Crowd Counting via Segmentation Guided Attention Networks and Curriculum Loss

Qian Wang, *Member, IEEE,* Toby P. Breckon, *Senior Member, IEEE,*

***Abstract*—Automatic crowd behaviour analysis is an important task for intelligent transportation systems to enable effective flow control and dynamic route planning for varying road participants. Crowd counting is one of the keys to automatic crowd behaviour analysis. Crowd counting using deep convolutional neural networks (CNN) has achieved encouraging progress in recent years. Researchers have devoted much effort to the design of variant CNN architectures and most of them are based on the pre-trained VGG16 model. Due to the insufficient expressive capacity, the backbone network of VGG16 is usually followed by another cumbersome network specially designed for good counting performance. Although VGG models have been outperformed by Inception models in image classification tasks, the existing crowd counting networks built with Inception modules still only have a small number of layers with basic types of Inception modules. To fill in this gap, in this paper, we firstly benchmark the baseline Inception-v3 model on commonly used crowd counting datasets and achieve surprisingly good performance comparable with or better than most existing crowd counting models. Subsequently, we push the boundary of this disruptive work further by proposing a Segmentation Guided Attention Network (SGANet) with Inception-v3 as the backbone and a novel curriculum loss for crowd counting. We conduct thorough experiments to compare the performance of our SGANet with prior arts and the proposed model can achieve state-of-the-art performance with MAE of 57.6, 6.3 and 87.6 on ShanghaiTechA, ShanghaiTechB and UCF_QNRF, respectively.**

***Index Terms*—Crowd counting, Curriculum loss, Inception-v3, Segmentation guided attention networks**

I. Introduction

AUTOMATIC crowd counting has attracted increasing attention in the research community since its valuable impacts in public surveillance and intelligent transportation systems [1], [2], [3], [4], [5]. Crowd behaviour can have a big effect in the efficiency of public transportation. Intelligent transportation systems deployed in a smart city should be able to capture real-time crowd behaviour information from public surveillance and dynamically adjust the planning for effective transportation. Accurate people and vehicle counting in varying conditions provide basic information for automatic crowd behaviour analysis. People and vehicle counting can be formulated in a unified object counting framework which aims to estimate the number of target objects in still images or video frames and has been applied in many real-world applications. For instance, there have been works focusing on automatic counting different objects including cells [6], vehicles [7], [8], leaves [9], [10] and people [4].

In earlier years, crowd counting in images was implemented by detection [11], [12], [13] or direct count regression [14], [15]. Counting by detection methods assume people signatures (i.e. the whole body or the head) in images are detectable and the count can be easily achieved from the detection results. This assumption, however, does not always hold in real scenarios, especially when the crowd is extremely dense. Counting by direct count regression aims to learn a regression model (e.g., support vector machine [15] or neural networks [14]) mapping the hand-crafted image features directly to the count of people in the image. Methods falling into this category only give the final counts hence lack of explainability and reliability.

Recently, crowd counting has been overwhelmingly dominated by density estimation based methods since the idea of density map was first proposed in [16]. The use of deep Convolutional Neural Networks [17] to estimate the density map along with the availability of large-scale datasets [18], [19] further improved the accuracy of crowd counting in more challenging real-world scenarios. Recent works in crowd counting have been focusing on the design of novel architectures of deep neural networks (e.g., multi-column CNN [18], [20] and attention mechanism [21], [22]) for accurate density map estimation. The motivations of these designs are usually to improve the generalization to scale-variant crowd images. Among them, the *Inception* module [23] has been employed and showed effectiveness in crowd counting [24], [25], although only the basic *Inception* modules are used and the networks are relative shallow compared with the state-of-the-art deep CNN models for image classification such as *Inception-v3* [23] which uses heterogeneous *Inception* modules to improve the expressive power of the network. Although VGG16, VGG19 and ResNet101 have been used as the backbone networks for crowd counting in [26], [27], [28], to our best knowledge, the *Inception* models have not been investigated.

In this paper, we make the first attempt to investigate the effectiveness of *Inception-v3* model for crowd counting. We modify the original *Inception-v3* to make it suitable for crowd density estimation. Without bells and whistles, the *Inception-v3* model can achieve surprisingly good performance comparable with or even better than most existing crowd counting models on commonly used crowd counting datasets. Subsequently, we add a segmentation map guided attention layer to the *Inception-v3* model to enhance the salient feature extraction for accurate density map estimation and propose a

Q. Wang is with the Department of Computer Science, Durham University, United Kingdom, e-mail: qian.wang173@hotmail.com

TP. Breckon is with the Department of Computer Science and Department of Engineering, Durham University, United Kingdom, e-mail: toby.breckon@durham.ac.uk

novel curriculum loss strategy to address the issues caused by extremely dense regions in crowd counting. As a result, the proposed SGANet with curriculum loss is able to achieve state-of-the-art performance for crowd counting with the embarrassingly simple design. The contributions of this paper are summarized as follows:

- We make the first attempt to investigate the effectiveness of *Inception-v3* in crowd counting and achieve disruptive results which are important for the research community.
- We present a Segmentation Guided Attention Network (SGANet) with a novel curriculum loss function based on the *Inception-v3* model for crowd counting.
- Extensive evaluations are conducted on benchmark datasets and the results demonstrate the superior performance of SGANet and the effectiveness of curriculum loss in crowd counting.

The remainder of this paper is organized as follows. Section II reviews related work of crowd counting and curriculum learning. In Section III we introduce our proposed segmentation guided attention networks for crowd counting with curriculum loss. Section IV presents the experiments and results on several benchmark datasets and we conclude our work in Section V.

II. Related Work

In this section, we first review related works on CNN based crowd counting and focus mainly on the diverse network architectures against which our proposed crowd counting model is compared. Subsequently, we introduce works related to curriculum learning and how they can potentially be used in the task of crowd counting.

A. Crowd Counting Networks

Successful efforts have been devoted to the design of novel network architectures to improve the performance of crowd counting. Commonly used principles of network design for crowd counting include multi-column networks, rich feature fusion and attention mechanism.

Multi-column neural networks were employed to address the scale-variant issue in crowd counting [18], [29], [30]. As one of the earliest CNN based models for crowd counting, MCNN [18] consists of three branches aiming to handle crowds of different densities. Following this idea, Sam et al. [29] proposed SwitchCNN which employs a classifier to explicitly select one of the three branches for a given input patch based on its level of crowd density. While these methods aim to use different kernel sizes in different branches to capture scale-variant information, Liu et al. [31] proposed a model consisting of multiple branches of VGG16 networks with shared weights to process scaled input images respectively. Similarly, Ranjan et al. [32] devised a two-column network which learns the low- and high-resolution density maps iteratively via two branches of CNN. The success of these specially designed network architectures has validated that multi-column CNN models are capable of capturing scale-variant features for crowd counting.

