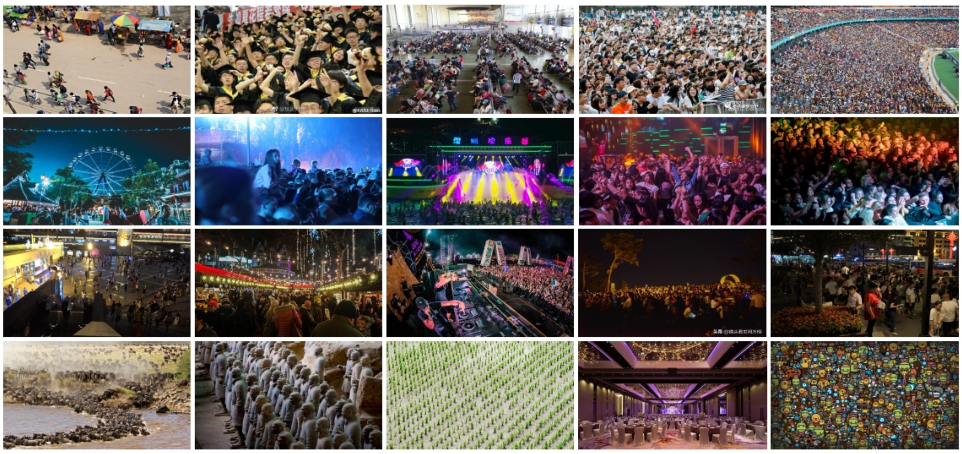

Figure 4. Images from the NWPU-Crowd [47] counting dataset. Adapted from ref. [47].

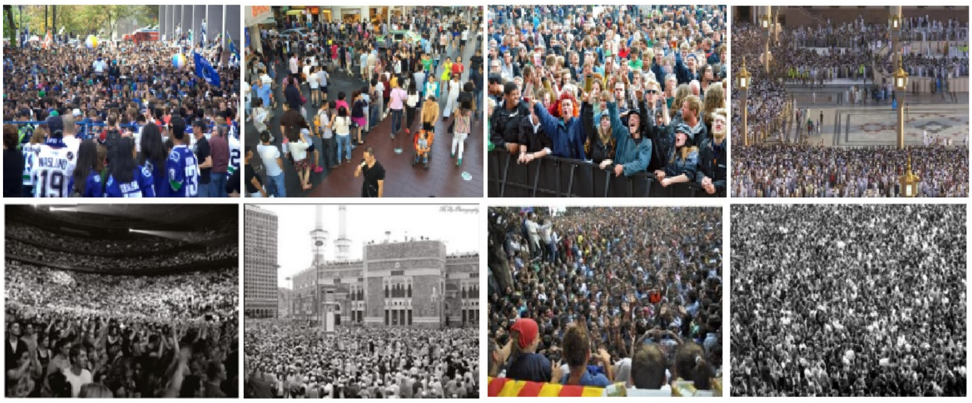

Figure 5. Sample images from the UCF-QNRF [74] dataset. Adapted from ref. [74].

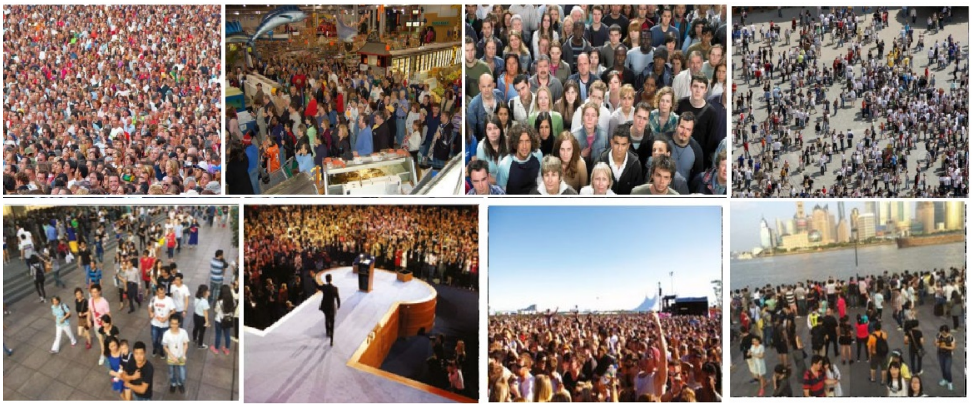

Figure 6. Images from the Shanghai Tech [2] dataset. Adapted from ref. [2].

