Pushing the Frontiers of Unconstrained Crowd Counting: New Dataset and Benchmark Method

Vishwanath A. Sindagi Rajeev Yasarla Vishal M. Patel
Department of Electrical and Computer Engineering,
Johns Hopkins University, 3400 N. Charles St, Baltimore, MD 21218, USA
{vishwanathsindagi,rajeevyasarla,vpatel36}@jhu.edu

Abstract

In this work, we propose a novel crowd counting network that progressively generates crowd density maps via residual error estimation. The proposed method uses VGG16 as the backbone network and employs density map generated by the final layer as a coarse prediction to refine and generate finer density maps in a progressive fashion using residual learning. Additionally, the residual learning is guided by an uncertainty-based confidence weighting mechanism that permits the flow of only high-confidence residuals in the refinement path. The proposed Confidence Guided Deep Residual Counting Network (CG-DRCN) is evaluated on recent complex datasets, and it achieves significant improvements in errors.

Furthermore, we introduce a new large scale unconstrained crowd counting dataset (JHU-CROWD) that is $\sim$2.8 $\times$ larger than the most recent crowd counting datasets in terms of the number of images. It contains 4,250 images with 1.11 million annotations. In comparison to existing datasets, the proposed dataset is collected under a variety of diverse scenarios and environmental conditions. Specifically, the dataset includes several images with weather-based degradations and illumination variations in addition to many distractor images, making it a very challenging dataset. Additionally, the dataset consists of rich annotations at both image-level and head-level. Several recent methods are evaluated and compared on this dataset.

1. Introduction

With burgeoning population and rapid urbanization, crowd gatherings have become more prominent in the recent years. Consequently, computer vision-based crowd analytics and surveillance [5, 10, 18, 19, 27, 28, 34, 37, 38, 44, 46, 57, 59, 61, 63] have received increased interest. Furthermore, algorithms developed for the purpose of crowd analytics have found applications in other fields such as agriculture monitoring [26], microscopic biology [16], urban planning and environmental survey [8, 57]. Current state-of-the-art counting networks achieve impressive error rates on a variety of datasets that contain numerous challenges. Their success can be broadly attributed to two major factors: (i) design of novel convolutional neural network (CNN) architectures specifically for improving count performance [4, 29, 33, 36, 38, 43, 50, 59], and (ii) development and publication of challenging datasets [10, 11, 59, 61]. In this paper, we consider both of the above factors in an attempt to further improve the crowd counting performance.

Design of novel networks specifically for the task of counting has improved the counting error by leaps and bounds. Architectures have evolved from the simple ones like [59] which consisted of a set of convolutional and fully connected layers, to the most recent complex architectures like SA-Net [4] which consists of a set of scale aggregation modules. Typically, most existing works ([2, 4, 4, 29, 33, 38, 43, 44, 47, 50, 59, 61]) have designed their networks by laying a strong emphasis on addressing large variations of scale in crowd images. While this strategy of developing robustness towards scale changes has resulted in significant performance gains, it is nevertheless important to exploit other properties like in [33, 39, 41] to further the improvements.

In a similar attempt, we exploit residual learning mechanism for the purpose of improving crowd counting. Specifically, we present a novel design based on the VGG16 network [42], which employs residual learning to progressively generate better quality crowd density maps. This use of residual learning is inspired by its success in several other tasks like super-resolution [13, 15, 15, 21, 49]. Although this technique results in improvements in performance, it is important to ensure that only highly confident residuals are used in order to ensure the effectiveness of residual learning. To address this issue, we draw inspiration from the success of uncertainty-based learning mechanism [7, 14, 65]. We propose an uncertainty-based confidence weighting module that captures high-confidence regions in

the feature maps to focus on during the residual learning. The confidence weights ensure that only highly confident residuals get propagated to the output, thereby increasing the effectiveness of the residual learning mechanism.

In addition to the new network design, we identify the next set of challenges that require attention from the crowd counting research community and collect a large-scale dataset collected under a variety of conditions. Existing efforts like UCF_CROWD_50 [10], World Expo '10 [59] and ShanghaiTech [58] have progressively increased the complexity of the datasets in terms of average count per image, image diversity *etc*. While these datasets have enabled rapid progress in the counting task, they suffer from shortcomings such as limited number of training samples, limited diversity in terms of environmental conditions, dataset bias in terms of positive samples, and limited set of annotations. More recently, Idrees *et al*. [11] proposed a new dataset called UCF-QNRF that alleviates some of these challenges. Nevertheless, they do not specifically consider some of the challenges such as adverse environmental conditions, dataset bias and limited annotation data.

To address these issues, we propose a new large-scale unconstrained dataset with a total of 4,250 images (containing 1,114,785 head annotations) that are collected under a variety of conditions. Specific care is taken to include images captured under various weather-based degradations. Additionally, we include a set of distractor images that are similar to the crowd images but do not contain any crowd. Furthermore, the dataset also provides a much richer set of annotations at both image-level and head-level. We also benchmark several representative counting networks, providing an overview of the state-of-the-art performance.

Following are our key contributions in this paper:

- We propose a crowd counting network that progressively incorporates residual mechanism to estimate high quality density maps. Furthermore, a set of uncertainty-based confidence weighting modules are introduced in the network to improve the efficacy of residual learning.
- We propose a new large-scale unconstrained crowd counting dataset with the largest number of images till date. The dataset specifically includes a number of images collected under adverse weather conditions. Furthermore, this is the first counting dataset that provides a rich set of annotations such as occlusion, blur, image-level labels, *etc*.