The second direction of network design is to pursue effective fusion of rich features from different layers [33], [25]. These attempts are based on the fact different layers have variant receptive fields hence capturing features of variant-scale information. Different feature fusion strategies including direct fusion [33], top-down fusion [34] and bidirectional fusion [35] have been employed in crowd counting.

To take advantage of the two aforementioned ideas for crowd counting, one straightforward solution is to utilise the *Inception* module [23] which was firstly proposed in [36] and has evolved into a variety of more efficient forms to date. The *Inception* modules have been employed in crowd counting models before in SANet [24] and the TEDNet [25]. Both of them use only the basic types of *Inception* modules similar to those used in the first version of *Inception* net (i.e. GoogLeNet [36]). In our work, we aim to explore the more advanced *Inception* modules in the framework of *Inception-v3*.

The attention mechanism is another useful technique considered when designing network architectures for crowd counting [21], [33], [26], [37]. Attention layers are usually combined with multi-column structures so that regions of different semantic information (e.g., background, sparse, dense, etc.) can be attended and processed by different branches respectively. Attention maps learned by these models have proved to be aware of semantic regions [26], however, they cannot provide fine-grained scale awareness within the images. To address this issue explicitly, perspective maps have been employed to guide the accurate estimation of density maps [38], [39], [40]. In many scenarios where the perspective maps are not available, it is possible to estimate these perspective maps from the crowd images via a specially designed and trained network [41].

Alternatively, binary segmentation maps generated from point annotation [42] are introduced as additional supervision for the training of crowd counting networks via multi-task learning [42]. In our work, binary segmentation maps are treated as explicit attention maps guiding the learning of salient visual features for density map estimation. In this sense, our work is more related to [43] and [44] in which the segmentation maps are also used as attention maps but in essentially different ways as explained in Section III-B and validated in Section IV-F.

B. Curriculum Learning

Curriculum learning is a strategy of model training (e.g., neural networks) in machine learning and was proposed by Bengio et al. [45]. The idea of curriculum learning can date back to no later than 1993 when Elman [46] proved the benefit of training neural networks to learn a simple grammar by “starting small". The strategy of curriculum learning is inspired by the way how humans learn knowledge from easy concepts to hard abstractions gradually. In the specific case of training a machine learning model, curriculum learning selects easy examples at the beginning of training and allows more difficult ones added to the training set gradually. A curriculum is usually defined as a ranking of training examples by some prior knowledge to determine the level of difficulty of a given example. Jiang et al. [47] extended curriculum learning to a so-called self-paced curriculum learning by integrated the ideas of original curriculum learning and self-paced learning [48] in a unified framework.

In this work, we apply the strategy of curriculum learning in crowd counting to address the issue of large variance of the crowd density in the images. Curriculum learning has been employed for crowd counting in [49] where the curriculum is designed on the *image level*, i.e., a difficulty score is calculated for each training image. The training images are divided into multiple subsets based on their difficulty scores and the easiest subset is added into the training set first. By contrast, our curriculum learning strategy is characterized by a novel curriculum loss defined on the *pixel level* as described in Section III-D. We define that density map pixels of higher values than a *threshold* have higher difficulty scores because these pixels are within regions of denser crowds. We use all training images throughout the training process but set the threshold to a low value at the beginning and increase it gradually so that the difficult pixels become easy ones and contribute more to the training. As a result, our curriculum learning strategy is simple to implement with zero extra cost and has been proved effective especially when there exist extremely dense crowd regions in the images.

III. Segmentation Guided Attention Network

Crowd counting is formulated as a density map regression problem in this study. Given a crowd image I, we aim to learn a Fully Convolutional Network (FCN) denoted as $\mathcal{F}$ so that the corresponding density map $\boldsymbol{M}^{den}$ can be estimated by:

$$\hat{\boldsymbol{M}}^{den} = \mathcal{F}(I; \boldsymbol{\Theta}). \tag{1}$$

where $\boldsymbol{\Theta}$ is a collection of parameters of the FCN.

As shown in Figure 1, our proposed network is adapted from the famous *Inception-v3* originally designed for image classification by Google Research [23]. We first modify *Inception-v3* to an FCN so that it can process images of arbitrary sizes and generates the estimated density maps $\boldsymbol{M}^{den}$ as the outputs. An attention layer is added to the network to filter out features within the background region and concentrate on the foreground features for accurate density map estimation. Since the attention maps generated by this attention layer aim to discriminate the regions of background and foreground of the feature maps, we use a ground truth segmentation map, which can be easily derived from point annotations, as extra guidance for the training of the attention layer. As a result, the learned attention maps are forced to be similar to the segmentation maps during training.

We also investigate the use of curriculum loss in the training of crowd counting networks. Specifically, we define a curriculum based on the pixel-wise difficulty level so that the network starts training by focusing more on the “easy" regions (sparse) within the density maps and down-weighting the “hard" pixels (dense). During training, the “hard" pixels are gradually exposed to the model and finally, the learned model can perform well for all situations.

A. Density and Segmentation Maps

In this study, we use simple ways to generate density and segmentation maps from the point annotations although more complicated ones [44] might benefit the performance. For density maps $\boldsymbol{M}^{den} \in \mathbb{R}^{+H\times W}$, where H and W are the height and width of the image, we follow [18] using a Gaussian kernel $G_\sigma \in \mathbb{R}^{+15\times 15}$ with a kernel size of 15×15 and fixed $\sigma = 4$:

$$\boldsymbol{M}^{den}(\boldsymbol{x}) = H(\boldsymbol{x}) * G_\sigma(\boldsymbol{x}), \tag{2}$$

where $H(\boldsymbol{x}) = \sum_{i=1}^{N} \delta(\boldsymbol{x} - \boldsymbol{x}_i)$, N is the number of point annotations in the image and $\delta(\cdot)$ is the Delta function. As a result, $H(\boldsymbol{x})$ is a binary matrix of the same size as the image and only has values of ones at the positions of point annotations. The density map is derived by the convolution between $H(\boldsymbol{x})$ and the Gaussian kernel $G_\sigma(\boldsymbol{x})$.

For segmentation maps $\boldsymbol{M}^{seg} \in \{0, 1\}^{H\times W}$, we use a similar method:

$$\boldsymbol{M}^{seg}(\boldsymbol{x}) = H(\boldsymbol{x}) * J_n(\boldsymbol{x}), \tag{3}$$

where $J_n(\boldsymbol{x})$ is an all-one matrix of size $n \times n$ centred at the position $\boldsymbol{x}$. As a result, ones and zeros in the matrix $\boldsymbol{M}^{seg}$ denote the pixels belong to the foreground and background regions, respectively. We empirically set $n = 25$ across all our experiments to ensure that a specific head within an image is characterized by more pixels in the segmentation map than in the density map to avoid losing useful contextual information. Our choice of $n = 25$ cannot guarantee precise foreground segmentation maps due to the head-scale variance. A larger value would affect the discrimination of the foreground and background in crowded regions. Our experimental results demonstrate the our choice is suitable and can benefits the density map estimation.

