Generating High-Quality Crowd Density Maps using Contextual Pyramid CNNs

Vishwanath A. Sindagi and Vishal M. Patel
Rutgers University, Department of Electrical and Computer Engineering
94 Brett Road, Piscataway, NJ 08854, USA
vishwanath.sindagi@rutgers.edu, vishal.m.patel@rutgers.edu

Abstract

We present a novel method called Contextual Pyramid CNN (CP-CNN) for generating high-quality crowd density and count estimation by explicitly incorporating global and local contextual information of crowd images. The proposed CP-CNN consists of four modules: Global Context Estimator (GCE), Local Context Estimator (LCE), Density Map Estimator (DME) and a Fusion-CNN (F-CNN). GCE is a VGG-16 based CNN that encodes global context and it is trained to classify input images into different density classes, whereas LCE is another CNN that encodes local context information and it is trained to perform patch-wise classification of input images into different density classes. DME is a multi-column architecture-based CNN that aims to generate high-dimensional feature maps from the input image which are fused with the contextual information estimated by GCE and LCE using F-CNN. To generate high resolution and high-quality density maps, F-CNN uses a set of convolutional and fractionally-strided convolutional layers and it is trained along with the DME in an end-to-end fashion using a combination of adversarial loss and pixel-level Euclidean loss. Extensive experiments on highly challenging datasets show that the proposed method achieves significant improvements over the state-of-the-art methods.

1. Introduction

With ubiquitous usage of surveillance cameras and advances in computer vision, crowd scene analysis [18, 43] has gained a lot of interest in the recent years. In this paper, we focus on the task of estimating crowd count and high-quality density maps which has wide applications in video surveillance [15, 41], traffic monitoring, public safety, urban planning [43], scene understanding and flow monitoring. Also, the methods developed for crowd counting can be extended to counting tasks in other fields such as cell microscopy [38, 36, 16, 6], vehicle counting [23, 49, 48, 11, 34], environmental survey [8, 43], etc. The task of crowd counting and density estimation has seen a significant progress in the recent years. However, due to the presence of various complexities such as occlusions, high clutter, non-uniform distribution of people, non-uniform illumination, intra-scene and inter-scene variations in appearance, scale and perspective, the resulting accuracies are far from optimal.

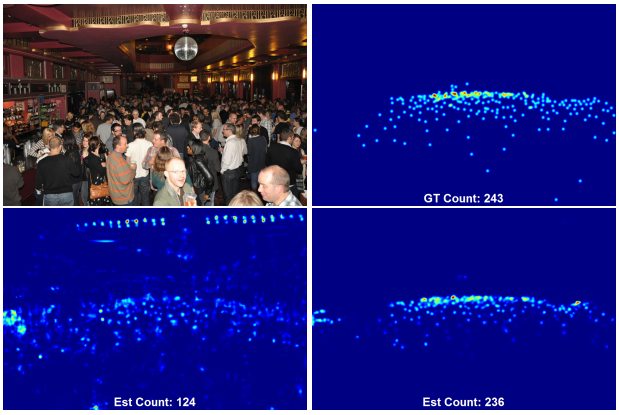

Figure 1: Density estimation results. Top Left: Input image (from the ShanghaiTech dataset [50]). Top Right: Ground truth. Bottom Left: Zhang *et al*. [50] (PSNR: 22.7 dB SSIM: 0.68). Bottom Right: CP-CNN (PSNR: 26.8 dB SSIM: 0.91).

Recent CNN-based methods using different multi-scale architectures [50, 23, 29] have achieved significant success in addressing some of the above issues, especially in the high-density complex crowded scenes. However, these methods tend to under-estimate or over-estimate count in the presence of high-density and low-density crowd images, respectively (as shown in Fig. 2). A potential solution is to use contextual information during the learning process. Several recent works for semantic segmentation [21], scene parsing [51] and visual saliency [52] have demonstrated that incorporating contextual information can provide significant improvements in the results. Motivated by their success, we believe that availability of global context shall aid the learning process and help us achieve better

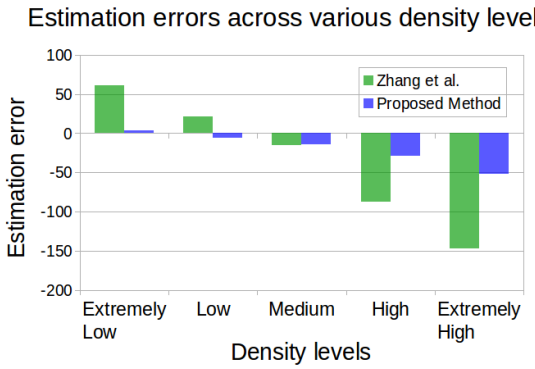

Figure 2: Average estimation errors across various density levels. Current state-of-the-art method [50] overestimates/underestimates count in the presence of low-density/high-density crowd.

count estimation. In addition, existing approaches employ max-pooling layers to achieve minor translation invariance resulting in low-resolution and hence low-quality density maps. Also, to the best of our knowledge, most existing methods concentrate only on the quality of count rather than that of density map. Considering these observations, we propose to incorporate global context into the learning process while improving the quality of density maps.