2. Related work

Crowd Counting. Traditional approaches for crowd counting from single images are based on hand-crafted representations and different regression techniques. Loy *et al*. [25] categorized these methods into (1) detection-based methods [17] (2) regression-based methods [6, 10, 35] and (3) density estimation-based methods [16, 31, 55]. Interested readers are referred to [6, 18] for more comprehensive study of different crowd counting methods.

Recent advances in CNNs have been exploited for the task of crowd counting and these methods [1, 3, 29, 30, 38, 38, 44, 50, 52, 54, 59, 61] have demonstrated significant improvements over the traditional methods. A recent survey [45] categorizes these approaches based on the network property and the inference process. Walach *et al*. [50] used CNNs with layered boosting approach to learn a non-linear function between an image patch and count. Recent work [29, 61] addressed the scale issue using different architectures. Sam *et al*. [38] proposed a VGG16-based switching classifier that first identifies appropriate regressor based on the content of the input image patch. More recently, Sindagi *et al*. [44] proposed to incorporate global and local context from the input image into the density estimation network. In another approach, Cao *et al*. [4] proposed a encoder-decoder network with scale aggregation modules.

In contrast to these methods that emphasize on specifically addressing large-scale variations in head sizes, the most recent methods ([2] ,[39], [41], [24], [33]) have focused on other properties of the problem. For instance, Babu *et al*. [2] proposed a mechanism to incrementally increase the network capacity conditioned on the dataset. Shen *et al*. [39] overcame the issue of blurred density maps by utilizing adversarial loss. In a more recent approach, Ranjan *et al*. [33] proposed a two-branch network to estimate density map in a cascaded manner. Shi *et al*. [41] employed deep negative correlation based learning for more generalizable features. Liu *et al*. [24] used unlabeled data for counting by proposing a new framework that involves learning to rank.

Recent approaches like [22, 47, 48, 51, 62] have aimed at incorporating various forms of related information like attention [22], semantic priors [51], segmentation [62], inverse attention [48], and hierarchical attention [47] respectively into the network. Other techniques such as [12, 23, 40, 60] leverage features from different layers of the network using different techniques like trellis style encoder decoder [12], explicitly considering perspective [40], context information [23], and multiple views [60].

Crowd Datasets. Crowd counting datasets have evolved over time with respect to a number of factors such as size, crowd densities, image resolution, and diversity. UCSD [5] is among one of the early datasets proposed for counting and it contains 2000 video frames of low resolution with 49,885 annotations. The video frames are collected from a single frame and typically contain low density crowds. Zhang *et al*. [59] addressed the limitations of UCSD dataset by introducing the WorldExpo dataset that contains 108 videos with a total of 3,980 frames belonging to 5 different scenes. While the UCSD and WorldExpo datasets contain only low/low-medium densities, Idrees *et al*. [10] proposed

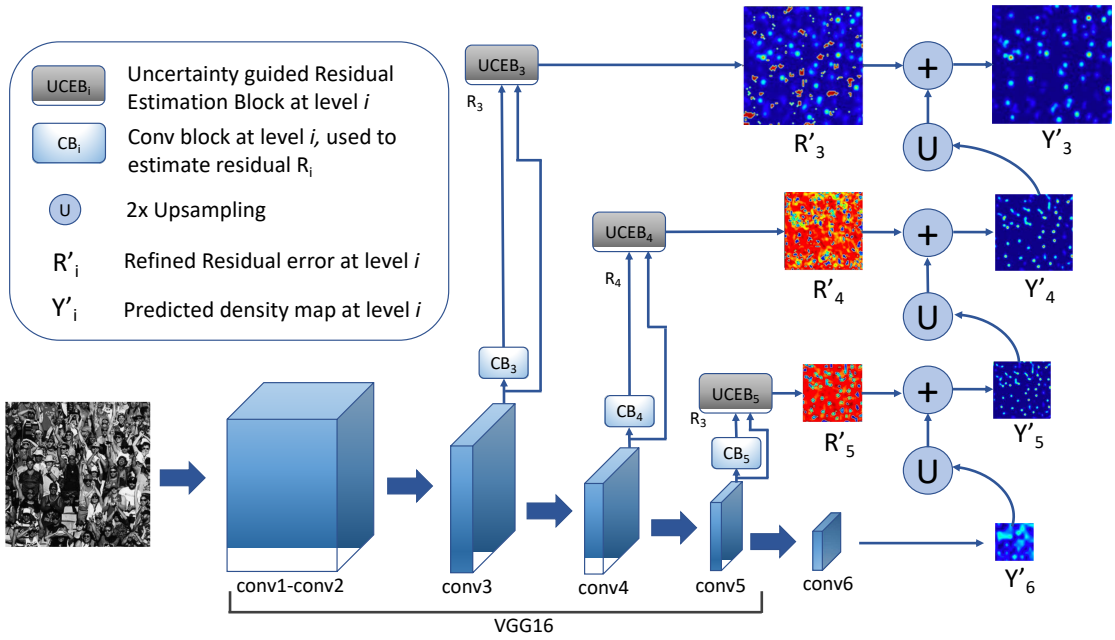


Figure 1. Overview of the proposed method. Coarse density map from the deepest layer of the base network is refined using the residual map estimated by the shallower layer. The residual estimation is performed by convolutional block, CB_i and is further refined in $UCEB_i$. Note that, the conv features from the main branch are first reduced to 32 dimensions using 1×1 conv before forwarding them to $UCEB_i$ along with R_i. In the residual maps, red indicates negative values and cyan indicates positive value.