4.3. Quantification of Tasks

We represented the count estimation as i by C_i. C_i is a single metric that does not provide information, in particular about the distribution of people in an image. However, this metric helps predict the size of a crowd, which may span many kilometers. An idea was presented in [75], which divides the occupied area of the crowd into further smaller sections. The average number of participants in that particular area is then further estimated. The method also estimates the mean density of the covered area. However, counting for several crowded images at many locations is comparatively difficult. Due to a much more complicated nature, two additional metrics, Mean Absolute Error (MAE) and Mean Squared Error (MSE), are frequently used by researchers. We also reported our work with these two measures. Mathematically, these measures can be defined as:

$$MAE = \frac{1}{\mathrm{M}} \sum_{j=1}^{M} |Y_j - Y_j^{'}| \tag{6}$$

$$MSE = \sqrt{\frac{1}{\mathrm{M}} \sum_{j=1}^{M} |Y_j - Y_j^{'}|^2} \tag{7}$$

In Equations (6) and (7), M represents the test samples, Y_j the ground truth count, and $Y_j^{'}$ the count estimated for the jth sample.

We noticed that crowd localization is a less explored area. Similarly, researchers have not yet firmly established evaluation metrics for localization problems. We observed that [54] was the only research work that proposed a one-to-one matching. However, this idea [54] also leads to some optimistic issues. Similarly, the authors defined no penalization method if overdetection cases occur. We noticed that the lastly discussed method has failed to be acknowledged widely. We evaluated our method with precision, recall, and the F-measure. These evaluation metrics are further defined as True Positive (*TP*), False Positive (*FP*), and False Negative (*FN*). The *TP* is the number of heads that are correctly detected. Similarly, the *FN* is the number of heads that are incorrectly detected as non-heads, whereas the FP is the number incorrectly detected as heads. Mathematically we can write precision, recall, and the F-measure as:

$$Precession = \frac{(TP)}{(TP) + (FP)} \tag{8}$$

$$Recall = \frac{(TP)}{(TP) + (FN)} \tag{9}$$

$$\text{F-measure} = 2 \times \frac{(Precision) \times (Recall)}{(Precision) + (Recall)} \tag{10}$$

4.4. Comparative Analysis

The reported results with the proposed method and its comparison are presented in Table 4. From Table 4, it is clear that we have better results in most of the cases as compared to the previous results. We present a summary of the concluding remarks in the paragraphs follows:

- We reported our results in the form of precision, recall, and the F-measure. We also used other evaluation metrics, the MAE and MSE. All these values are reported in Tables 4 and 5. From both tables, it is clear that we had much better results as compared to previous results.
- In the last ten years, crowd counting has been explored by researchers significantly. A summary of the results can be seen in Table 4. Researchers have introduced several datasets that address the crowd-counting problem. We noticed that less emphasis has been given to crowd behavior analysis and localization of a crowd. Due to many

more applications, crowd counting has been targeted in a better way than other crowd analysis tasks. Due to the diversity of applications, our work also focused on crowd counting.