To incorporate global context, a CNN-based Global Context Estimator (*GCE*) is trained to encode the context of an input image that is eventually used to aid the density map estimation process. *GCE* is a CNN-based on VGG-16 architecture. A Density Map Estimator (*DME*), which is a multi-column architecture-based CNN with appropriate max-pooling layers, is used to transform the image into high-dimensional feature maps. Furthermore, we believe that use of local context in the image will guide the *DME* to estimate better quality maps. To this effect, a Local Context Estimator CNN (*LCE*) is trained on input image patches to encode local context information. Finally, the contextual information obtained by LCE and GCE is combined with the output of DME using a Fusion-CNN (*F-CNN*). Noting that the use of max-pooling layers in *DME* results in low-resolution density maps, *F-CNN* is constructed using a set of fractionally-strided convolutions [22] to increase the output resolution, thereby generating high-quality maps. In a further attempt to improve the quality of density maps, the *F-CNN* is trained using a weighted combination of pixel-wise Euclidean loss and adversarial loss [10]. The use of adversarial loss helps us combat the widely acknowledge issue of blurred results obtained by minimizing only the Euclidean loss [13].

The proposed method uses CNN networks to estimate context at various levels for achieving lower count error and better quality density maps. It can be considered as a set of CNNs to estimate pyramid of contexts, hence, the proposed method is dubbed as Contextual Pyramid CNN (CP-CNN).

To summarize, the following are our main contributions:

- We propose a novel Contextual Pyramid CNN (CP-CNN) for crowd count and density estimation that encodes local and global context into the density estimation process.
- To the best of our knowledge, ours is the first attempt to concentrate on generating high-quality density maps. Also, in contrast to the existing methods, we evaluate the quality of density maps generated by the proposed method using different quality measures such as PSNR/SSIM and report state-of-the-art results.
- We use adversarial loss in addition to Euclidean loss for the purpose of crowd density estimation.
- Extensive experiments are conducted on three highly challenging datasets ([50, 44, 12]) and comparisons are performed against several recent state-of-the-art approaches. Further, an ablation study is conducted to demonstrate the improvements obtained by including contextual information and adversarial loss.

2. Related work

Various approaches have been proposed to tackle the problem of crowd counting in images [12, 5, 16, 44, 50] and videos [2, 9, 26, 7]. Initial research focussed on detection style [17] and segmentation framework [35]. These methods were adversely affected by the presence of occlusions and high clutter in the background. Recent approaches can be broadly categorized into regression-based, density estimation-based and CNN-based methods. We briefly review various methods among these cateogries as follows:

Regression-based approaches. To overcome the issues of occlusion and high background clutter, researchers attempted to count by regression where they learn a mapping between features extracted from local image patches to their counts [3, 27, 6]. These methods have two major components: low-level feature extraction and regression modeling. Using a similar approach, Idrees *et al*. [12] fused count from multiple sources such as head detections, texture elements and frequency domain analysis.

Density estimation-based approaches. While regression-based approaches were successful in addressing the issues of occlusion and clutter, they ignored important spatial information as they were regressing on the global count. Lempitsky *et al*. [16] introduced a new approach of learning a linear mapping between local patch features and corresponding object density maps using regression. Observing that it is difficult to learn a linear mapping, Pham *et al*. in [24] proposed to learn a non-linear mapping between local patch features and density maps using a random forest framework. Many recent approaches have proposed methods based on density map regression [38, 42, 40]. A more comprehensive survey of different crowd counting methods

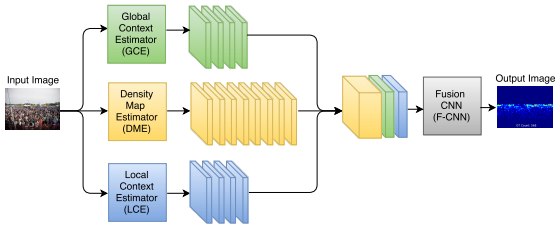

Figure 3: Overview of the proposed CP-CNN architecture. The network incorporates global and local context using *GCE* and *LCE* respectively. The context maps are concatenated with the output of *DME* and further processed by *F-CNN* to estimate high-quality density maps.

can be found in [33, 6, 18, 28].

CNN-based methods. Recent success of CNN-based methods in classification and recognition tasks has inspired researchers to employ them for the purpose of crowd counting and density estimation [37, 44, 36, 30]. Walach *et al*. [36] used CNNs with layered training approach. In contrast to the existing patch-based estimation methods, Shang *et al*. [30] proposed an end-to-end estimation method using CNNs by simultaneously learning local and global count on the whole sized input images. Zhang *et al*. [50] proposed a multi-column architecture to extract features at different scales. Similarly, Onoro-Rubio and López-Sastre in [23] addressed the scale issue by proposing a scale-aware counting model called Hydra CNN to estimate the object density maps. Boominathan *et al*. in [1] proposed to tackle the issue of scale variation using a combination of shallow and deep networks along with an extensive data augmentation by sampling patches from multi-scale image representations. Marsden *et al*. explored fully convolutional networks [19] and multi-task learning [20] for the purpose of crowd counting.

Inspired by cascaded multi-task learning [25, 4], Sindagi *et al*. [32] proposed to learn a high-level prior and perform density estimation in a cascaded setting. In contrast to [32], the work in this paper is specifically aimed at reducing overestimation/underestimation of count error by systemically leveraging context in the form of crowd density levels at various levels using different networks. Additionally, we incorporate several elements such as local context and adversarial loss aimed at improving the quality of density maps. Most recently, Sam *et al*. [29] proposed a Switching-CNN network that intelligently chooses the most optimal regressor among several independent regressors for a particular input patch. A comprehensive survey of recent cnn-based methods for crowd counting can be found in [33]. Recent works using multi-scale and multi-column architectures [50, 23, 36] have demonstrated considerable success in achieving lower count errors. We make the following observations regarding these recent state-of-the-art approaches: 1. These methods do not explicitly incorporate contextual information which is essential for achieving further improvements. 2. Though existing approaches regress on density maps, they are more focussed on improving count errors rather than quality of the density maps, and 3. Existing CNN-based approaches are trained using a pixel-wise Euclidean loss which results in blurred density maps. In view of these observations, we propose a novel method to learn global and local contextual information from images for achieving better count estimates and high-quality density maps. Furthermore, we train the CNNs in a Generative Adversarial Network (GAN) based framework [10] to exploit the recent success of adversarial loss to achieve high-quality and sharper density maps.