the UCF_CROWD_50 dataset specifically for very high density crowd scenarios. However, the dataset consists of only 50 images rendering it impractical for training deep networks. Zhang *et al*. [61] introduced the ShanghaiTech dataset which has better diversity in terms of scenes and density levels as compared to earlier datasets. The dataset is split into two parts: Part A (containing high density crowd images) and Part B (containing low density crowd images). The entire dataset contains 1,198 images with 330,165 annotations. Recently, Idrees *et al*. [11] proposed a new large-scale crowd dataset containing 1,535 high density images images with a total of 1.25 million annotations. Wang *et al*. [53] introduced a synthetic crowd dataset that contains diverse scenes. In addition, they proposed a SSIM based CycleGAN [64] for adapting the network trained on synthetic images to real world images.

3. Proposed method

In this section, we present the details of the proposed Confidence Guided Deep Residual Crowd Counting (CG-DRCN) along with the training and inference specifics. Fig. 1 shows the architecture of the proposed network.

3.1. Base network

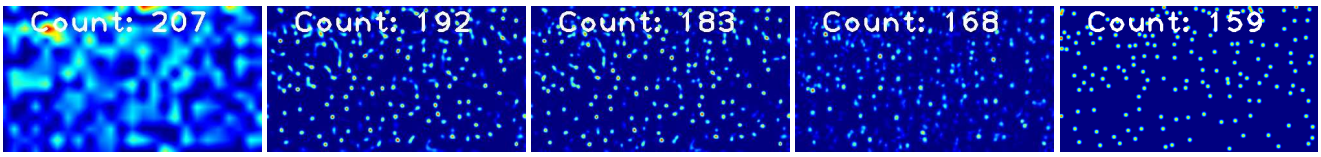
Following recent approaches [4, 38, 44], we perform counting based on the density estimation framework. In this framework, the network is trained to estimate the density map ($\hat{Y}$) from an input crowd image (X). The target density map (Y) for training the network is generated by imposing normalized 2D Gaussian at head locations provided by the dataset annotations: $Y(x) = \sum_{x_g \in S} \mathcal{N}(x - x_g, \sigma)$, where, S is the set of all head locations (x_g) in the input image and σ is scale parameter of 2D Gaussian kernel. Due to this formulation, the density map contains per-pixel density information of the scene, which when integrated results in the count of people in the image.

The proposed network consists of conv1$\sim$conv5 layers ($C_1 - C_5$) of the VGG16 architecture as a part of the backbone, followed by a conv block (CB_6) and a max-pooling layer with stride 2. First, the input image (of size $W \times H$) is passed through $C_1 - C_5$, CB_6 and the max pooling layer to produce the corresponding density map ($\hat{Y}_6$) of size $\frac{W}{32} \times \frac{H}{32}$. CB_6 is defined by $\{\text{conv}_{512,32,1}\text{-relu-conv}_{32,32,3}\text{-relu-conv}_{32,1,3}\}$[1]). Due to its low resolution, ($\hat{Y}_6$) can be considered as a coarse estimation, and learning this will implicitly incorporate global context in the image due the large receptive field at the deepest layer in the network.

3.2. Residual learning

Although $\hat{Y}_6$ provides a good estimate of the number of people in the image, the density map lacks several local de-

[1] $conv_{N_i,N_o,k}$ denotes conv layer (with N_i input channels, N_o output channels, $k\times k$ filter size), *relu* denotes ReLU activation



(a) (b) (c) (d) (e)

Figure 2. Density maps estimated by different layers of the proposed network. (a) $\hat{Y}_6$ (b) $\hat{Y}_5$ (c) $\hat{Y}_4$ (d) $\hat{Y}_3$ (e) Y(ground-truth). It can be observed that the output of the deepest layer ($\hat{Y}_6$) looks very coarse, and it is refined in a progressive manner using the residual learned by the conv blocks CB_5, CB_4, CB_3 to obtain the $\hat{Y}_5, \hat{Y}_4, \hat{Y}_3$ respectively. Note that fine details and the total count in the density maps improve as we move from $\hat{Y}_6$ to $\hat{Y}_3$.

tails as shown in Fig. 2 (a). This is because deeper layers learn to capture abstract concepts and tend to lose low level details in the image. On the other hand, the shallower layers have relatively more detailed local information as compared to their deeper counterparts [32]. Based on this observation, we propose to refine the coarser density maps by employing shallower layers in a residual learning framework. This refinement mechanism is inspired in part by several leading work on super-resolution [15, 21, 49] that incorporate residual learning to learn finer details required to generate a high quality super-resolved image. Specifically, features from C_5 are forwarded through a conv-block (CB_5) to generate a residual map R_5, which is then added to an appropriately up-sampled version of $\hat{Y}_6$ to produce the density map $\hat{Y}_5$ of size $\frac{W}{16} \times \frac{H}{16}$, *i.e.*,

$$\hat{Y}_5 = R_5 + up(\hat{Y}_6). \tag{1}$$

Here, $up()$ denotes up-sampling by a factor of $2\times$ via bilinear interpolation. By enforcing CB_5 to learn a residual map, the network focuses on the local errors emanating from the deeper layer, resulting in better learning of the offsets required to refined the coarser density map. CB_5 is defined by $\{\text{conv}_{512,32,1}\text{-relu-conv}_{32,32,3}\text{-relu-conv}_{32,1,3}\}$[1].