B. Network Configuration

Instead of designing a novel network from scratch, we exploit the state-of-the-art CNN model for image classification *Inception-v3* in our study. To apply the original *Inception-v3* network in crowd counting, some favourable modifications have been made. Firstly, we remove the final fully-connected layers and reserve all the convolutional layers. The input size of the original *Inception-v3* network is 299×299 and the output size of the final convolutional layer is 8×8. That is to say, feature maps generated by the last convolutional layer have approximately $\frac{1}{2^5}$ spatial resolutions of the input image. This is achieved by the first convolutional layers (stride of 2), two max-pooling layers (stride of 2) and two *Inception* modules in which max-pooling (stride of 2) is employed. To ensure the outputs of the network (i.e. estimated density maps) have sufficient spatial resolutions, we remove the first two max-pooling layers from the original network and add one upsample layer before the final *Inception* module. As a result, the output of the modified network has exactly $\frac{1}{4}$ spatial resolution of the input image when the input size is 2^n (e.g., 128×128 in our case). Such modification does not change the number of parameters of the network hence the pre-trained weights can be directly loaded and used. However, since the spatial resolutions of intermediate feature maps have been increased, the number of operations is also increased. This modified model will also denoted as *Inception-v3* without introducing ambiguity and used as a baseline method in our experiments.

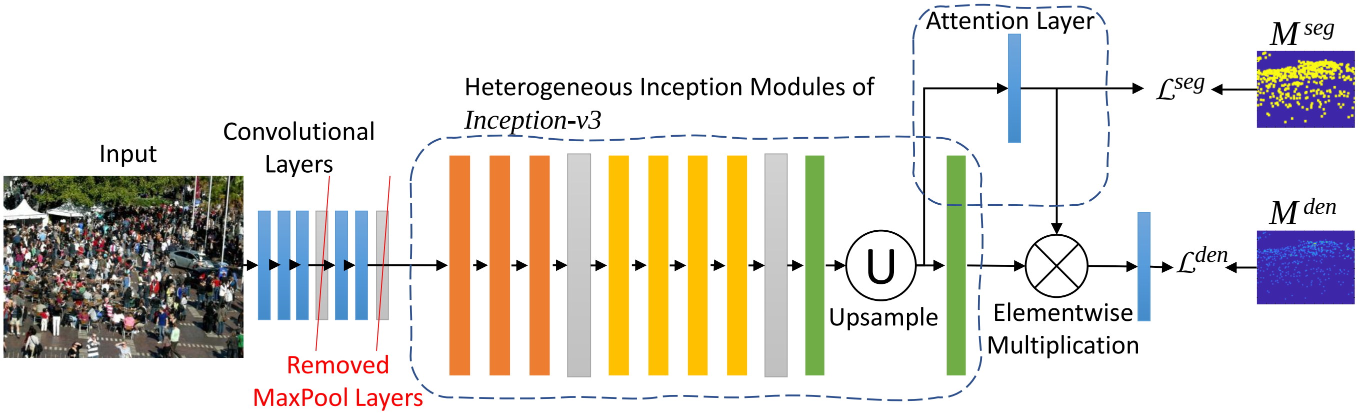


Fig. 1. The framework of our proposed Segmentation Guided Attention Network (SGANet) which is adapted from *Inception-v3* by: 1) removing the fully-connected layers; 2) removing two maxpooling layers to reserve high spatial resolution feature maps; 3) adding an upsampling layer before the last *Inception* module; 4) adding an attention layer whose output is applied to the feature maps generated by the last *Inception* module; 5) adding a convolutional layer for density map estimation.

Distinct from existing works using the segmentation map in the framework of multi-task learning [42] to extract more salient features for density map estimation, we claim that the segmentation map can be used as an ideal attention map to emphasize the contributions of features within the foreground regions to the density map estimation whilst compressing the effects of features within the background regions. To this end, we add an attention layer to estimate the attention map. The attention layer is a convolutional layer followed by a sigmoid layer which restricts the output values in the range of 0–1. The attention layer takes the feature maps generated by the second last *Inception* module as input and outputs a one-channel attention map of the same spatial resolution as the input. Subsequently, the attention map estimated by the attention layer is applied to the feature map generated by the last *Inception* module by an elementwise multiplication with each channel of the feature map.

$$F^l = F^{l-1} \odot M_{att}, \tag{4}$$

where $\odot$ denotes the operation of element-wise product.

The attention layer designed in our framework is similar to that in [43], [44]. However, a so-called inverse attention map is estimated in [43] while our attention layer generates an attention map directly applied to the feature map. Also, the foreground regions in the ground truth segmentation map in [43] are derived by thresholding the density map hence both maps have the same positive fields for each head while ours are different (c.f. Eq.(2-3)). In [44], the attention layer takes the feature map as input to estimate an attention map which again is applied to the same feature map. This may limit the capacity of the model since it is forced to learn two different maps from the same feature map via two convolutional layers which have limited parameters. In contrast, as mentioned above, the input of our attention layer is the feature map from the previous layer which has higher spatial resolutions and is different from the one the generated attention map will be applied to. These favourable distinctions collectively benefit the estimation of the density map and will be empirically evaluated in our experiments.

C. Loss Function

We first describe the loss function used to train the SGANet without curriculum loss in this section and describe the curriculum loss in the following section. The loss function consists of two components. The first one is the Mean Squared Error (MSE) loss applied to the estimation of the density map and is denoted as $\mathcal{L}^{den}$. The density map loss can be calculated as follows:

$$\mathcal{L}^{den}(\boldsymbol{\Theta}) = \frac{1}{2N}\sum_{i=1}^{N} ||\hat{\boldsymbol{M}}_i^{den} - \boldsymbol{M}_i^{den}||_F^2, \tag{5}$$

where $||\cdot||_F^2$ is a Frobenius norm of a matrix. The second component of the loss function is the segmentation map loss $\mathcal{L}^{seg}$ which is defined as the cross-entropy loss:

$$\mathcal{L}^{seg}(\boldsymbol{\Theta}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j,k} \boldsymbol{H}_i(j,k), \tag{6}$$

$$\begin{aligned} \boldsymbol{H}_i = & \boldsymbol{M}_i^{seg} \odot log(\hat{\boldsymbol{M}}_i^{seg}) \\ & + (1 - \boldsymbol{M}_i^{seg}) \odot log(1 - \hat{\boldsymbol{M}}_i^{seg}), \end{aligned} \tag{7}$$

where $\odot$ denotes elementwise multiplication of two matrices with the same size and $\boldsymbol{H}(j,k)$ denotes an element of the matrix $\boldsymbol{H}$. These two components are combined during network training and the compositional loss function is:

$$\mathcal{L}(\boldsymbol{\Theta}) = \mathcal{L}^{den}(\boldsymbol{\Theta}) + \lambda\mathcal{L}^{seg}(\boldsymbol{\Theta}) \tag{8}$$

where λ is a hyper-parameter which ensures the two components to have comparable values and is set 20 across our experiments.