3. Proposed method (CP-CNN)

The proposed CP-CNN method consists of a pyramid of context estimators and a Fusion-CNN as illustrated in Fig. 3. It consists of four modules: *GCE*, *LCE*, *DME*, and *F-CNN*. *GCE* and *LCE* are CNN-based networks that encode global and local context present in the input image respectively. *DME* is a multi-column CNN that performs the initial task of transforming the input image to high-dimensional feature maps. Finally, *F-CNN* combines contextual information from *GCE* and *LCE* with high-dimensional feature maps from *DME* to produce high-resolution and high-quality density maps. These modules are discussed in detail as follows.

3.1. Global Context Estimator (GCE)

As discussed in Section 1, though recent state-of-the-art multi-column or multi-scale methods [50, 23, 36] achieve significant improvements in the task of crowd count estimation, they either underestimate or overestimate counts in high-density and low-density crowd images respectively (as explained in Fig. 2). We believe it is important to explicilty model context present in the image to reduce the estimation error. To this end, we associate global context with the level of density present in the image by considering the task of learning global context as classifying the input image into five different classes: extremely low-density (ex-lo), low-density (lo), medium-density (med), high-density (hi) and extremely high-density (ex-hi). Note that the number of classes required is dependent on the crowd density variation in the dataset. A dataset containing large variations may require higher number of classes. In our experiments, we obtained significant improvements using five categories of density levels.

In order to learn the classification task, a VGG-16 [31] based network is fine-tuned with the crowd training data. Network used for *GCE* is as shown in Fig. 4. The convolutional layers from the VGG-16 network are retained, however, the last three fully connected layers are replaced

with a different configuration of fully connected layers in order to cater to our task of classification into five categories. Weights of the last two convolutional layers are fine-tuned while keeping the weights fixed for the earlier layers. The use of pre-trained VGG network results in faster convergence as well as better performance in terms of context estimation.

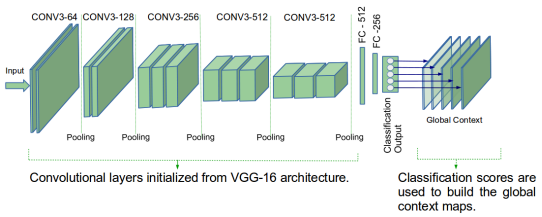

Figure 4: Global context estimator based on VGG-16 architecture. The network is trained to classify the input images into various density levels thereby encoding the global context present in the image.

3.2. Local Context Estimator (LCE)

Existing methods for crowd density estimation have primarily focussed on achieving lower count errors rather than estimating better quality density maps. As a result, these methods produce low-quality density maps as shown in Fig. 1. After an analysis of these results, we believe that some kind of local contextual information can aid us to achieve better quality maps. To this effect, similar to *GCE*, we propose to learn an image's local context by learning to classify it's local patches into one of the five classes: {ex-lo, lo, med, hi, ex-hi}. The local context is learned by the *LCE* whose architecture shown in Fig. 5. It is composed of a set of convolutional and max-pooling layers followed by 3 fully connected layers with appropriate drop-out layers after the first two fully connected layers. Every convolutional and fully connected layer is followed by a ReLU layer except for the last fully connected layer which is followed by a sigmoid layer.

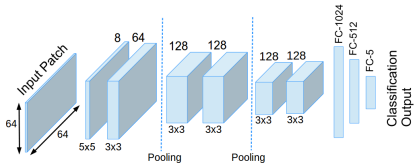

Figure 5: Local context estimator: The network is trained to classify local input patches into various density levels thereby encoding the local context present in the image.

3.3. Density Map Estimator (DME)

The aim of *DME* is to transform the input image into a set of high-dimensional feature maps which will be concatenated with the contextual information provided by *GCE* and *LCE*. Estimating density maps from high-density crowd images is especially challenging due to the presence of heads with varying sizes in and across images. Previous works on multi-scale [23] or multi-column [50] architectures have demonstrated abilities to handle the presence of considerably large variations in object sizes by achieving significant improvements in such scenarios. Inspired by the success of these methods, we use a multi-column architecture similar to [50]. However, notable differences compared to their work are that our columns are much deeper and have different number of filters and filter sizes that are optimized for lower count estimation error. Also, in this work, the multi-column architecture is used to transform the input into a set of high-dimensional feature map rather than using them directly to estimate the density map. Network details for *DME* are illustrated in Fig. 6.

It may be argued that since the *DME* has a pyramid of filter sizes, one may be able to increase the filter sizes and number of columns to address larger variation in scales. However, note that addition of more columns and the filter sizes will have to be decided based on the scale variation present in the dataset, resulting in new network designs that cater to different datasets containing different scale variations. Additionally, deciding the filter sizes will require time consuming experiments. With our network, the design remains consistent across all datasets, as the context estimators can be considered to perform the task of coarse crowd counting.