The above refinement is further repeated to recursively generate finer density maps $\hat{Y}_4$ and $\hat{Y}_3$ using the feature maps from the shallower layers C_4 and C_3, respectively. Specifically, the output of C_4 and C_3 are forwarded through CB_4, CB_3 to learn residual maps R_4 and R_3, which are then added to the appropriately up-sampled versions of the coarser maps $\hat{Y}_5$ and $\hat{Y}_4$ to produce $\hat{Y}_4$ and $\hat{Y}_3$ respectively in that order. CB_4 is defined by $\{\text{conv}_{512,32,1}\text{-relu-conv}_{32,32,3}\text{-relu-conv}_{32,1,3}\}$[1]. CB_3 is defined by $\{\text{conv}_{256,32,1}\text{-relu-conv}_{32,32,3}\text{-relu-conv}_{32,1,3}\}$[1]. Specifically, $\hat{Y}_4$ and $\hat{Y}_3$ are obtained as follows: $\hat{Y}_4 = R_4 + up(\hat{Y}_5), \hat{Y}_3 = R_3 + up(\hat{Y}_4)$.

3.3. Confidence guided residual learning

In order to improve the efficacy of the residual learning mechanism discussed above, we propose an uncertainty guided confidence estimation block (UCEB) to guide the refinement process. The task of conv blocks CB_5, CB_4, CB_3 is to capture residual errors that can be incorporated into the coarser density maps to produce high quality density maps in the end. For this purpose, these conv blocks employ feature maps from shallower conv layers C_5, C_4, C_3. Since these conv layers primarily trained for estimating the coarsest density map, their features have high responses in regions where crowd is present, and hence, they may not necessarily produce effective residuals. In order to overcome this issue, we propose to gate the residuals that are not effective using uncertainty estimation. Inspired by uncertainty estimation in CNNs [7, 14, 56, 65], we aim to model pixel-wise aleatoric uncertainty of the residuals estimated by CB_5, CB_4, CB_3. That is we, predict the pixel-wise confidence (inverse of the uncertainties) of the residuals which are then used to gate the residuals before being passed on to the subsequent outputs. This ensures that only highly confident residuals get propagated to the output.

In terms of the overall architecture, we introduce a set of UCEBs as shown in Fig. 1. Each residual branch consists of one such block. The UCEB$_i$ takes the residual R_i and dimensionality reduced features from the main branch as input, concatenates them, and forwards it through a set of conv layers ($\{\text{conv}_{33,32,1}\text{-relu-conv}_{32,16,3}\text{-relu-conv}_{16,16,3}\text{-relu-conv}_{16,1,1}\}$) and produces a confidence map CM_i which is then multiplied element-wise with the input to form the refined residual map: $\hat{R}_i = R_i \odot CM_i$. Here $\odot$ denotes element-wise multiplication.

In order to learn these confidence maps, the loss function L_f used to train the network is defined as follows,

$$L_f = L_d - \lambda_c L_c, \tag{2}$$

where, λ_c is a regularization constant, L_d is the pixel-wise regression loss to minimize the density map prediction error and is defined as:

$$L_d = \sum_{i \in \{3,4,5,6\}} \|(CM_i \odot Y_i) - (CM_i \odot \hat{Y}_i)\|_2, \tag{3}$$

where, $\hat{Y}_i$ is the predicted density map, i indicates the index of the conv layer from which the predicted density map is taken, Y_i is the corresponding target.

L_c is the confidence guiding loss, defined as,

$$L_c = \sum_{i \in \{3,4,5,6\}} \sum_{j=1}^{H} \sum_{k=1}^{W} \log(CM_i^{j,k}), \tag{4}$$

where, $W \times H$ is the dimension of the confidence map CM_i. As it can be seen from Eq. (2), the loss L_f has two parts L_d and L_c. The first term minimizes the Euclidean distance between the prediction and target features, whereas L_c maximizes the confidence scores CM_i by making them closer to 1.

Fig. 2 illustrates the output density maps ($\hat{Y}_6, \hat{Y}_5, \hat{Y}_4, \hat{Y}_3$) generated by the proposed network for a sample crowd image. It can be observed that the density maps progressively improve in terms of fine details and the count value.

3.4. Training and inference details

The training dataset is obtained by cropping patches from multiple random locations in each training image. The cropped patch-size is 224×224. We randomly sub-sample 10% of the training set (before cropping) and keep it aside for validating the training models. We use the Adam optimizer to train the network. We use a learning rate of 0.00001 and a momentum of 0.9.

For inference, the density map $\hat{Y}_3$ is considered as the final output. The count performance is measured using the standard error metrics: mean absolute error (MAE) and mean squared error (MSE). These metrics are defined as follows: $MAE = \frac{1}{N}\sum_{i=1}^{N}|Y_i - Y_i'|$ and $MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|Y_i - Y_i'|^2}$ respectively, where N is the number of test samples, Y_i is the ground-truth count and Y_i' is the estimated count corresponding to the i^{th} sample.

4. JHU-CROWD: Unconstrained Crowd Counting Dataset

In this section, we first motivate the need for a new crowd counting dataset, followed by a detailed description of the various factors and conditions while collecting the dataset.