D. Curriculum Loss

To benefit from the strategy of curriculum learning, we present a novel curriculum loss function in this section to replace the traditional density map loss function defined in Eq. (5). The curriculum loss function is designed to be aware of the pixel-wise difficulty level when computing the density map loss. Based on the fact that dense crowds are generally more

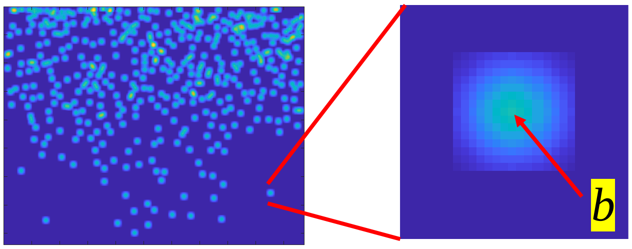


Fig. 2. An illustration of how to determine the value of b in Eq.(10).

difficult to count than sparse ones, we design a curriculum where pixels of higher values than a dynamic threshold in the density map are defined as difficult pixels. We set the dynamic threshold and assign variant weights to different pixels of the density map when calculating the density map loss. Specifically, we define a weight matrix $\boldsymbol{W}$ as follows:

$$\boldsymbol{W} = \frac{T(e)}{max\{\boldsymbol{M}^{den} - T(e), 0\} + T(e)}. \tag{9}$$

The weight matrix $\boldsymbol{W}$ has the same size as the density map matrix $\boldsymbol{M}^{den}$ used in Eq.(5) and the pixel-wise weights are determined by the dynamic threshold $T(e)$ and the pixel values in the density map. If the pixel value of the density map is higher than the threshold, this pixel is treated as a difficult one and the corresponding weight is set less than one, otherwise the weight is equal to one. The higher the pixel values are, the smaller the weights will be. As a result, the training will focus more on the pixels of lower density value than $T(e)$.

The dynamic threshold $T(e)$ is defined as a function of the training epoch index e in the form of:

$$T(e) = ke + b \tag{10}$$

where k and b can be determined empirically in the following way. The value of b is the initial threshold when $epoch = 0$ and it is set to be equivalent to the maximum density value in the region characterizing a single head (as shown in Figure 2). As a result, all the density values within the regions where heads are not overlapped are smaller than the threshold $T(e)$ throughout the training (i.e., e = 0, 1, 2, ...). On the other hand, the value of k controls the speed of increasing the difficulty and its value is determined so that $T(e)$ increases to a value higher than the maximum density values before training stops. This guarantees the final weight matrix $\boldsymbol{W}$ has all one values hence all pixels contribute equally to the training. In practice, we just need to find the density value in the center of a single-head region in a ground truth density map to determine b, whilst the maximum pixel value of the density maps in the training data and the number of training epochs collaboratively determine k.

Finally, the curriculum loss function for density map can be derived by modifying Eq.(5) as:

$$\mathcal{L}^{den}(\boldsymbol{\Theta}) = \frac{1}{2N}\sum_{i=1}^{N} ||\boldsymbol{W}(e) \odot (\hat{\boldsymbol{M}}_i^{den} - \boldsymbol{M}_i^{den})||_F^2. \tag{11}$$

where $\boldsymbol{W}(e)$ is also a function with respect to the training epoch index e.

IV. Experiments

Extensive experiments have been conducted on benchmark datasets to evaluate the performance of SGANet and the effectiveness of curriculum loss in crowd counting. We will briefly describe the datasets and evaluation metrics used in our experiments, details of experimental protocols and network training. Experimental results are compared with state-of-the-art methods and analysed. We also present an ablation study to investigate the contributions of different components to the performance of the proposed framework.

A. Datasets

ShanghaiTech dataset was collected and published by Zhang et al. [18] consisting of two parts. Part A consists of 300 and 182 images of different resolutions for training and testing respectively. The minimum and maximum counts are 33 and 3139 respectively, and the average count is 501.4. Part B consists of 400 and 316 images of a unique resolution (768×1024) for training and testing respectively. Compared with part A, the numbers of people in these images are much smaller with the minimum and maximum counts of 9 and 578 respectively, and the average count is 123.6.

UCF_QNRF dataset [19] contains 1,535 high-quality images, among which 1201 images are used for training and 334 images for testing. The minimum and maximum counts are 49 and 12,865 respectively, and the average count is 815.

UCF_CC_50 dataset [50] contains 50 images with the minimum and maximum counts of 94 and 4,534 respectively. It is a challenging dataset due to the limited number of images. Following the suggestion in [50] and many other works, we use 5-fold cross-validation in our experiments.

B. Evaluation Metrics

We follow the previous works using two metrics, i.e., the mean absolute error (MAE) and the root mean squared error (RMSE), to evaluate the performance of different models in our experiments. The two metrics can be calculated as follows:

$$MAE = \frac{1}{N_{test}}\sum_{i=1}^{N_{test}} |y_i - \hat{y}_i| \tag{12}$$

$$RMSE = \sqrt{\frac{1}{N_{test}}\sum_{i=1}^{N_{test}} (y_i - \hat{y}_i)^2} \tag{13}$$

where y_i and $\hat{y}_i$ are the ground truth and predicted count for i-th test image respectively, N_{test} is the number of test images.

C. Network Training

SGANet is implemented in PyTorch [51] and the source code is publicly available [1]. The "Adam" optimizer [52] is employed for training. The initial learning rate is set to 1e-4 and reduced by a factor of 0.5 after every 50 epochs. The total number of training epochs is set 500 since the model can always converge much earlier than that. The network is trained

[1] https://github.com/hellowangqian/sganet-crowd-counting

with image patches with a size of 128×128 randomly cropped from the training images. Instead of preparing the patches in advance, we do the random patch cropping on-the-fly during training. Specifically, we randomly select 8 images from the training set and 4 patches are randomly cropped from each selected image. This leads to a batch of 32 training patches in each iteration of training. The training patches generated in this way can be more diverse and help to alleviate the potential over-fitting problem. Since the output of SGANet has the size of 32×32 (i.e. 1/4 of the input size), we use sum-pooling to adapt the ground truth density and segmentation map so that they have the same size of 32×32 as the output. The training patches, as well as their corresponding density and segmentation maps, are horizontally flipped with a probability of 0.5 for data augmentation which has been shown beneficial in many works [26], [53]. For testing, we feed the whole image into the network and obtain the density map from which the predicted count can be computed. For the UCF_QNRF dataset, to save the memory usage during testing, we also resize the images from both training and test sets so that all images are limited to have their longer sides no higher than 2048 whilst the original aspect ratios are kept, if not specified otherwise.