- The labeling process for creating GT annotation data was performed by a manual process. We observed that this was a time-consuming process, and also, more chances for error existed. Such a labeling process entirely depends on the subjective perception of the person involved in labeling. Compared to this manual labeling, automatic labeling is a comparatively better option, but it is still not a mature case to be used effectively for research. We, in our work, also introduced an automatic labeling process for creating GT data.
- As discussed earlier, crowd counting is an active area of research due to diverse applications. Table 4 shows a detailed summary of the research conducted on SOTA datasets. We reported all the metrics, including the MAE, MSE, precision, recall, and F-measure, from the original research papers. It is very clear from the Table that all these metrics have improved with the passage of time. Much more improvement is brought in particular with the introduction of improved deep learning methods.
- Some research papers reported that TML methods showed better performance as compared to the DLM. Even though through this comparison, it was not claimed that hand-crafted features are better than deep learning-based methods. We argue that a better understanding of the deep learning-based methods is needed for the crowd-counting task. For example, a limited data scenario is a major drawback faced by the deep learning-based methods. We noticed that the performance of the traditional machine learning methods is acceptable with data that are collected in controlled environmental conditions. However, when these TML methods were applied to data collected in-the-wild, a drop in performance by a huge amount was noticed. On the other hand, the DLM extracts a higher level of abstraction from the data. As a result, the DLM outperforms traditional methods. The need for feature engineering is reduced with deep learning-based methods. It is also worth noting that the DLM is facing some concerns from the research community. For instance, the DLM and its applications are complicated procedures that require various inputs from the practitioner’s end. Most of the researchers rely on a trial and error strategy. Hence, these methods are time consuming and more engineered. However, it must be noted that the DLM is the only definitive choice for the crowd-counting task.
- As can be seen from Table 4, most of the DLMs for crowd counting use DCNNs. However, most of these DCNN-based methods employ the pooling layer, which results in comparatively low resolution and some feature loss as well. It is clear that the deeper layers extract some high-level information, whereas comparatively shallower layers somehow extract low-level information and features, which include spatial information. We suggest that both deeper and shallower layer information showed be combined for better results. More reasonable accuracy will be reported with this, and the count error will also be reduced.
- Crowd counting is an active area of research in computer vision. Tremendous progress has been reported in the last couple of years. From the reported results to date, it is evident that most of the metrics such as the MAE, MSE, and F-measure have improved. However, noting the trend of the computer vision developments in various application scenarios with the DLM, it is clear that crowd counting is not a mature research area. As the training phase in the DLM is facing problems due to limited data, an option for researchers to explore is knowledge transfer [76,77].

Table 4. CMS performance in the form of average precision, recall, and the F1-measure metrics.

Database	Year	Method	Precision	Recall	F1-Measure
NWPU-crowd	**2021**	**proposed method**	**96.4**	**62.3**	**57.8**
	2019	Liu et al. [78]	66.6	54.3	59.8
	2019	Gao et al. [79]	55.8	49.6	52.5
	2017	Hu et al. [80]	52.9	61.1	56.7
UCF_QNRF	**2021**	**proposed method**	**84.5**	**76.42**	**80.25**
	2020	Xue et al. [81]	82.4	78.3	80.00
	2019	Liu et al. [78]	81.5	71.1	75.0
	2018	Liu et al. [78]	59.3	63.0	61.09
	2018	Shen et al. [82]	75.6	59.7	66.71
	2018	Idrees et al. [54]	75.8	63.5	69.10
	2016	Zhang et al. [2]	71.0	72.4	71.69
	2016	He et al. [83]	61.6	66.9	64.14
	2016	Huang et al. [84]	70.1	58.1	63.53
	2015	Badrinarayanan et al. [85]	71.8	62.9	67.05
	2015	Zhang et al. [66]	78.1	65.1	71.17
Shanghai Tech. A	**2021**	**proposed method**	**88.2**	**78.6**	**83.12**
	2020	Xue et al. [81]	87.3	79.2	82.05
	2019	Liu et al. [78]	86.5	69.7	77.12
	2018	Idrees et al. [54]	79.0	72.3	75.51
	2018	Liu et al. [78]	82.2	73.3	77.49
	2018	Shen et al. [82]	79.2	82.2	80.67
	2016	Zhang et al. [2]	76.5	81.7	78.92
	2015	Zhang et al. [66]	81.9	77.9	79.84
Shanghai Tech. B	**2021**	**proposed method**	**87.7**	**81.1**	**84.27**
	2020	Xue et al. [81]	86.7	80.5	83.80
	2019	Liu et al. [78]	79.1	60.1	68.30
	2019	Zhang et al. [2]	82.4	76.0	79.07
	2019	Idrees et al. [54]	76.8	78.0	77.39
	2019	Liu et al. [78]	78.1	73.9	75.94
	2018	Liu et al. [78]	75.4	79.3	77.30
	2018	Shen et al. [82]	80.2	78.8	77.34
	2015	Zhang et al. [66]	84.1	75.8	79.73
World Expo	**2021**	**proposed method**	**83.3**	**84.5**	**83.89**
	2020	Xue et al. [81]	82.0	81.5	81.74
	2019	Idrees et al. [54]	72.4	78.3	75.23
	2019	Liu et al. [78]	73.7	79.6	76.52
	2019	Liu et al. [78]	71.6	75.4	73.45
	2019	Zhang et al. [2]	80.9	77.5	79.16
	2018	Liu et al. [78]	73.8	78.2	75.93
	2018	Shen et al. [82]	68.5	81.2	74.31
	2015	Zhang et al. [66]	79.5	73.1	76.16