4.1. Motivation and dataset details

As discussed earlier, existing datasets (such as UCF_CROWD_50 [10], World Expo '10 [59] and ShanghaiTech [58]) have enabled researchers to develop novel counting networks that are robust to several factors such as variations in scale, pose, view *etc*. Several recent methods have specifically addressed the large variations in scale by proposing different approaches such as multi-column networks [61], incorporating global and local context [44], scale aggregation network [4], *etc*. These methods are largely successful in addressing issues in the existing datasets, and there is pressing need to identify newer set of challenges that require attention from the crowd counting community.

Table 1. Summary of images collected under adverse conditions.

Degradation type	Rain	Snow	Fog/Haze	Total
Num. of images	151	190	175	516
Num. of annotations	32,832	32,659	37,070	102,561

In what follows, we describe the shortcomings of existing datasets and discuss the ways in which we overcome them:

(i) *Limited number of training samples*: Typically, crowd counting datasets have limited number of images available for training and testing. For example, ShanghaiTech dataset [61] has only 1,198 images and this low number of images results in lower diversity of the training samples. Due to this issue, networks trained on this dataset will have reduced generalization capabilities. Although datasets like Mall [6], WorldExpo '10 [59] have higher number of images, it is important to note that these images are from a set of video sequences from surveillance cameras and hence, they have limited diversity in terms of background scenes and number of people. Most recently, Idrees *et al*. [11] addressed this issue by introducing a high-quality dataset (UCF-QNRF) that has images collected from various geographical locations under a variety of conditions and scenarios. Although it has a large set of diverse scenarios, the number of samples is still limited from the perspective of training deep neural networks.

To address this issue, we collect a new large scale unconstrained dataset with a total of 4,250 images that are collected under a variety of conditions. Such a large number of images results in increased diversity in terms of count, background regions, scenarios *etc*. as compared to existing datasets. The images are collected from several sources on the Internet using different keywords such as crowd, crowd+marathon, crowd+walking, crowd+India, *etc*.

(ii) *Absence of adverse conditions*: Typical application of crowd counting is video surveillance in outdoor scenarios which involve regular weather-based degradations such as haze, snow, rain *etc*. It is crucial that networks, deployed under such conditions, achieve more than satisfactory performance.

To overcome this issue, specific care is taken during our dataset collection efforts to include images captured under various weather-based degradations such as rain, haze, snow, *etc*. (as as shown in Fig. 3(b-d)). Table 1 summarizes images collected under adverse conditions.

(iii) *Dataset bias*: Existing datasets focus on collecting only images with crowd, due to which a deep network trained on such a dataset may end up learning bias in the dataset. Due to this error, the network will erroneously predict crowd even in scenes that do not contain crowd.

In order to address this, we include a set of distractor images that are similar to crowd images but do not contain any crowd. These images can enable the network to avoid learning bias in the dataset. The total number of distractor images in the dataset is 100. Fig 3(e) shows sample distractor images.

(a) (b) (c) (d) (e)

Figure 3. Representative samples of the images in the JHU-CROWD dataset. (a) Overall (b) Rain (c) Snow (d) Haze (e) Distractors.

Table 2. Comparison of different datasets. P: Point-wise annotations for head locations, O: Occlusion level per head, B: Blur level per head, S: Size indicator per head, I: Image level labels.

Dataset	Num of Images	Num of Annotations	Avg Count	Max Count	Avg Resolution	Weather degradations	Distractors	Type of annotations
UCSD [5]	2000	49,885	25	46	158×238	✗	✗	P
Mall [6]	2000	62,325	-	53	320×240	✗	✗	P
UCF_CROWD_50 [10]	50	63,974	1279	4543	2101×2888	✗	✗	P
WorldExpo '10 [59]	3980	199,923	50	253	576×720	✗	✗	P
ShanghaiTech [61]	1198	330,165	275	3139	598×868	✗	✗	P
UCF-QNRF [11]	1535	1,251,642	815	12865	2013×2902	✗	✗	P
JHU-CROWD (proposed)	4250	1,114,785	262	7286	1450×900	✓	✓	P, O, B, S, I

(iv) *Limited annotations*: Typically, crowd counting datasets provide point-wise annotations for every head/person in the image, *i.e.*, each image is provided with a list of x, y locations of the head centers. While these annotations enable the networks to learn the counting task, absence of more information such as occlusion level, head sizes, blur level *etc*. limits the learning ability of the networks. For instance, due to the presence of large variations in perspective, size of the head is crucial to determine the precise count. One of the reasons for these missing annotations is that crowd images typically contain several people and it is highly labor intensive to obtain detailed annotations such as size.

To enable more effective learning, we collect a much richer set of annotations at both image-level and head-level. Head-level annotation include x, y locations of heads and corresponding occlusion level, blur level and size level. Occlusion label has three levels: {*un-occluded, partially occluded, fully occluded*}. Blur level has two labels: {*blur, no-blur*}. Since obtaining the size is a much harder issue, each head is labeled with a size indicator. Annotators were instructed to first annotate the largest and smallest head in the image with a bounding box. The annotators were then instructed to assign a size level to every head in the image such that this size level is indicative of the relative size with respect to the smallest and largest annotated bounding box. Image level annotations include labels (such as *marathon, mall, walking, stadium* etc.) and the weather conditions under which the images were captured. The total number of point-level annotations in the dataset are 1,114,785.