D. Comparative Study

We select both classic and state-of-the-art models for the comparison, including **MCNN** [18] which is a three-column CNN, **CSRNet** [54] which uses VGG16 as the front-end and dilated convolutional layers as the back-end, **SANet** [24] which employs the basic *Inception* modules but has a relatively shallow depth, **DADNet** [26] which employs the ideas of dilated convolution, attention map and deformable convolution in the framework, **CANNet** [55] which captures context-aware feature by multiple branches, **TEDNet** [25] which also uses *Inception*-style modules, **RANet** [53] which uses an iterative distillation algorithm, **ANF** [57] which uses conditional random fields (CRFs) to aggregate multi-scale features, and **SPANet** [58] which incorporates the spatial context within images into the crowd counting model.

The experimental results are listed in Table I where the best result in each column is highlighted in **bold** and the second best in underscored *italic*. From Table I, we can see our modified *Inception-v3* can achieve very competitive performance on all four datasets. Especially on ShanghaiTech part B, it achieves the second best MAE of 6.4 and the best RMSE of 9.8. On the UCF_QNRF dataset, *Inception-v3* also achieves significantly better results than most existing models including TEDNet (MAE: 95.6 vs 113 and MSE: 165.4 vs 188) which also employs the *Inception* modules. These results demonstrate the superiority of heterogeneous *Inception* modules in classification problems can be transferred to the task of crowd counting hence different *Inception* modules deserve more attention when designing a novel CNN architecture for crowd counting as well as other tasks suffering from the issue of scale variance. On the other hand, the disruptive performance of *Inception-v3* in crowd counting provides more insight for the research community regarding the selection of backbone models when designing novel network architectures for crowd counting.

By adding the segmentation guided attention layer, our SGANet can achieve better performance on all datasets in terms of MAE (i.e. 58.0 vs 60.1 for ShanghaiTech part A, 89.1 vs 95.6 for UCF-QNRF and 224.6 vs 236.0 for UCF-CC-50), although the improvement on ShanghaiTech part B dataset is very marginal (i.e. 6.3 vs 6.4). Regarding RMSE, SGANet achieves better performance on ShanghaiTech part A (i.e. 100.4 vs 105.0) and UCF_QNRF (i.e. 150.6 vs 165.4) but worse results on the other two datasets (i.e. 10.6 vs 9.8 for ShanghaiTech part B and 314.6 vs 304.9 for UCF_CC_50). Overall, our proposed SGANet with the combination of *Inception-v3* and a segmentation guided attention layer can achieve state-of-the-art performance on several benchmark datasets.

The use of curriculum loss (SGANet+CL) further improves the performance of SGANet on three out of four datasets in terms of MAE and these three datasets (i.e. ShanghaiTech part A, UCF_QNRF and UCF_CC_50) consist of crowds with significant density variations. On the ShanghaiTech part B dataset, the use of curriculum loss does not improve the performance because the images from this dataset contain crowds with a relatively small variance of head scales. However, we also observe slight increases of MSE on ShanghaiTech part A and UCF-QNRF datasets when curriculum loss is applied. This demonstrates the limitation of curriculum loss in the cases where extreme crowds exist. Curriculum loss cares more about regions with lower density from the very beginning of the training process and gradually attends the regions with higher density. As a result, the regions with very high density can be less exposed to the learning process. The resultant model performs well for the majority of the regions but also suffers from large errors in the regions of extreme density. These sparse large errors can contribute to MSE more significantly than to MAE on datasets containing very crowded images. In summary, these results provide evidence that the issue of large scale variance can be partially alleviated by the use of our proposed curriculum loss. We will provide more evidence for the effectiveness of curriculum loss in the following ablation study.

E. Results on Curriculum Loss

The use of curriculum loss has shown a positive effect when applied to SGANet for crowd counting (Table I). In this section, we attempt to explore the effectiveness of curriculum learning in the training of other crowd counting networks. To this end, we consider “MCNN", “CSRNet", “SANet", “CANNet", “DADNet" and our modified “Inception-v3" and use the curriculum loss when training these networks on ShanghaiTech part A. Firstly, we try to reproduce the results of these crowd counting models using conventional density map loss under our training protocols to remove the effects of various factors such as the ways of density map generation, patch cropping, data augmentation and so on for a fair comparison and focus on the effect of curriculum loss. It is noteworthy that the generated density maps have different sizes for these models (e.g., the size ratio between input and output is 1 for “SANet", 2 for “DADNet", 4 four “MCNN" and “Inception-v3", 8 for

TABLE I
COMPARISON RESULTS WITH STATE-OF-THE-ART MODELS FOR CROWD COUNTING (– DENOTES THE RESULTS ARE NOT AVAILABLE; CL DENOTES CURRICULUM LOSS).

Model	ShTechA		ShTechB		UCF-QNRF		UCF-CC-50	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN [18]	110.2	173.2	26.4	41.3	–	–	377.6	509.1
CSRNet [54]	68.2	115.0	10.6	16.0	–	–	266.1	397.5
SANet [24]	67.0	104.5	8.4	13.6	–	–	258.4	334.9
DADNet [26]	64.2	99.9	8.8	13.5	113.2	189.4	285.5	389.7
CANNet [55]	62.3	100.0	7.8	12.2	107	183	**212.2**	**243.7**
TEDNet [25]	64.2	109.1	8.2	12.8	113	188	249.4	354.5
Wan et al. [56]	64.7	*97.1*	8.1	13.6	101	176	–	–
RANet[53]	59.4	102.0	7.9	12.9	111	190	239.8	319.4
ANF [57]	63.9	99.4	8.3	13.2	110	174	250.2	340.0
SPANet [58]	59.4	**92.5**	6.5	*9.9*	–	–	232.6	311.7
Inception-v3	60.1	105.0	*6.4*	**9.8**	95.6	165.4	236.0	304.9
SGANet	*58.0*	100.4	**6.3**	10.6	*89.1*	**150.6**	224.6	314.6
SGANet + CL	**57.6**	101.1	6.6	10.2	**87.6**	*152.5*	*221.9*	*289.8*

TABLE II
THE EFFECT OF CURRICULUM LEARNING IN DIFFERENT MODELS ON SHANGHAITECH PART A (THE SYMBOL ↓ MEANS THE ERROR DECREASES WITH THE USE OF CURRICULUM LOSS).