Table 5. CMS performance in the form of the MAE and MSE metrics.

Database	Year	Method	MAE	MSE
NWPU-crowd	**2021**	**proposed method**	**78.2**	**301.9**
	2021	Abousamra et al. [86]	107.8	438.5
	2021	Liang et al. [87]	86.0	312.5
	2019	Gao et al. [79]	127.3	439.9
	2019	Liu et al. [78]	151.5	634.7
	2017	Peiyun et al. [88]	272.4	764.9
	2015	Ren et al. [49]	414.2	1063.7
UCF_QNRF	**2021**	**proposed method**	**96.4**	**162.2**
	2019	Liu et al. [89]	107	183
	2019	Jiang et al. [90]	113	188
	2019	Wan et al. [91]	101	176
	2018	Li et al. [13]	113.2	189.4
	2017	Zhang et al. [92]	111	190
Shanghai Tech. A	**2021**	**proposed method**	**56.7**	**93.1**
	2019	Guo et al. [93]	64.2	99.9
	2019	Liu et al. [89]	62.3	100.0
	2019	Jiang et al. [90]	64.2	109.1
	2019	Wan et al. [94]	64.7	97.1
	2019	Zhang et al. [92]	59.4	102.0
	2018	Li et al. [13]	68.2	115.0
	2018	Cao et al. [12]	67.0	104.5
	2017	Zhang et al. [2]	110.2	173.2
Shanghai Tech. B	**2021**	**proposed method**	**7.4**	**10.2**
	2019	Guo et al. [93]	8.8	13.5
	2019	Lie et al. [89]	7.8	12.2
	2019	Jiang et al. [90]	8.2	12.8
	2019	Wan et al. [94]	8.1	13.6
	2019	Zhang et al. [92]	7.9	12.9
	2017	Zhang et al. [2]	26.4	41.3
	2018	Li et al. [13]	10.6	16.0
	2018	Cao et al. [12]	8.4	13.6
World Expo	**2021**	**proposed method**	**8.3**	–
	2020	Gao et al. [95]	21.6	–
	2019	Gao et al. [79]	17.4	–
	2019	Gao et al. [79]	10.8	–
	2019	Wang et al. [91]	26.3	–
	2017	Zhu et al. [96]	32.4	–

5. Summary and Concluding Remarks

Crowd counting is an essential task in crowd image analysis due to the very diverse applications. Crowd counting is challenging when the proposed algorithm is particularly exposed to data collected in diverse conditions and in-the-wild. However, researchers from the CV community have shown significant progress, in particular in the last 10 years. We proposed an end-to-end SSS-based crowded-counting algorithm. We introduced the idea of using semantic segmentation of an image showing a crowd. The proposed method also adapted the changes occurring in crowd counts through optimization of the classification part. We introduced a novel loss function, which improved the performance of the proposed method on SOTA datasets. We validated our method on four standard datasets including NWPU_Crowd, Shanghai Tech, UCF_CC, and World Expo and obtained better results as compared to the previous ones. At the end, we are expecting some excellent evaluations using more optimized deep learning-based techniques for crowd counting.

Author Contributions: Conceptualization, K.K. and R.U.K.; methodology, K.K.; software, R.U.K.; validation, K.K. and W.A.; formal analysis, K.K.; investigation, R.U.K.; resources, M.I.; data curation, D.N.; writing—original draft preparation, K.K.; writing—review and editing, K.K., S.H., M.I. and A.M.Q.; visualization, S.H.; supervision, M.I.; project administration, K.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Education in Saudi Arabia through project number QURDO001.