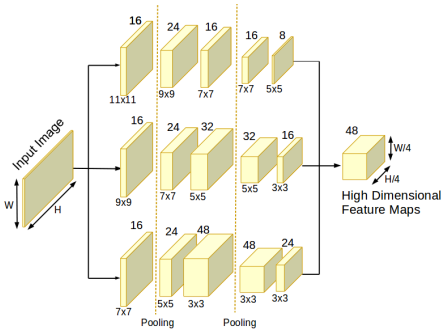

Figure 6: Density Map Estimator: Inspired by Zhang *et al.* [50], DME is a multi-column architecture. In contrast to [50], we use slightly deeper columns with different number of filters and filter sizes.

3.4. Fusion-CNN (F-CNN)

The contextual information from *GCE* and *LCE* are combined with the high-dimensional feature maps from *DME* using *F-CNN*. The *F-CNN* automatically learns to incorporate the contextual information estimated by context estimators. The presence of max-pooling layers in the *DME* network (which are essential to achieve translation invariance) results in down-sampled feature maps and loss of details.

Since, the aim of this work is to estimate high-resolution and high-quality density maps, *F-CNN* is constructed using a set of convolutional and fractionally-strided convolutional layers. The set of fractionally-strided convolutional layers help us to restore details in the output density maps. The following structure is used for *F-CNN*: *CR(64,9)-CR(32,7)-TR(32)-CR(16,5)-TR(16)-C(1,1)*, where, C is convolutional layer, R is ReLU layer, T is fractionally-strided convolution layer and the first number inside every brace indicates the number of filters while the second number indicates filter size. Every fractionally-strided convolution layer increases the input resolution by a factor of 2, thereby ensuring that the output resolution is the same as that of input.

Once the context estimators are trained, *DME* and *F-CNN* are trained in an end-to-end fashion. Existing methods for crowd density estimation use Euclidean loss to train their networks. It has been widely acknowledged that minimization of L_2 error results in blurred results especially for image reconstruction tasks [13, 14, 45, 46, 47]. Motivated by these observations and the recent success of GANs for overcoming the issues of L2-minimization [13], we attempt to further improve the quality of density maps by minimizing a weighted combination of pixel-wise Euclidean loss and adversarial loss. The loss for training *F-CNN* and *DME* is defined as follows:

$$L_T = L_E + \lambda_a L_A, \tag{1}$$

$$L_E = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\|\phi(X^{w,h}) - (Y^{w,h})\|_2, \tag{2}$$

$$L_A = -\log(\phi_D(\phi(X)), \tag{3}$$

where, L_T is the overall loss, L_E is the pixel-wise Euclidean loss between estimated density map and it's corresponding ground truth, λ_a is a weighting factor, L_A is the adversarial loss, X is the input image of dimensions $W \times H$, Y is the ground truth density map, ϕ is the network consisting of *DME* and *F-CNN* and ϕ_D is the discriminator sub-network for calculating the adversarial loss. Following structure is used for the discriminator sub-network: *CP(64)-CP(128)-M-CP(256)-M-CP(256)-CP(256)-M-C(1)-Sigmoid*, where C represents convolutional layer, P represents PReLU layer and M is max-pooling layer.

4. Training and evaluation details

In this section, we discuss details of the training and evaluation procedures.

Training details: Let D be the original training dataset. Patches $1/4^{th}$ the size of original image are cropped from 100 random locations from every image in D. Other augmentation techniques like horizontal flipping and noise addition are used to create another 200 patches. The random cropping and augmentation resulted in a total of 300 patches per image in the training dataset. Let this set of images be called as D_{dme}. Another training set D_{lc} is formed by cropping patches of size 64×64 from 100 random locations in every training image in D.

GCE is trained using the dataset D_{dme}. The corresponding ground truth categories for each image is determined based on the number of people present in it. Note that the images are resized to 224×224 before feeding them into the VGG-based *GCE* network. The network is then trained using the standard cross-entropy loss. *LCE* is trained using the 64×64 patches in D_{lc}. The ground truth categories of the training patches is determined based on the number of people present in them. The network is then trained using the standard cross-entropy loss.

Next, the *DME* and *F-CNN* networks are trained in an end-to-end fashion using input training images from D_{dme} and their corresponding global and local contexts[1]. The global context (F^i_{gc}) for an input training image X^i is obtained in the following way. First, an empty global context F^i_{gc} of dimension $5 \times W_i/4 \times H_i/4$ is created, where $W_i \times H_i$ is the dimension of X_i. Next, a set of classification scores $y^{i,j}_{gc}(j = 1...5)$ is obtained by feeding X_i to *GCE*. Each feature map in global context $F^{i,j}_{gc}$ is then filled with the corresponding classification score $y^{i,j}_{g}$. The local context (F^i_{lc}) for X^i is obtained in the following way. An empty local context F^i_{lc} of dimension $5 \times W_i \times H_i$ is first created. A sliding window classifier (*LCE*) of size 64×64 is run on X_i to obtain the classification score $y^{i,j,w}_{lc}(j = 1...5)$ where w is the window location. The classification scores $y^{i,j,w}_{lc}$ are used to fill the corresponding window location w in the respective local context map $F^{i,j}_{gc}$. $F^{i,j}_{gc}$ is then resized to a size of $W_i/4 \times H_i/4$. After the context maps are estimated, X_i is fed to *DME* to obtain a high-dimensional feature map F^i_{dme} which is concatenated with F^i_{gc} and F^i_{lc}. These concatenated feature maps are then fed into *F-CNN*. The two CNNs (*DME* and *F-CNN*) are trained in an end-to-end fashion by minimizing the weighted combination of pixel-wise Euclidean loss and adversarial loss (given by (1)) between the estimated and ground truth density maps.