Model	Without CL		With CL	
	MAE	RMSE	MAE	RMSE
MCNN	91.8	144.9	89.1 ↓	142.3 ↓
CSRNet	67.2	110.5	66.7 ↓	113.7
SANet	64.0	103.4	62.1 ↓	100.3 ↓
CANNet	65.6	106.7	63.9 ↓	103.9 ↓
DADNet	63.7	107.4	64.2	102.1 ↓
Inception-v3	60.1	105.0	58.2 ↓	97.9 ↓
SGANet	58.0	100.4	57.6 ↓	101.1

"CSRNet" and "CANNet"). The ground truth density maps need to be resized by sum pooling to have the same size as the corresponding outputs. As a result, the pixel values of the ground truth density maps for different models will have different distributions. This leads to model-specific curriculum designs (i.e. the parameter values in Eq.(10)). Specifically, we set b as the maximum value in the Gaussian kernel matrix G_σ used for density map generation (c.f. Eq. (2)) so that the sparse crowd regions without annotation overlapping will not be affected throughout the training process. The value of k in Eq. (10) is determined by the number of epochs so that all the crowd regions will contribute to the loss equally before training is finished. In our experiments, we set $k = 1e-3$ and $b = 0.1$ for SGANet. Experimental results are shown in Table II. The use of curriculum loss improves the performance of most models. Specifically, the MAE decreases for all models except "DADNet" and the RMSE decrease for all models except "CSRNet". These experimental results demonstrate the curriculum loss is useful not only for our SGANet but also many other crowd counting models.

To evaluate the effect of crowd density in the performance of SGANet and the curriculum loss, an additional experiment is conducted on the UCF_QNRF dataset. As mentioned above, we have changed the image resolutions in this dataset to be no higher than 2048 for computation efficiency. In this experiment, we create two more datasets by setting the image resolution thresholds as 1024 and 512 respectively. As a result, the images in the UCF_QNRF_512 dataset will have higher crowd density than those in the UCF_QNRF_1024 dataset which again consists of denser crowds than the UCF_QNRF_2048 dataset. We use SGANet on these three datasets and the experimental results are shown in Table III. It is obvious the image resolutions make a significant different in the performance and the models perform the best on the UCF_QNRF_2048 dataset whose image resolutions are higher hence have less crowded images. By comparing the performance of SGANet without and with curriculum loss, the use of curriculum loss leads to better results on all three datasets in terms of both MAE and RMSE except that in the last column of Table III. The performance gains achieved by the use of curriculum loss are also related to the image resolutions or the crowd densities in the datasets. Specifically, the MAE and RMSE are reduced by 7.6 and 18.9 respectively on UCF_QNRF_512, 5.0 and 9.2 on UCF_QNRF_1024, 1.5 and -1.9 on UCF_QNRF_2048. These results provide more evidence that the use of curriculum loss is more effective when the crowds are denser in the images.

In summary, the experimental results in Tables I–III provide sufficient evidence that the use of curriculum learning can benefit the training of crowd counting models in most cases especially when the head scales vary a lot in the crowd images.

F. Results on Segmentation Guided Attention

From Table I we can see the performance enhancement contributed by the segmentation guided attention layer by comparing the performance between *Inception-v3* and SGANet. To validate the superiority of our segmentation guided attention layer to other similar designs [44], we conduct an experiment on ShanghaiTech part A and UCF_QNRF. In this experiment, we follow [44] and modify the SGANet by feeding the feature

TABLE III
THE EFFECT OF IMAGE RESOLUTIONS IN THE PERFORMANCE OF SGANET ON UCF_QNRF DATASET.

Model	UCF_QNRF_512		UCF_QNRF_1024		UCF_QNRF_2048	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
SGANet	126.1	236.1	102.5	178.4	89.1	150.6
SGANet + CL	118.5	217.2	97.5	169.2	87.6	152.5
Performance gain	7.6	18.9	5.0	9.2	1.5	-1.9

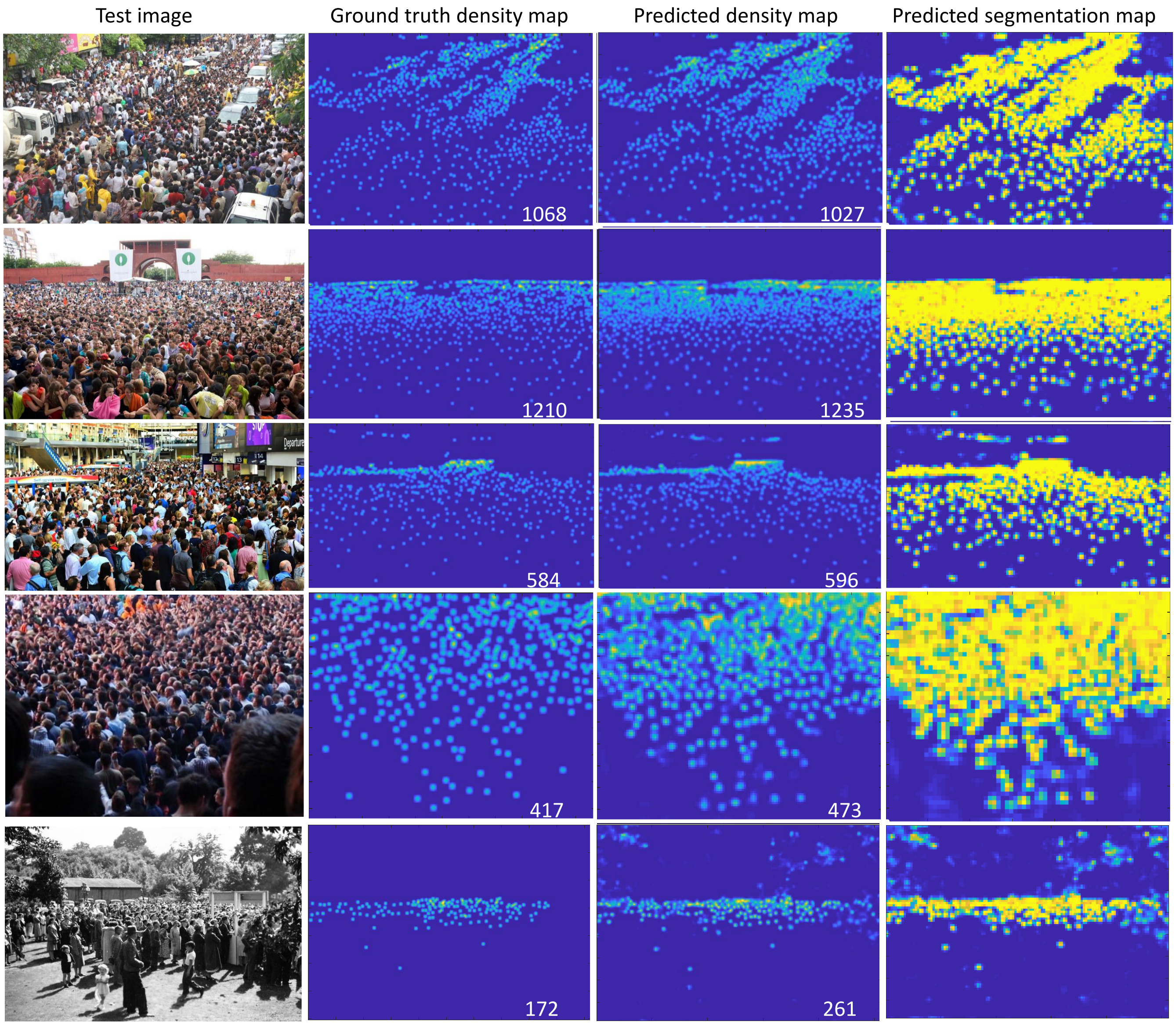


Fig. 3. Visualization of estimated density and segmentation maps for five test images from ShanghaiTech part A. The numbers shown on the images in the second and third columns are the ground truth and estimated counts respectively.