Acknowledgments: The authors extend their appreciation to the Deputyship for Research and Innovation,Ministry of Education in Saudi Arabia for funding this research work through the project number QURDO001.Project title: Intelligent Real-Time Crowd Monitoring System Using Unmanned Aerial Vehicle (UAV) Video andGlobal Positioning Systems (GPS) Data.

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

References

1. Arteta, C.; Lempitsky, V.; Zisserman, A. Counting in the wild. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 483–498.
2. Zhang, Y.; Zhou, D.; Chen, S.; Gao, S.; Yi, M. 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.
3. Onoro-Rubio, D.; Lopez-Sastre, R.J. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 615–629.
4. Boominathan, L.; Kruthiventi, S.S.; Babu, R.V. Crowdnet: A deep convolutional network for dense crowd counting. In *ACM MM*; ACM: New York, NY, USA, 2016; pp. 640–644.
5. Kang, D.; Chan, A. Crowd counting by adaptively fusing predictions from an image pyramid. *arXiv* **2018**, arXiv:1805.06115.
6. Sam, D.B.; Surya, S.; Babu, R.V. Switching convolutional neural network for crowd counting. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4031–4039.
7. Sindagi, V.A.; Patel, V.M. Generating high-quality crowd density maps using contextual pyramid CNNs. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1861–1870.
8. Liu, J.; Gao, C.; Meng, D.; Hauptmann, A.G. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5197–5206.
9. 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.
10. 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.
11. Sang, J.; Wu, W.; Luo, H.; Xiang, H.; Zhang, Q.; Hu, H.; Xia, X. Improved crowd counting method based on scale-adaptive convolutional neural network. *IEEE Access* **2019**, *1*, 24411–24419. [CrossRef]
12. Cao, X.; Wang, Z.; Zhao, Y.; Su, F. Scale aggregation network for accurate and efficient crowd counting. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 734–750.
13. Li, Y.; Zhang, X.; Chen, D. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1091–1100.
14. Varior, R.R.; Shuai, B.; Tighe, J.; Modolo, D. Scale-aware attention network for crowd counting. In Proceedings of the Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
15. Aich, S.; Stavness, I. Leaf counting with deep convolutional and deconvolutional networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), Venice, Italy, 22–29 October 2017; pp. 2080–2089.
16. Giuffrida, M.V.; Minervini, M.; Tsaftaris, S.A. Learning to count leaves in rosette plants. In *Proceedings of the Computer Vision Problems in Plant Phenotyping (CVPPP)*; BMVC: York, UK, 2015; pp. 1.1–1.13.
17. Wang, Y.; Zou, Y. Fast visual object counting via example-based density estimation. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3653–3657.
18. Walach, E.; Wolf, L. Learning to count with cnn boosting. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 660–676.
19. Lempitsky, V.; Zisserman, A. Learning to count objects in images. In Proceedings of the Neural Information Processing Systems, Hyatt Regency, Vancouver, BC, Canada, 6–11 December 2010; pp. 1324–1332.
20. Shao, J.; Kang, K.; Change Loy, C.; Wang, X. Deeply learned attributes for crowded scene understanding. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 4657–4666.