Inference details: Here, we describe the process to estimate the density map of a test image X^t_i. First, the global context map F^i_{tgc} for X^t_i is calculated in the following way. The test image X^t_i is divided into non-overlapping blocks of size $W^t_i/4 \times H^t_i/4$. All blocks are then fed into *GCE* to obtain their respective classification scores. As in training, the classification scores are used to build the context maps for each block to obtain the final global context feature map F^i_{tgc}. Next, the local context map F^i_{tlc} for X^t_i is

[1] Once *GCE* and *LCE* are trained, their weights are frozen.

calculated in the following way: A sliding window classifier (*LCE*) of size 64×64 is run across X_i^t and the classification scores from every window are used to build the local context F_{tlc}^i. Once the context information is obtained, X_i^t is fed into *DME* to obtain high-dimensional feature maps F_{tdme}^i. F_{tdme}^i is concatenated with F_{tgc}^i and F_{tlc}^i and fed into *F-CNN* to obtain the output density map. Note that due to additional context processing, inference using the proposed method is computationally expensive as compared to earlier methods such as [50, 29].

5. Experimental results

In this section, we present the experimental details and evaluation results on three publicly available datasets. First, the results of an ablation study conducted to demonstrate the effects of each module in the architecture is discussed. Along with the ablation study, we also perform a detailed comparison of the proposed method against a recent state-of-the-art-method [50]. This detailed analysis contains comparison of count metrics defined by (4), along with qualitative and quantitative comparison of the estimated density maps. The quality of density maps is measured using two standard metrics: PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity in Image [39]). The count error is measured using Mean Absolute Error (MAE) and Mean Squared Error (MSE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|, MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|^2}, \tag{4}$$

where N is number of test samples, y_i is the ground truth count and y_i' is the estimated count corresponding to the i^{th} sample. The ablation study is followed by a discussion and comparison of proposed method's results against several recent state-of-the-art methods on three datasets: ShanghaiTech [50], WorldExpo '10 [44] and UCF_CROWD_50 [12].

5.1. Ablation study using ShanghaiTech Part A

In this section, we perform an ablation study to demonstrate the effects of different modules in the proposed method. Each module is added sequentially to the network and results for each configuration are compared. Following four configurations are evaluated: (1) *DME*: The high-dimensional feature maps of *DME* are combined using 1×1 conv layer whose output is used to estimate the density map. L_E loss is minimized to train the network. (2) *DME* with only *GCE* and *F-CNN*: The output of *DME* is concatenated with the global context. *DME* and *F-CNN* are trained to estimate the density maps by minimizing L_E loss. (3) *DME* with *GCE*, *LCE* and *F-CNN*. In addition to the third configuration, local context is also used in this case and the network is trained using L_E loss. (4) *DME* with *GCE*, *LCE* and *F-CNN* with $L_A + L_E$ (entire network). These results are compared with a fifth configuration: Zhang *et al.* [50] (which is a recent state-of-the-art method) in order to gain a perspective of the improvements achieved by the proposed method and its various modules.

	Count estimation error		Density map quality	
Method	MAE	MSE	PSNR	SSIM
Zhang *et al.*[50]	110.2	173.2	20.91	0.52
DME	104.3	154.2	20.92	0.54
DME+GCE+FCNN	89.9	127.9	20.97	0.61
DME + GCE + LCE + FCNN	76.1	110.2	21.4	0.65
DME+GCE+LCE+ FCNN with L_A+L_E	**73.6**	**106.4**	**21.72**	**0.72**

Table 1: Estimation errors for different configurations of the proposed network on ShanghaiTech Part A[50]. Addition of contextual information and the use of adversarial loss progressively improves the count error and the quality of density maps.

The evaluation is performed on Part A of ShanghaiTech [50] dataset which contains 1198 annotated images with a total of 330,165 people. This dataset consists of two parts: Part A with 482 images and Part B with 716 images. Both parts are further divided into training and test datasets with training set of Part A containing 300 images and that of Part B containing 400 images. Rest of the images are used as test set. Due to the presence of large variations in density, scale and appearance of people across images in the Part A of this dataset, estimating the count with high degree of accuracy is difficult. Hence, this dataset was chosen for the detailed analysis of performance of the proposed architecture.

Count estimation errors and quality metrics of the estimated density images for the various configurations are tabulated in Table 1. We make the following observations: (1) The network architecture for *DME* used in this work is different from Zhang *et al*. [50] in terms of column depths, number of filters and filter sizes. These changes improve the count estimation error as compared to [50]. However, no significant improvements are observed in the quality of density maps. (2) The use of global context in (*DME + GCE + F-CNN*) greatly reduces the count error from the previous configurations. Also, the use of *F-CNN* (which is composed of fractionally-strided convolutional layers), results in considerable improvement in the quality of density maps. (3) The addition of local context and the use of adversarial loss progressively reduces the count error while achieving better quality in terms of PSNR and SSIM.