4.2. Summary and evaluation protocol

Fig. 3 illustrates representative samples of the images in the JHU-CROWD dataset under various categories. Table 2 summarizes the proposed JHU-CROWD dataset in comparison with the existing ones. It can be observed that the proposed dataset is the largest till date in terms of the number of images and enjoys a host of other properties such as a richer set of annotations, weather-based degradations and distractor images. With these properties, the proposed dataset will serve as a good complementary to other datasets such as UCF-QNRF. The dataset is randomly split into training and test sets, which contain 3,188 and 1,062 images respectively.

(a) (b) (c)

Figure 4. Results of the proposed dataset on sample images from the JHU-CROWD dataset. (a) Input image (b) Ground-truth density map (c) Estimated density map.

Following the existing work, we perform evaluation using the standard MAE and MSE metrics. Furthermore, these metrics are calculated for the following sub-categories of images: (i) Low density: images containing count between 0 and 50, (ii) Medium density: images containing count between 51 and 500, (iii) High density: images with count more than 500 people, (iv) Distractors: images containing 0 count, (v) Weather-based degradations, and (vi) Overall. The metrics under these sub-categories will enable a deeper understanding of the network performance.

5. Experimental details and results

In this section, we first discuss the results of an ablation study conducted to analyze the effect of different components in the proposed network. This is followed by a discussion on benchmarking of recent crowd counting algorithms including the proposed residual-based counting network on the JHU-CROWD dataset. Finally, we compared the proposed method with recent approaches on the ShanghaiTech [61] and UCF-QNRF [11] datasets.

5.1. Ablative Study

Due to the presence of various complexities such as high density crowds, large variations in scales, presence of occlusion, etc, we chose to perform the ablation study on JHU-CROWD dataset.

The ablation study consisted of evaluating the following configurations of the proposed method: (i) Base network: VGG16 network with an additional conv block (CB_6) at the end, (ii) Base network + R: the base network with residual learning as discussed in Section 3.2 , (iii) Base network + R + UCEB ($\lambda_c = 0$): the base network with residual learning guided by the confidence estimation blocks as discussed in Section 3.3. In this configuration, we aim to measure the performance due to the addition of the confidence estimation blocks without the uncertainty estimation mechanism by setting λ_c is set to 0, (iv) Base network + R + UCEB ($\lambda_c = 1$): the base network with residual learning guided by the confidence estimation blocks as discussed in Section 3.3. The results of these experiments are shown in Table 3. It can be seen that there are considerable improvements in the performance due to the inclusion of residual learning into the network. The use of confidence-based weighting of the residuals results in further improvements, thus highlighting its significance in improving the efficacy of uncertainty-based residual learning.

Table 3. Results of ablation study on the JHU-CROWD dataset.

Method	MAE	MSE
Base network	81.1	248.5
Base network + R	76.4	218.6
Base network + R + UCEB ($\lambda_c = 0$)	74.6	215.5
Base network + R + UCEB ($\lambda_c = 1$)	66.1	195.5

5.2. JHU-CROWD dataset

In this section, we discuss the benchmarking of recent algorithms including the proposed method on the new dataset.

Benchmarking and comparison. We benchmark recent algorithms on the newly proposed JHU-CROWD dataset. Specifically, we evaluate the following recent works: mulit-

Table 4. Results on JHU-CROWD dataset.

Category	Distractors		Low		Medium		High		Weather		Overall	
Method	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
MCNN [61]	103.8	238.5	37.7	92.5	84.1	185.2	499.6	795.5	128.2	288.3	109.3	291.0
CMTL [43]	135.8	263.8	47.0	106.0	82.4	198.3	407.8	660.2	117.8	260.1	102.5	262.6
Switching CNN [38]	100.5	235.5	32.1	80.5	76.1	173.1	395.1	640.1	105.1	245.2	99.1	255.1
SA-Net(image-based) [4]	71.9	167.7	30.0	76.6	65.4	121.5	516.3	762.7	99.4	234.9	98.0	260.3
CSR-Net [20]	44.3	102.4	15.8	39.9	48.4	77.7	463.5	746.1	96.5	284.6	78.4	242.7
CG-DRCN (proposed)	**43.4**	**97.8**	**15.7**	**38.9**	**44.0**	**73.2**	**346.2**	**569.5**	**80.9**	**227.31**	**66.1**	**195.5**

Table 5. Results on ShanghaiTech dataset [61].

	Part-A		Part-B	
Method	MAE	MSE	MAE	MSE
Cascaded-MTL [43]	101.3	152.4	20.0	31.1
Switching-CNN [38]	90.4	135.0	21.6	33.4
CP-CNN [44]	73.6	106.4	20.1	30.1
IG-CNN [2]	72.5	118.2	13.6	21.1
Liu *et al*. [24]	73.6	112.0	13.7	21.4
D-ConvNet [41]	73.5	112.3	18.7	26.0
CSRNet [20]	68.2	115.0	10.6	16.0
ic-CNN [33]	69.8	117.3	10.7	16.0
SA-Net (image-based) [4]	88.1	134.3	-	-
SA-Net (patch-based) [4]	67.0	104.5	8.4	13.6
ACSCP [39]	75.7	102.7	17.2	27.4
Jian *et al*. [12]	64.2	109.1	**8.2**	**12.8**
CG-DRCN (proposed)	**64.0**	**98.4**	8.5	14.4

Table 6. Results on UCF-QNRF datastet [11].