maps of the last *Inception* module into the attention layer and keeping the rest unchanged. The experimental results are shown in Table IV from which we conclude the way segmentation maps are used in our SGANet outperforms that in [44].

To give an intuitive evidence on how the attention layer helps for density map estimation, we visualize the estimated attention maps and density maps for five exemplar test images from ShanghaiTech part A. In Figure 3, we show the original

TABLE IV
RESULTS OF DIFFERENT APPROACHES TO SEGMENTATION MAP SUPERVISION.

Model	ShTechA		UCF_QNRF	
	MAE	RMSE	MAE	RMSE
W/o Seg. map	60.1	105.0	95.6	165.4
W/ Seg. map as [44]	59.5	102.2	92.3	155.3
W/ Seg. map as SGANet	**58.0**	**100.4**	**89.1**	**150.6**

TABLE V
COMPARISON RESULTS OF DIFFERENT TYPICAL DEEP NEURAL NETWORKS (MEAN±STD OVER FIVE TRIALS ARE REPORTED).

Model	#Param	ShTechA		ShTechB		UCF-QNRF		UCF-CC-50	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
VGG16-bn [59]	14.8M	72.2±0.7	124.5±3.2	9.0±0.2	15.6±1.1	125.7±0.7	202.3±4.8	319.6±6.7	437.4±14.1
VGG16-bn*	14.8M	72.3±0.8	124.0±1.6	8.8±0.3	14.4±0.5	121.1±2.8	193.6±6.1	326.4±11.9	448.4±13.7
ResNet50 [60]	23.8M	73.1±1.3	117.9±2.6	8.6±0.3	13.9±0.5	93.1±1.0	163.9±2.6	301.9±9.2	408.3±23.5
ResNet50*	23.8M	70.3±1.3	109.4±4.1	8.8±0.2	13.4±0.4	91.3±0.6	158.4±2.2	278.8±11.5	375.6±18.9
ResNet101 [60]	42.8M	81.0±2.6	127.3±4.1	8.6±0.1	14.3±0.5	94.5±1.4	166.2±3.4	321.1±13.4	432.7±24.3
ResNet101*	42.8M	76.8±6.5	119.1±7.5	8.5±0.2	14.1±0.4	91.9±2.3	161.2±7.3	299.2±10.0	379.8±18.8
DenseNet121 [61]	7.1M	78.4±5.4	128.1±13.	9.4±0.5	15.3±0.5	93.5±1.3	166.2±2.2	446.5±12.1	616.1±24.5
DenseNet121*	7.1M	68.5±3.5	113.3±8.5	8.8±0.7	14.5±1.0	91.0±1.8	161.8±4.1	294.3±24.2	404.3±33.0
ShuffleNet-v2 [62]	0.5M	96.8±0.7	148.9±1.4	15.1±0.5	23.8±0.4	137.4±1.2	226.2±3.2	618.2±37.8	806.3±37.1
ShuffleNet-v2*	0.5M	93.2±2.0	145.9±2.9	14.5±0.7	23.4±1.5	127.4±1.8	217.9±2.1	626.2±55.4	798.2±57.4
Inception-v3 [23]	21.8M	*60.1±1.2*	*105.0±1.8*	**6.4±0.3**	**9.8±0.8**	95.6±1.3	165.4±2.8	236.0±5.8	**304.9±17.8**
SGANet	21.8M	**58.0±0.9**	**100.4±1.3**	**6.3±0.3**	**10.6±0.7**	*89.1±1.1*	**150.6±3.3**	**224.6±6.6**	**314.6±19.2**
SGANet + CL	21.8M	**57.6±0.7**	**101.1±1.5**	**6.6±0.4**	**10.2±0.5**	**87.6±0.9**	**152.5±2.5**	**221.9±6.3**	**289.8±15.6**

images, ground truth density maps, predicted density maps and predicted segmentation maps in four columns respectively. The real and predicted counts are also shown on the density maps for a direct comparison. We can see that the prediction errors for the top three examples are relatively low given the accurately predicted segmentation maps. However, the bottom two images suffer from higher errors since the model can not predict accurate foreground regions. For example, the image in the fourth row contains people raising their hands in the air and the hands are easy to be counted since they have similar colours with human faces. In the bottom image, the trees in the background are mistakenly recognised as foreground and result in the over-estimated count.

G. Results on Typical Deep Neural Networks

In this experiment, we compare the performance of *Inception-v3* and our proposed variants with several typical deep neural networks. Specifically, we consider the most popular and performant models including VGG16bn (VGG16 with batch normalisation layers) [59], ResNet50 [60], ResNet101 [60], DenseNet121 [60] and ShuffleNet-v2-0.5x [62]. Similar to the modifications we have made on Inception-v3, we replace the final fully-connected layers with convolutional layers for density map estimation. For each model, we also consider their counterpart with the segmentation guided attention map (those marked with * in Table V). Specifically, an attention map is learned from an intermediate feature map close to the final output layer by two convolutional layers. Similar to our SGANet, the attention map is learned by the supervision of the segmentation map and applied to the final feature map before the density map estimation.

Experiments are conducted on the four crowd counting benchmark datasets and the results are shown in Table V. For each model, we repeat the experiment for five times with random initialisation to get the statistics (i.e. mean ± standard deviation) as reported in Table V.

In general, Inception-v3 performs significantly better than other five models on all three out of four datasets. On the UCF_QNRF dataset, Inception-v3 performs comparably with ResNet50, ResNet101 and DenseNet121. This demonstrates that inception modules in Inception-v3 are beneficial to crowd counting since the inception module was designed to capture different scales of contextual information in each convolutional layer. ResNet50 achieves the second best overall performance over four benchmark datasets whilst the deeper version ResNet101 performs consistently worse than ResNet50. This phenomenon is also observed when a deeper version of DenseNet121 was employed in our preliminary experiments on the ShanghaiTech A dataset which are not presented here. This may be due to the fact that the outputs of deeper models have lower spatial resolution and lead to less accurate density map estimation. Among six investigated DNN models, ShuffleNet-v2 performs the worst and this is expected since this model has a significant smaller number of parameters (0.7M) than others (7.1-42.8M). Our proposed SGANet (a variant of Inception-v3) with or without the curriculum loss can generally achieve statistically significant better performance than the original Inception-v3 with only negligible additional parameters.