Estimated density maps from various configurations on sample input images are shown in Fig. 7. It can be observed that the density maps generated using Zhang *et al*. [50] and

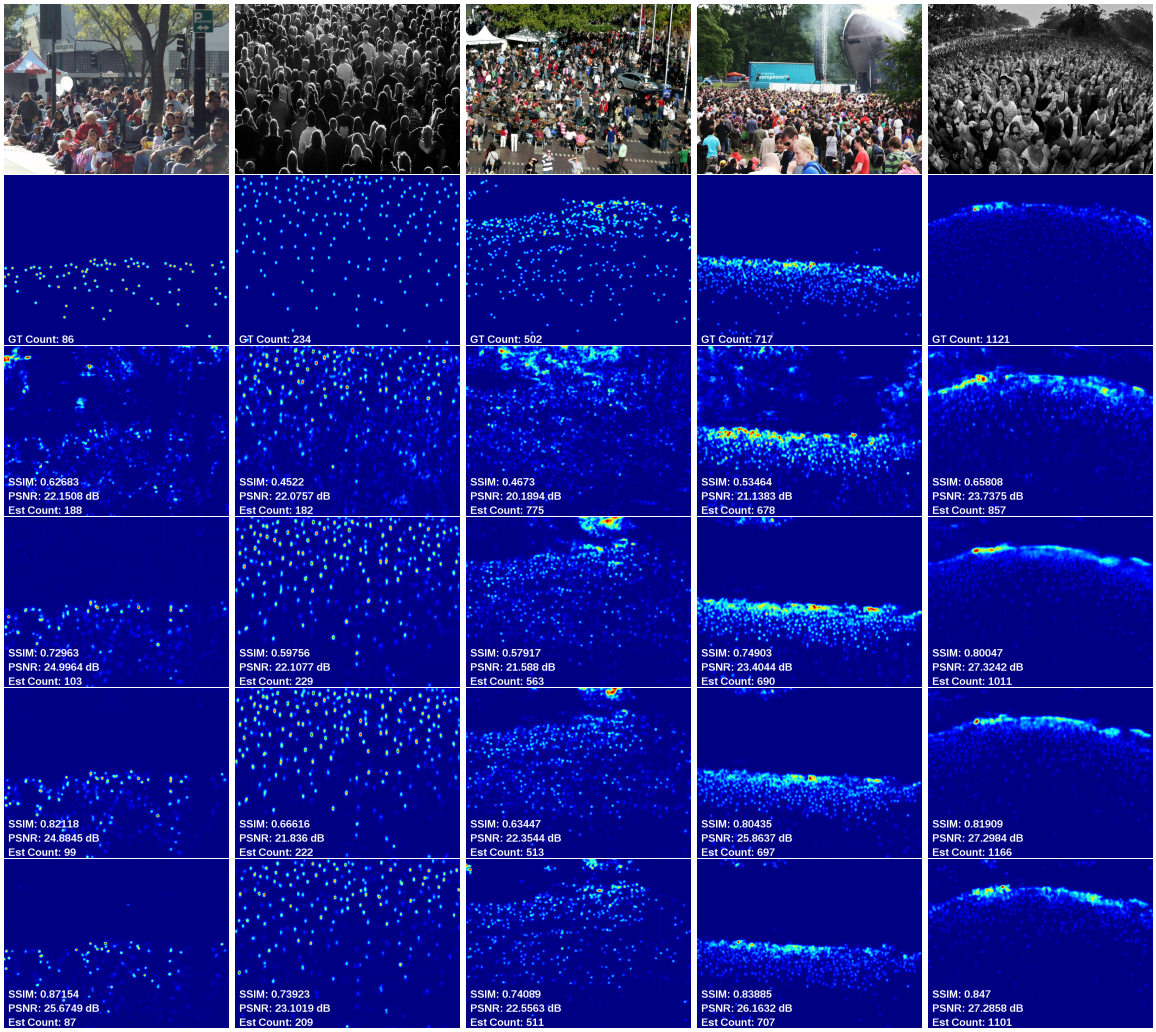

Figure 7: Comparison of results from different configurations of the proposed network along with Zhang *et al*. [50]. Top Row: Sample input images from the ShanghaiTech dataset. Second Row: Ground truth. Third Row: Zhang *et al*. [50]. (Loss of details can be observed). Fourth Row: *DME*. Fifth Row: *DME + GCE + F-CNN*. Sixth Row:*DME + GCE + LCE + F-CNN*. Bottom Row: *DME + GCE + LCE + F-CNN* with adversarial loss. Count estimates and the quality of density maps improve after inclusion of contextual information and adversarial loss.

DME (which regress on low-resolution maps) suffer from loss of details. The use of global context information and fractionally-strided convolutional layers results in better estimation quality. Additionally, the use of local context and minimization over a weighted combination of L_A and L_E further improves the quality and reduces the estimation error.

5.2. Evaluations and comparisons

In this section, the results of the proposed method are compared against recent state-of-the-art methods on three challenging datasets.

ShanghaiTech. The proposed method is evaluated against four recent approaches: Zhang *et al*. [44], MCNN [50], Cascaded-MTL [32] and Switching-CNN [29] on Part A and Part B of the ShanghaiTech dataset are shown in Table 2. The authors in [44] proposed a switchable learning function where they learned their network by alternatively training on two objective functions: crowd count and density estimation. They made use of perspective maps for appropriate ground truth density maps. In another approach, Zhang *et al*. [50] proposed a multi-column convolutional network (MCNN) to address scale issues and a sophisticated ground truth density map generation technique. Instead of using the responses of all the columns, Sam *et al*. [29] proposed a switching-CNN classifier that chooses the optimal regressor. Sindagi *et al*. [32] incorporate high-level prior in the form of crowd density levels and perform a cascaded multi-

task learning of estimating prior and density map. It can be observed from Table 2, that the proposed method is able to achieve superior results as compared to the other methods, which highlights the importance of contextual processing in our framework.