Method	MAE	MSE
Idrees *et al*. [10]	315.0	508.0
Zhang *et al*. [59]	277.0	426.0
CMTL *et al*. [43]	252.0	514.0
Switching-CNN [38]	228.0	445.0
Idrees *et al*. [11]	132.0	191.0
Jian *et al*. [12]	113.0	188.0
CG-DRCN (proposed)	**112.2**	**176.3**

column network (MCNN) [61], cascaded multi-task learning for crowd counting (CMTL) [43], Switching-CNN [38], CSR-Net [20] and SANet [4] [2]. Furthermore, we also evaluate the proposed method (CG-DRCN) and demonstrate its effectiveness over the other methods.

All the networks are trained using the entire training set and evaluated under six different categories. For a fair comparison, the same training strategy (in terms of cropping patches), as described in Section 3.4, is used. Table 4 shows the results of the above experiments for various sub-categories of images in the test set. It can be observed that the proposed method outperforms the other methods in general. Furthermore, it may also be noted that the overall performance does not necessarily indicate the proposed method performs well in all the sub-categories. Hence, it is essential to compare the methods for each of the sub-category.

[2]We used the implementation provided by [9]

5.3. Comparison on other datasets

ShanghaiTech: The proposed network is trained on the train splits using the same strategy as discussed in Section 3.4. Table 5 shows the results of the proposed method on ShanghaiTech as compared with several recent approaches ([38], [44], [2], [41], [24], [20], [33] , [4], [39] and [12]). It can be observed that the proposed method outperforms all existing methods on Part A of the dataset, while achieving comparable performance on Part B.

UCF-QNRF: Results on the UCF-QNRF [11] dataset as compared with recent methods ([10],[61],[43]) are shown in Table 6. The proposed method is compared against different approaches: [10], [61], [43],[38], [11] and [12]. It can be observed that the proposed method outperforms other methods by a considerable margin.

6. Conclusions

In this paper, we presented a novel crowd counting network that employs residual learning mechanism in a progressive fashion to estimate coarse to fine density maps. The efficacy of residual learning is further improved by introducing an uncertainty-based confidence weighting mechanism that is designed to enable the network to propagate only high-confident residuals to the output. Experiments on recent datasets demonstrate the effectiveness of the proposed approach. Furthermore, we also introduced a new large scale unconstrained crowd counting dataset (JHU-CROWD) consisting of 4,250 images with 1.11 million annotations. The new dataset is collected under a variety of conditions and includes images with weather-based degradations and other distractors. Additionally, the dataset provides a rich set of annotations such as head locations, blur-level, occlusion-level, size-level and other image-level labels.

Acknowledgment

This work was supported by the NSF grant 1910141.

References

[1] Carlos Arteta, Victor Lempitsky, and Andrew Zisserman. Counting in the wild. In *European Conference on Computer Vision*, pages 483–498. Springer, 2016. 2

[2] Deepak Babu Sam, Neeraj N Sajjan, R Venkatesh Babu, and Mukundhan Srinivasan. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3626, 2018. 1, 2, 8

[3] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 640–644. ACM, 2016. 2

[4] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *European Conference on Computer Vision*, pages 757–773. Springer, 2018. 1, 2, 3, 5, 8

[5] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008. 1, 2, 6

[6] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *European Conference on Computer Vision*, 2012. 2, 5, 6

[7] Terrance DeVries and Graham W Taylor. Learning confidence for out-of-distribution detection in neural networks. *arXiv preprint arXiv:1802.04865*, 2018. 1, 4

[8] Geoffrey French, Mark Fisher, Michal Mackiewicz, and Coby Needle. Convolutional neural networks for counting fish in fisheries surveillance video. In *British Machine Vision Conference Workshop*. BMVA Press, 2015. 1

[9] Junyu Gao, Wei Lin, Bin Zhao, Dong Wang, Chenyu Gao, and Jun Wen. C^ 3 framework: An open-source pytorch code for crowd counting. *arXiv preprint arXiv:1907.02724*, 2019. 8

[10] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554, 2013. 1, 2, 5, 6, 8

[11] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *European Conference on Computer Vision*, pages 544–559. Springer, 2018. 1, 2, 3, 5, 6, 7, 8

[12] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder network. *arXiv preprint arXiv:1903.00853*, 2019. 2, 8

[13] Wei Ke, Jie Chen, Jianbin Jiao, Guoying Zhao, and Qixiang Ye. Srn: Side-output residual network for object symmetry detection in the wild. *arXiv preprint arXiv:1703.02243*, 2017. 1

[14] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pages 5574–5584, 2017. 1, 4

[15] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1646–1654, 2016. 1, 4

[16] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, pages 1324–1332, 2010. 1, 2

[17] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008. 2

[18] Teng Li, Huan Chang, Meng Wang, Bingbing Ni, Richang Hong, and Shuicheng Yan. Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(3):367–386, 2015. 1, 2

[19] Weixin Li, Vijay Mahadevan, and Nuno Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2014. 1

[20] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1100, 2018. 8

[21] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, volume 1, page 4, 2017. 1, 4

[22] Ning Liu, Yongchao Long, Changqing Zou, Qun Niu, Li Pan, and Hefeng Wu. Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding. *arXiv preprint arXiv:1811.11968*, 2018. 2

