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,rajeeyasarla,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 ∼2.8 × 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, 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 an 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
the UCF_CROWD50 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. [51] 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 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: $Y(x) = \sum_{x,y \in S} N(x - x_y, \sigma)$, where, $S$ is the set of all head locations ($x_y$) in the input image and $\sigma$ 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~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 $\{conv_{512,32,1},relu-conv_{32,3,3},relu-conv_{32,1,3}\}^1$. 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-
The coarser density maps to produce high quality density is to capture residual errors that can be incorporated into the refinement process. The task of conv blocks guided confidence estimation block (UCEB) to guide the refining mechanism discussed above, we propose an uncertainty estimation in CNNs that incorporates 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 $Y_6$ to produce the density map $\hat{Y}_5$ of size $\frac{W}{16} \times \frac{H}{16}$, i.e.,

$$\hat{Y}_5 = R_5 + up(Y_6). \quad (1)$$

Here, $up()$ denotes up-sampling by a factor of $2 \times 2$ 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 refine the coarser density map. $CB_5$ is defined by \{conv$_{32,32,1}$-relu-conv$_{32,32,3}$-relu-conv$_{32,3,1}$\}. 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 branches $C_4$ 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 \{conv$_{32,32,1}$-relu-conv$_{32,32,3}$-relu-conv$_{32,3,1}$\}. $CB_3$ is defined by \{conv$_{56,32,1}$-relu-conv$_{56,32,3}$-relu-conv$_{32,3,1}$\}. 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 efficiency 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, 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 Eq. 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 ({conv$_{33,32,1}$-relu-conv$_{32,16,3}$-relu-conv$_{16,16,3}$}) and produces a confidence map $CM_i$, which is then multiplied element-wise with the input to form the refined residual map: $R_i = r_i \odot CM_i$. 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,

$$L_f = L_d - \lambda_c L_c, \quad (2)$$

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

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

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,

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

![Figure 2. Density maps estimated by different layers of the proposed network.](image)
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.

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 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.

Table 1. Summary of images collected under adverse conditions.

<table>
<thead>
<tr>
<th>Degradation type</th>
<th>Rain</th>
<th>Snow</th>
<th>Fog/Haze</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. of images</td>
<td>131</td>
<td>190</td>
<td>175</td>
<td>516</td>
</tr>
<tr>
<td>Num. of annotations</td>
<td>32,832</td>
<td>32,659</td>
<td>37,070</td>
<td>102,561</td>
</tr>
</tbody>
</table>
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.
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 \((C_{B_0})\) 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 \(\lambda_c\) 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>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Base network</td>
<td>81.1</td>
<td>248.5</td>
</tr>
<tr>
<td>Base network + R</td>
<td>76.4</td>
<td>218.6</td>
</tr>
<tr>
<td>Base network + R + UCEB ((\lambda_c = 0))</td>
<td>74.6</td>
<td>215.5</td>
</tr>
<tr>
<td>Base network + R + UCEB ((\lambda_c = 1))</td>
<td>66.1</td>
<td>195.5</td>
</tr>
</tbody>
</table>

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: mult-
Table 4. Results on JHU-CROWD dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Category</th>
<th>Distractors</th>
<th>Low</th>
<th>Medium</th>
<th>High</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MSE</td>
<td>MAE</td>
<td>MSE</td>
<td>MAE</td>
<td>MSE</td>
</tr>
<tr>
<td>MCNN [61]</td>
<td>103.8</td>
<td>238.5</td>
<td>37.7</td>
<td>92.5</td>
<td>84.1</td>
<td>185.2</td>
</tr>
<tr>
<td>CMTL [43]</td>
<td>135.8</td>
<td>263.8</td>
<td>47.0</td>
<td>106.0</td>
<td>82.4</td>
<td>198.3</td>
</tr>
<tr>
<td>Switching CNN [38]</td>
<td>100.5</td>
<td>235.5</td>
<td>32.1</td>
<td>80.5</td>
<td>76.1</td>
<td>173.1</td>
</tr>
<tr>
<td>CP-CNN [44]</td>
<td>90.5</td>
<td>210.3</td>
<td>30.2</td>
<td>71.3</td>
<td>71.2</td>
<td>155.3</td>
</tr>
<tr>
<td>IG-CNN [2]</td>
<td>72.5</td>
<td>182.8</td>
<td>13.6</td>
<td>21.1</td>
<td>48.4</td>
<td>77.7</td>
</tr>
<tr>
<td>Liu et al. [24]</td>
<td>73.6</td>
<td>112.0</td>
<td>13.7</td>
<td>21.4</td>
<td>64.0</td>
<td>102.4</td>
</tr>
<tr>
<td>D-ConvNet [41]</td>
<td>73.5</td>
<td>112.3</td>
<td>18.7</td>
<td>26.0</td>
<td>97.8</td>
<td>195.5</td>
</tr>
<tr>
<td>CSRNet [20]</td>
<td>44.3</td>
<td>102.4</td>
<td>15.8</td>
<td>39.9</td>
<td>44.3</td>
<td>102.4</td>
</tr>
<tr>
<td>CG-DRCN (proposed)</td>
<td>43.4</td>
<td>97.8</td>
<td>15.7</td>
<td>38.9</td>
<td>44.0</td>
<td>73.2</td>
</tr>
</tbody>
</table>