	Part A		Part B	
Method	MAE	MSE	MAE	MSE
Zhang *et al*. [44]	181.8	277.7	32.0	49.8
MCNN [50]	110.2	173.2	26.4	41.3
Cascaded-MTL [32]	101.3	152.4	20.0	31.1
Switching-CNN [29]	90.4	135.0	21.6	33.4
CP-CNN (ours)	**73.6**	**106.4**	**20.1**	**30.1**

Table 2: Estimation errors on the ShanghaiTech dataset.

WorldExpo'10. The WorldExpo'10 dataset was introduced by Zhang *et al*. [44] and it contains 3,980 annotated frames from 1,132 video sequences captured by 108 surveillance cameras. The frames are divided into training and test sets. The training set contains 3,380 frames and the test set contains 600 frames from five different scenes with 120 frames per scene. They also provided Region of Interest (ROI) map for each of the five scenes. For a fair comparison, perspective maps were used to generate the ground truth maps similar to the work of [44]. Also, similar to [44], ROI maps are considered for post processing the output density map generated by the network.

The proposed method is evaluated against five recent state-of-the-art approaches: Chen *et al*. [5], Zhang *et al*. [44], MCNN [50], Shang *et al*. [30] and Switching-CNN [29] is presented in Table 3. The authors in [5] introduced cumulative attributive concept for learning a regression model for crowd density and age estimation. Shang *et al*. [30] proposed an end-to-end CNN architecture consisting of three parts: pre-trained GoogLeNet model for feature generation, long short term memory (LSTM) decoders for local count and fully connected layers for the final count. It can be observed from Table 3 that the proposed method outperforms existing approaches on an average while achieving comparable performance in individual scene estimations.

Method	Scene1	Scene2	Scene3	Scene4	Scene5	Avgerage
Chen *et al*. [5]	**2.1**	55.9	**9.6**	11.3	**3.4**	16.5
Zhang *et al*. [44]	9.8	**14.1**	14.3	22.2	3.7	12.9
MCNN [50]	3.4	20.6	12.9	13.0	8.1	11.6
Shang *et al*. [30]	7.8	15.4	14.9	11.8	5.8	11.7
Switching-CNN [29]	4.4	15.7	10.0	11.0	5.9	9.4
CP-CNN (ours)	2.9	14.7	10.5	**10.4**	5.8	**8.86**

Table 3: Average estimation errors on the WorldExpo'10 dataset.

UCF_CC_50. The UCF_CC_50 is an extremely challenging dataset introduced by Idrees *et al*. [12]. The dataset contains 50 annotated images of different resolutions and aspect ratios crawled from the internet. There is a large variation in densities across images. Following the standard protocol discussed in [12], a 5-fold cross-validation was performed for evaluating the proposed method. Results are compared with seven recent approaches: Idrees *et al*. [12], Zhang *et al*. [44], MCNN [50], Onoro *et al*. [23], Walach *et al*. [36], Cascaded-MTL [32] and Switching-CNN [29]. The authors in [12] proposed to combine information from multiple sources such as head detections, Fourier analysis and texture features (SIFT). Onoro *et al*. in [23] proposed a scale-aware CNN to learn a multi-scale non-linear regression model using a pyramid of image patches extracted at multiple scales. Walach *et al*. [36] proposed a layered approach of learning CNNs for crowd counting by iteratively adding CNNs where every new CNN is trained on residual error of the previous layer. It can be observed from Table 4 that our network achieves the lowest MAE and MSE count errors. This experiment clearly shows the significance of using context especially in images with widely varying densities.

Method	MAE	MSE
Idrees *et al*. [12]	419.5	541.6
Zhang *et al*. [44]	467.0	498.5
MCNN [50]	377.6	509.1
Onoro *et al*. [23] Hydra-2s	333.7	425.2
Onoro *et al*. [23] Hydra-3s	465.7	371.8
Walach *et al*. [36]	364.4	341.4
Cascaded-MTL [32]	322.8.4	341.4
Switching-CNN [29]	318.1	439.2
CP-CNN (ours)	**295.8**	**320.9**

Table 4: Estimation errors on the UCF_CC_50 dataset.

6. Conclusion

We presented contextual pyramid of CNNs for incorporating global and local contextual information in an image to generate high-quality crowd density maps and lower count estimation errors. The global and local contexts are obtained by learning to classify the input images and its patches into various density levels. This context information is then fused with the output of a multi-column DME by a Fusion-CNN. In contrast to the existing methods, this work focuses on generating better quality density maps in addition to achieving lower count errors. In this attempt, the Fusion-CNN is constructed with fractionally-strided convolutional layers and it is trained along with the DME in an end-to-end fashion by optimizing a weighted combination of adversarial loss and pixel-wise Euclidean loss. Extensive experiments performed on challenging datasets and comparison with recent state-of-the-art approaches demonstrated the significant improvements achieved by the proposed method.

Acknowledgement

This work was supported by US Office of Naval Research (ONR) Grant YIP N00014-16-1-3134.