21. Zhou, B.; Wang, X.; Tang, X. Understanding collective crowd behaviors: Learning a mixture model of dynamic pedestrian-agents. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2871–2878.
22. Chan, A.B.; Liang, Z.-S.J.; Vasconcelos, N. Privacy preserving crowd monitoring: Counting people without people models or tracking. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–7.
23. French, G.; Fisher, M.; Mackiewicz, M.; Needle, C. Convolutional neural networks for counting fish in fisheries surveillance video. In Proceedings of the Workshop: Machine Vision of Animals and their Behaviour MVAB, Swansea, UK, 10 September 2015; pp. 1–7.
24. Zhan, B.; Monekosso, D.N.; Remagnino, P.; Velastin, S.A.; Xu, L.-Q. Crowd analysis: A survey. *MVA* **2008**, *19*, 345–357. [CrossRef]
25. Fiaschi, L.; Köthe, U.; Nair, R.; Hamprecht, F.A. Learning to count with regression forest and structured labels. In Proceedings of the 2012 21st International Conference on Pattern Recognition (ICPR), Tsukuba Science City, Japan, 11–15 November 2012; pp. 2685–2688.
26. Rabaud, V.; Belongie, S. Counting crowded moving objects. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 1, pp. 705–711.
27. Bharti, Y.; Saharan, R.; Saxena, A. Counting the Number of People in Crowd as a Part of Automatic Crowd Monitoring: A Combined Approach. In *Information and Communication Technology for Intelligent Systems*; Springer: Singapore, 2019; pp. 545–552.
28. Boulos, M.N.K.; Resch, B.; Crowley, D.N.; Breslin, J.G.; Sohn, G.; Burtner, R.; Pike, W.A.; Eduardo Jezierski, E.; Chuang, K.-Y.S. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: Trends, OGC standards and application examples. *Int. J. Health Geogr.* **2011**, *10*, 1–29.
29. Lv, Y.; Duan, Y.; Kang, W.; Li, Z.; Wang, F.-Y. Traffic flow prediction with big data: A deep learning approach. *IEEE Trans. Intell. Transp. Syst.* **2014**, *16*, 865–873. [CrossRef]
30. Sadeghian, A.; Alahi, A.; Savarese, S. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 300–311.
31. Barr, J.R.; Bowyer, K.W.; Flynn, P.J. The effectiveness of face detection algorithms in unconstrained crowd scenes. In Proceedings of the 2014 IEEE Winter Conference on Applications of Computer Vision (WACV), Steamboat Springs, CO, USA, 24–26 March 2014; pp. 1020–1027.
32. Ng, H.W.; Nguyen, V.D.; Vonikakis, V.; Winkler, S. Deep learning for emotion recognition on small datasets using transfer learning. In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, WA, USA, 9–13 November 2015; pp. 443–449.
33. Chackravarthy, S.; Schmitt, S.; Yang, L. Intelligent Crime Anomaly Detection in Smart Cities Using Deep Learning. In Proceedings of the 2018 IEEE 4th International Conference on Collaboration and Internet Computing (CIC), Philadelphia, PA, USA, 18–20 October 2018; pp. 399–404.
34. Zainuddin, Z.; Thinakaran, K.; Shuaib, M. Simulation of the Pedestrian Flow in the Tawaf Area Using the Social Force Model. *World Acad. Sci. Eng. Technol. Int. J. Math. Comput. Sci.* **2010**, *4*, 789–794.
35. Zainuddin, Z.; Thinakaran, K.; Abu-Sulyman, I.M. Simulating the Circumbulation of the Ka'aba using SimWalk. *Eur. J. Sci. Res.* **2009**, *38*, 454–464.
36. Al-Ahmadi, H.M.; Alhalabi, W.S.; Malkawi, R.H.; Reza, I. Statistical analysis of the crowd dynamics in Al-Masjid Al-Nabawi in the city of Medina, Saudi Arabia. *Int. J. Crowd Sci.* **2018**, *2*, 64–73. [CrossRef]
37. Zhou, B.; Tang, X.; Wang, X. Learning collective crowd behaviors with dynamic pedestrian-agents. *Int. J. Comput. Vis.* **2015**, *111*, 50–68. [CrossRef]
38. Perez, H.; Hernandez, B.; Rudomin, I.; Ayguade, E. Task-based crowd simulation for heterogeneous architectures. In *Innovative Research and Applications in Next-Generation High Performance Computing*; IGI Global: Harrisburg, PA, USA, 2016; pp. 194–219.
39. Martani, C.; Stent, S.; Acikgoz, S.; Soga, K.; Bain, D.; Jin, Y. Pedestrian monitoring techniques for crowd-flow prediction. *Proc. Inst. Civ. Eng.-Smart Infrastruct. Constr.* **2017**, *2*, 17–27. [CrossRef]
40. Khouj, M.; López, C.; Sarkaria, S.; Marti, J. Disaster management in real time simulation using machine learning. In Proceedings of the 24th Canadian Conference on Electrical and Computer Engineering (CCECE), Niagara Falls, ON, Canada, 8–11 May 2011; pp. 1507–1510.
41. Khan, K.; Waleed, A.; Rehan, U.K.; Ali Mustafa, Q.; Durre, N. Advances and trends in real time visual crowd analysis. *Sensors* **2020**, *20*, 5073. [CrossRef]
42. Khan, A.; Shah, J.; Kadir, K.; Albattah, W.; Khan, F. Crowd Monitoring and Localization Using Deep Convolutional Neural Network: A Review. *Appl. Sci.* **2020**, *10*, 4781. [CrossRef]
43. 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.
44. 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 2008 19th International Conference on Pattern Recognition, Tampa, FL, USA, 8–11 December 2008; pp. 1–4.