[23] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5099–5108, 2019. 2

[24] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 8

[25] Chen Change Loy, Ke Chen, Shaogang Gong, and Tao Xiang. Crowd counting and profiling: Methodology and evaluation. In *Modeling, Simulation and Visual Analysis of Crowds*, pages 347–382. Springer, 2013. 2

[26] Hao Lu, Zhiguo Cao, Yang Xiao, Bohan Zhuang, and Chunhua Shen. Tasselnet: Counting maize tassels in the wild via local counts regression network. *Plant Methods*, 13(1):79, 2017. 1

[27] Vijay Mahadevan, Weixin Li, Viral Bhalodia, and Nuno Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, volume 249, page 250, 2010. 1

[28] Mark Marsden, Kevin McGuinness, Suzanne Little, Ciara E. Keogh, and Noel E. O'Connor. People, penguins and petri dishes: Adapting object counting models to new visual do-

mains and object types without forgetting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[29] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision*, pages 615–629. Springer, 2016. 1, 2

[30] Daniel Oñoro-Rubio, Roberto Javier López-Sastre, and Mathias Niepert. Learning short-cut connections for object counting. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, 2018. 2

[31] Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo Okada. Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3253–3261, 2015. 2

[32] R. Ranjan, V. M. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135, Jan 2019. 4

[33] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *European Conference on Computer Vision*, pages 278–293. Springer, 2018. 1, 2, 8

[34] Mikel Rodriguez, Ivan Laptev, Josef Sivic, and Jean-Yves Audibert. Density-aware person detection and tracking in crowds. In *2011 International Conference on Computer Vision*, pages 2423–2430. IEEE, 2011. 1

[35] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications, 2009. DICTA'09.*, pages 81–88. IEEE, 2009. 2

[36] Deepak Babu Sam and R Venkatesh Babu. Top-down feedback for crowd counting convolutional neural network. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 1

[37] Deepak Babu Sam, Neeraj N Sajjan, Himanshu Maurya, and R Venkatesh Babu. Almost unsupervised learning for dense crowd counting. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019. 1

[38] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 8

[39] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 8

[40] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting perspective information for efficient crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7279–7288, 2019. 2

[41] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1, 2, 8

[42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 1

[43] Vishwanath A. Sindagi and Vishal M. Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *Advanced Video and Signal Based Surveillance (AVSS), 2017 IEEE International Conference on*. IEEE, 2017. 1, 8

[44] Vishwanath A. Sindagi and Vishal M. Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1, 2, 3, 5, 8

[45] Vishwanath A Sindagi and Vishal M Patel. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 2017. 2

[46] Vishwanath A. Sindagi and Vishal M. Patel. Dafe-fd: Density aware feature enrichment for face detection. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2185–2195. IEEE, 2019. 1

[47] Vishwanath A. Sindagi and Vishal M. Patel. Ha-ccn: Hierarchical attention-based crowd counting network. *arXiv preprint arXiv:1907.10255*, 2019. 1, 2

[48] Vishwanath A. Sindagi and Vishal M. Patel. Inverse attention guided deep crowd counting network. *arXiv preprint*, 2019. 2

[49] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 5, 2017. 1, 4

[50] Elad Walach and Lior Wolf. Learning to count with cnn boosting. In *European Conference on Computer Vision*, pages 660–676. Springer, 2016. 1, 2

[51] Jia Wan, Wenhan Luo, Baoyuan Wu, Antoni B Chan, and Wei Liu. Residual regression with semantic prior for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4036–4045, 2019. 2

[52] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1299–1302. ACM, 2015. 2

[53] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. *arXiv preprint arXiv:1903.03303*, 2019. 3

[54] Ze Wang, Zehao Xiao, Kai Xie, Qiang Qiu, Xiantong Zhen, and Xianbin Cao. In defense of single-column networks for crowd counting. *arXiv preprint arXiv:1808.06133*, 2018. 2

[55] Bolei Xu and Guoping Qiu. Crowd density estimation based on rich features and random projection forest. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–8. IEEE, 2016. 2

[56] Rajeev Yasarla and Vishal M. Patel. Uncertainty guided multi-scale residual learning-using a cycle spinning cnn for single image de-raining. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 4

[57] Beibei Zhan, Dorothy N Monekosso, Paolo Remagnino, Sergio A Velastin, and Li-Qun Xu. Crowd analysis: a survey. *Machine Vision and Applications*, 19(5-6):345–357, 2008. 1

[58] Cong Zhang, Kai Kang, Hongsheng Li, Xiaogang Wang, Rong Xie, and Xiaokang Yang. Data-driven crowd under-

standing: A baseline for a large-scale crowd dataset. *IEEE Transactions on Multimedia*, 18(6):1048–1061, 2016. 2, 5

[59] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 833–841, 2015. 1, 2, 5, 6, 8

[60] Qi Zhang and Antoni B Chan. 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*, pages 8297–8306, 2019. 2

[61] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 589–597, 2016. 1, 2, 3, 5, 6, 7, 8

[62] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12736–12745, 2019. 2

[63] Feng Zhu, Xiaogang Wang, and Nenghai Yu. Crowd tracking with dynamic evolution of group structures. In *European Conference on Computer Vision*, pages 139–154. Springer, 2014. 1

[64] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 3

[65] Lingxue Zhu and Nikolay Laptev. Deep and confident prediction for time series at uber. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 103–110. IEEE, 2017. 1, 4