Table 5. Results on ShanghaiTech dataset [61].

<table>
<thead>
<tr>
<th>Method</th>
<th>Category</th>
<th>Part-A</th>
<th>Part-B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE</td>
<td>MSE</td>
<td>MAE</td>
</tr>
<tr>
<td>Cascaded-MTL [43]</td>
<td>101.3</td>
<td>152.4</td>
<td>20.0</td>
</tr>
<tr>
<td>Switching-CNN [38]</td>
<td>90.4</td>
<td>135.0</td>
<td>21.6</td>
</tr>
<tr>
<td>CP-CNN [44]</td>
<td>73.6</td>
<td>106.4</td>
<td>20.1</td>
</tr>
<tr>
<td>IG-CNN [2]</td>
<td>72.5</td>
<td>118.2</td>
<td>13.6</td>
</tr>
<tr>
<td>Liu et al. [24]</td>
<td>73.6</td>
<td>112.0</td>
<td>13.7</td>
</tr>
<tr>
<td>D-ConvNet [41]</td>
<td>73.5</td>
<td>112.3</td>
<td>18.7</td>
</tr>
<tr>
<td>CSRNet [20]</td>
<td>68.2</td>
<td>115.0</td>
<td>10.6</td>
</tr>
<tr>
<td>ic-CNN [33]</td>
<td>69.8</td>
<td>117.3</td>
<td>10.7</td>
</tr>
<tr>
<td>SA-Net (image-based) [4]</td>
<td>88.1</td>
<td>134.3</td>
<td>-</td>
</tr>
<tr>
<td>SA-Net (patch-based) [4]</td>
<td>67.0</td>
<td>104.5</td>
<td>8.4</td>
</tr>
<tr>
<td>ACSCP [39]</td>
<td>75.7</td>
<td>102.7</td>
<td>17.2</td>
</tr>
<tr>
<td>Jian et al. [12]</td>
<td>64.2</td>
<td>109.1</td>
<td>8.2</td>
</tr>
<tr>
<td>CG-DRCN (proposed)</td>
<td>64.0</td>
<td>98.4</td>
<td>8.5</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Method</th>
<th>MAE</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Idrees et al. [10]</td>
<td>315.0</td>
<td>508.0</td>
</tr>
<tr>
<td>Zhang et al. [59]</td>
<td>277.0</td>
<td>426.0</td>
</tr>
<tr>
<td>CMTL et al. [43]</td>
<td>252.0</td>
<td>514.0</td>
</tr>
<tr>
<td>Switching-CNN [38]</td>
<td>228.0</td>
<td>445.0</td>
</tr>
<tr>
<td>Idrees et al. [11]</td>
<td>132.0</td>
<td>191.0</td>
</tr>
<tr>
<td>Jian et al. [12]</td>
<td>113.0</td>
<td>188.0</td>
</tr>
<tr>
<td>CG-DRCN (proposed)</td>
<td><strong>112.2</strong></td>
<td><strong>176.3</strong></td>
</tr>
</tbody>
</table>

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.

**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


[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


[57] Beibei Zhan, Dorothy N Monekosso, Paolo Remagnino, Sergio A Velastin, and Li-Qun Xu. Crowd analysis: a survey,


[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