45. Leibe, B.; Seemann, E.; Schiele, B. Pedestrian detection in crowded scenes. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 878–885.
46. Enzweiler, M.; Gavrila, D.M. Monocular pedestrian detection: Survey and experiments. *TPAMI* **2009**, *31*, 2179–2195. [CrossRef]
47. Wang, Q.; Gao, J.; Lin, W.; Li, X. NWPU-crowd: A large-scale benchmark for crowd counting. *arXiv* **2020**, arXiv:2001.03360.
48. Girshick, R. Fast r-cnn. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
49. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards realtime object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 91–99. [CrossRef] [PubMed]
50. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
51. Redmon, J.; Divvala, S.; Girshick, R.; 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.
52. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
53. Chan, A.B.; Vasconcelos, N. Bayesian poisson regression for crowd counting. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 545–551.
54. Idrees, H.; Saleemi, I.; Seibert, C.; Shah, M. Multi-source Multi-scale counting in extremely dense crowd images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2547–2554.
55. Chen, K.; Loy, C.C.; Gong, S.; Xiang, T. Feature mining for localised crowd counting. *BMVC* **2012**, *1*, 3.
56. Ryan, D.; Denman, S.; Fookes, C.; Sridharan, S. Crowd counting using multiple local features. In Proceedings of the 2009 Digital Image Computing: Techniques and Applications, Melbourne, VIC, Australia, 1–3 December 2009; pp. 81–88.
57. Low, D.G. Object recognition from local scale-invariant features. *ICCV* **1999**, *99*, 1150–1157.
58. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns. In *Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2000*; Springer: Berlin/Heidelberg, Germany, 2000; pp. 404–420.
59. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. *IEEE Comput. Soc.* **2005**, *1*, 886–893.
60. Haralick, R.M.; Shanmugam, K. Textural features for image classification. *TSMC* **1973**, *6*, 610–621. [CrossRef]
61. Paragios, N.; Ramesh, V. A MRF-based approach for real-time subway monitoring. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, 8–14 December 2001; Volume 1, pp. 1–1034.
62. Tian, Y.; Sigal, L.; Badino, H.; De la Torre, F.; Liu, Y. Latent gaussian mixture regression for human pose estimation. In *ACCV*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 679–690.
63. Pham, V.-Q.; 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, 7–13 December 2015; pp. 3253–3261.
64. Wang, C.; Zhang, H.; Yang, L.; Liu, S.; Cao, X. Deep people counting in extremely dense crowds. In *ACM MM*; ACM: New York, NY, USA, 2015; pp. 1299–1302.
65. Fu, M.; Xu, P.; Li, X.; Liu, Q.; Ye, M.; Zhu, C. Fast crowd density estimation with convolutional neural networks. *EAAI* **2015**, *43*, 81–88. [CrossRef]
66. Zhang, C.; Li, H.; Wang, X.; Yang, X. Cross-scene crowd counting via deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015.
67. Gao, G.; Gao, J.; Liu, Q.; Wang, Q.; Wang, Y. Cnn-based density estimation and crowd counting: A survey. *arXiv* **2020**, arXiv:2003.12783.
68. Zhang, J.; Liu, J.; Wang, Z. Convolutional Neural Network for Crowd Counting on Metro Platforms. *Symmetry* **2021**, *13*, 703. [CrossRef]
69. Lei, Y.; Liu, Y.; Zhang, P.; Liu, L. Towards using count-level weak supervision for crowd counting. *Pattern Recognit.* **2021**, *109*, 107616. [CrossRef]
70. 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]
71. Songchenchen, G. Real-Time Implementation of Counting People in a Crowd on the Embedded Reconfigurable Architecture on the Unmanned Aerial Vehicle. Image Processing [eess.IV]. Ph.D. Thesis, Université Bourgogne Franche-Comté, Besancon, France, 13 November 2020.
72. Bhangale, U.; Patil, S.; Vishwanath, V.; Thakker, P.; Bansode, A.; Navandhar, D. Near Real-time Crowd Counting using Deep Learning Approach. *Procedia Comput. Sci.* **2020**, *171*, 770–779. [CrossRef]
73. Collobert, R.; Jason, W. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 160–167.
74. Idrees, H.; Tayyab, M.; Athrey, K.; Zhang, D.; Al-Maadeed, S.; Rajpoot, N. Composition loss for counting, density map estimation and localization in dense crowds. *arXiv* **2018**, arXiv:1808.01050.
75. Jacobs, H. To count a crowd. *Columbia J. Rev.* **1967**, *6*, 36–40.