References

[1] L. Boominathan, S. S. Kruthiventi, and R. V. Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 640–644. ACM, 2016. 3

[2] G. J. Brostow and R. Cipolla. Unsupervised bayesian detection of independent motion in crowds. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 594–601. IEEE, 2006. 2

[3] A. B. Chan and N. Vasconcelos. Bayesian poisson regression for crowd counting. In *2009 IEEE 12th International Conference on Computer Vision*, pages 545–551. IEEE, 2009. 2

[4] J.-C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa. A cascaded convolutional neural network for age estimation of unconstrained faces. In *International Conference on BTAS*, pages 1–8. IEEE, 2016. 3

[5] K. Chen, S. Gong, T. Xiang, and C. Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2467–2474, 2013. 2, 8

[6] K. Chen, C. C. Loy, S. Gong, and T. Xiang. Feature mining for localised crowd counting. In *European Conference on Computer Vision*, 2012. 1, 2, 3

[7] S. Chen, A. Fern, and S. Todorovic. Person count localization in videos from noisy foreground and detections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1364–1372, 2015. 2

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

[9] W. Ge and R. T. Collins. Marked point processes for crowd counting. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2913–2920. IEEE, 2009. 2

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2, 3

[11] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu. Drone-based object counting by spatially regularized regional proposal networks. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[12] H. Idrees, I. Saleemi, C. Seibert, and M. 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. 2, 6, 8

[13] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2017. 2, 5

[14] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[15] D. Kang, Z. Ma, and A. B. Chan. Beyond counting: Comparisons of density maps for crowd analysis tasks-counting, detection, and tracking. *arXiv preprint arXiv:1705.10118*, 2017. 1

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

[17] M. Li, Z. Zhang, K. Huang, and T. 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] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan. Crowded scene analysis: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(3):367–386, 2015. 1, 3

[19] M. Marsden, K. McGuiness, S. Little, and N. E. O'Connor. Fully convolutional crowd counting on highly congested scenes. *arXiv preprint arXiv:1612.00220*, 2016. 3

[20] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor. Resnetcrowd: A residual deep learning architecture for crowd counting, violent behaviour detection and crowd density level classification. *arXiv preprint arXiv:1705.10698*, 2017. 3

[21] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 1

[22] H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1520–1528, 2015. 2

[23] D. Onoro-Rubio and R. J. López-Sastre. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision*, pages 615–629. Springer, 2016. 1, 3, 4, 8

[24] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. 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

[25] R. Ranjan, V. Patel, and R. Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE transactions on PAMI*, 2016. 3

[26] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert. Density-aware person detection and tracking in crowds. In *2011 International Conference on Computer Vision*, pages 2423–2430. IEEE, 2011. 2

[27] D. Ryan, S. Denman, C. Fookes, and S. Sridharan. Crowd counting using multiple local features. In *Digital Image Computing: Techniques and Applications, 2009. DICTA'09.*, pages 81–88. IEEE, 2009. 2

[28] S. A. M. Saleh, S. A. Suandi, and H. Ibrahim. Recent survey on crowd density estimation and counting for visual surveillance. *Engineering Applications of Artificial Intelligence*, 41:103–114, 2015. 3

[29] D. B. Sam, S. Surya, and R. V. Babu. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1, 3, 6, 7, 8

[30] C. Shang, H. Ai, and B. Bai. End-to-end crowd counting via joint learning local and global count. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1215–1219. IEEE, 2016. 3, 8

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 3

[32] V. A. Sindagi and V. 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. 3, 7, 8

[33] V. A. Sindagi and V. M. Patel. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 2017. 3

[34] E. Toropov, L. Gui, S. Zhang, S. Kottur, and J. M. Moura. Traffic flow from a low frame rate city camera. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 3802–3806. IEEE, 2015. 1

[35] P. Tu, T. Sebastian, G. Doretto, N. Krahnstoever, J. Rittscher, and T. Yu. Unified crowd segmentation. In *European Conference on Computer Vision*, pages 691–704. Springer, 2008. 2

[36] E. Walach and L. Wolf. Learning to count with cnn boosting. In *European Conference on Computer Vision*, pages 660–676. Springer, 2016. 1, 3, 8

[37] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao. Deep people counting in extremely dense crowds. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1299–1302. ACM, 2015. 3

[38] Y. Wang and Y. Zou. Fast visual object counting via example-based density estimation. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 3653–3657. IEEE, 2016. 1, 2

[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 6

[40] F. Xia and S. Zhang. Block-coordinate frank-wolfe optimization for counting objects in images. 2

[41] F. Xiong, X. Shi, and D.-Y. Yeung. Spatiotemporal modeling for crowd counting in videos. In *IEEE International Conference on Computer Vision*. IEEE, 2017. 1

[42] B. Xu and G. 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

[43] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu. Crowd analysis: a survey. *Machine Vision and Applications*, 19(5-6):345–357, 2008. 1

[44] C. Zhang, H. Li, X. Wang, and X. 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. 2, 3, 6, 7, 8

[45] H. Zhang and K. Dana. Multi-style generative network for real-time transfer. *arXiv preprint*, 2017. 5

[46] H. Zhang, V. A. Sindagi, and V. M. Patel. Image de-raining using a conditional generative adversarial network. *arXiv preprint arXiv:1701.05957*, 2017. 5

[47] H. Zhang, V. A. Sindagi, and V. M. Patel. Joint transmission map estimation and dehazing using deep networks. *arXiv preprint arXiv:1701.05957*, 2017. 5

[48] S. Zhang, G. Wu, J. P. Costeira, and J. M. Moura. Understanding traffic density from large-scale web camera data. In *IEEE Computer Vision and Pattern Recognition*. IEEE, 2017. 1

[49] S. Zhang, G. Wu, J. P. Costeira, and J. M. F. Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *IEEE International Conference on Computer Vision*. IEEE, 2017. 1

[50] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. 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, 4, 6, 7, 8

[51] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[52] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015. 1