76. Tsai, Y.-H.H.; Yeh, Y.-R.; Wang, Y.-C.F. Learning Cross-Domain Landmarks Heterog. Domain Adaptation. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5081–5090.
77. Hoffman, J.; Rodner, E.; Donahue, J.; Kulis, B.; Saenko, K. Asymmetric and category invariant feature transformations for domain adaptation. *Int. J. Comput. Vis.* **2014**, *109*, 28–41. [CrossRef]
78. Liu, C.; Weng, X.; Mu, Y. Recurrent attentive zooming for joint crowd counting and precise localization. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1217–1226.
79. Gao, J.; Han, T.; Wang, Q.; Yuan, Y. Domain-adaptive crowd counting via inter-domain features segregation and gaussian-prior reconstruction. *arXiv* **2019** arXiv:1912.03677.
80. Hu, Y.; Chang, H.; Nian, F.; Wang, Y.; Li, T. Dense crowd counting from still images with convolutional neural networks. *J. Vis. Commun. Image Represent.* **2016**, *38*, 530–539. [CrossRef]
81. Xue, Y.; Liu, S.; Li, Y.; Qian, X. Crowd Scene Analysis by Output Encoding. *arXiv* **2020**, arXiv:2001.09556.
82. Liu, L.; Amirgholipour, S.; Jiang, J.; Jia, W.; Zeibots, M.; He, X. Performance-enhancing network pruning for crowd counting. *Neurocomputing* **2019**, *360*, 246–253. [CrossRef]
83. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
84. Huang, G.; Liu, Z.; Maaten, L.V.D.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
85. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
86. Shahira, A.; Minh, H.; Dimitris, S.; Chao, C. Localization in the crowd with topological constraints. *arXiv* **2021**, arXiv:2012.12482.
87. Liang, D.; Wei, X.; Zhu, Y.; Zhou, Y. Focal Inverse Distance Transform Maps for Crowd Localization and Counting in Dense Crowd. *arXiv* **2021**, arXiv:2102.07925.
88. Hu, P.; Deva, R. Finding tiny faces. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
89. Liu, W.; Salzmann, M.; Fua, P. Context-aware crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5099–5108.
90. Jiang, X.; Xiao, Z.; Zhang, B.; Zhen, X.; Cao, X.; Doermann, D.; Shao, L. Crowd counting and density estimation by trellis encoderdecoder networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6133–6142.
91. Wang, Q.; Gao, J.; Lin, W.; Yuan, Y. Learning from synthetic data for crowd counting in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8198–8207.
92. 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.
93. Guo, D.; Li, K.; Zha, Z.-J.; Meng, W. Dadnet: Dilated-attentiondeformable convnet for crowd counting. In Proceedings of the ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 1823–1832.
94. Wan, J.; Chan, A. Adaptive density map generation for crowd counting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1130–1139.
95. Gao, J.; Yuan, Y.; Wang, Q. Feature-aware adaptation and density alignment for crowd counting in video surveillance. *IEEE Trans. Cybernet.* **2020**, 1–12. [CrossRef]
96. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv* **2017**, arXiv:1703.10593.