The effectiveness of segmentation guided attention maps is also observed consistently when they are added to the other deep models. As shown in Table V, for five considered deep models, their variants with the use of segmentation guided attention maps achieve better performance in almost all cases.

H. Results on Cross-Dataset Transfer Learning

In this experiment, we investigate the capabilities of cross-dataset transfer learning of different baseline models and our proposed methods. To this end, we train the models on UCF_QNRF training data and test them on ShanghaiTech A and B test data. We choose UCF_QNRF as the training data due to the fact it consists of much more training images than other datasets and the training images have a large range of resolutions. Experimental results are shown in Table VI. Again, we repeat each experiment for five times to get the statistics. The experimental results show that models trained on UCF_QNRF perform slightly worse than those trained within datasets without the need of tranfer learning. This is due to the distribution shift across different datasets. One exception is DenseNet121 performs better on ShanghaiTech A test data when it is trained on the UCF_QNRF trainign

TABLE VI
RESULTS OF CROSS-DATASET TRANSFER LEARNING (MEAN±STD OVER FIVE TRIALS ARE REPORTED).

Model	ShTechA		shTechB	
	MAE	RMSE	MAE	RMSE
VGG16-bn [59]	78.4±1.9	138.5±2.2	12.0±0.8	19.3±1.3
ResNet50 [60]	75.6±2.8	130.1±5.1	13.6±1.2	22.5±1.5
ResNet101 [60]	77.8±1.2	139.1±2.6	11.8±1.1	20.0±1.3
DenseNet121 [61]	68.0±0.9	114.9±1.1	11.6±1.3	19.4±1.5
ShuffleNet-v2 0.5x [62]	102.9±2.9	167.8±5.6	19.1±1.1	28.6±1.0
Inception-v3 [23]	72.8±3.2	125.6±5.2	11.4±1.5	20.0±1.9
SGANet	74.6±1.8	128.2±3.4	9.8±0.5	18.2±1.1
SGANet+CL	71.3±2.4	122.4±3.9	10.1±0.8	19.1±1.4

data. Other than this exception, the performances of different models on cross-dataset transfer learning are consistent with the results in Table V. Our proposed methods SGANet with or without curriculum loss outperform other comparative models. These results provide evidence the proposed methods based on the Inception-v3 are more capable of transfer learning across datasets hence are more useful in practice.

V. Discussion and Conclusion

In this paper, we address an important problem in crowd counting which can be of great values to intelligent transportation systems. We investigated the effectiveness of *Inception-v3* in crowd counting and proposed a segmentation guided attention network using *Inception-v3* as the backbone. We also proposed a novel curriculum loss function for crowd counting by defining pixel-wise difficulty levels to resolve the issue of scale variance in crowd images. Experimental results on four commonly used datasets demonstrate the proposed SGANet can achieve superior performance due to the combination of *Inception-v3* and the segmentation guided attention layer. The proposed strategy of curriculum learning is also proved to be helpful for a variety of existing crowd counting models in general.

Although the proposed two strategies can promote the crowd counting performance in most scenarios, there exist cases where they could fail. For example, when the heads in images are less crowded (e.g., ShanghaiTech part B), both the segmentation guided attention and the curriculum loss will not make a difference to the counting accuracy since they can provide little additional information for the learning process in these situations. As a result, we can expect more benefit from the proposed two strategies when the images contain extremely dense crowds, otherwise a more powerful backbone such as *Inception-v3* will be the optimal solution to achieve high counting accuracy.

This is the first attempt to use the whole *Inception-v3* model for crowd counting and achieves state-of-the-art performance on commonly used datasets. Although the employed *Inception-v3* model (with our own modifications) is not designed from scratch, it is quantitatively shown to be able to achieve superior performance to many specially designed models in the recent couple of years. To these ends, our work is both disruptive and important to the crowd counting research community. Researchers in this community have devoted too much effort to the design of variant CNN architectures and most of them are based on the pre-trained VGG16 model which just has insufficient expressive capacity for crowd counting tasks. In this sense, we believe it is important and necessary to make the community aware of the fact *Inception-v3* is a more suitable architecture for effective crowd counting and divert the attention of the community to more diverse research directions.

Most existing crowd counting methods including ours in this paper rely on a large amount of training data which require extensive efforts of data collection and annotation. In real-world applications, it is challenging to get access to sufficient training data for various scenarios (e.g., different camera resolutions, illumination conditions, weather conditions and perspectives). To solve this realistic problem, our future work will focus on weakly supervised learning such as domain adaptation [63] and transfer learning [28] for which the method proposed in this paper can be served as a strong baseline. On the other hand, our proposed method using *Inception-v3* as the backbone also inherits its limitations that it suffers from gradient vanishing issues when becoming deeper. To resolve this issue, the skip connections [60] and self-attention modules [64] should be considered in the future work.

References

[1] Q. Zhou, J. Zhang, L. Che, H. Shan, and J. Z. Wang, "Crowd counting with limited labeling through submodular frame selection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1728–1738, 2018.

[2] B. Zhan, D. N. Monekosso, P. Remagnino, S. A. Velastin, and L.-Q. Xu, "Crowd analysis: a survey," *Machine Vision and Applications*, vol. 19, no. 5-6, pp. 345–357, 2008.

[3] D. Ryan, S. Denman, S. Sridharan, and C. Fookes, "An evaluation of crowd counting methods, features and regression models," *Computer Vision and Image Understanding*, vol. 130, pp. 1–17, 2015.

[4] 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*, vol. 107, pp. 3–16, 2018.

[5] X. Ding, F. He, Z. Lin, Y. Wang, H. Guo, and Y. Huang, "Crowd density estimation using fusion of multi-layer features," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[6] W. Xie, J. A. Noble, and A. Zisserman, "Microscopy cell counting and detection with fully convolutional regression networks," *Computer methods in biomechanics and biomedical engineering: Imaging & Visualization*, vol. 6, no. 3, pp. 283–292, 2018.

[7] M. Liang, X. Huang, C.-H. Chen, X. Chen, and A. O. Tokuta, "Counting and classification of highway vehicles by regression analysis." *IEEE Trans. Intelligent Transportation Systems*, vol. 16, no. 5, pp. 2878–2888, 2015.

[8] T. Moranduzzo and F. Melgani, "Automatic car counting method for unmanned aerial vehicle images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, no. 3, pp. 1635–1647, 2014.

[9] M. V. Giuffrida, M. Minervini, and S. A. Tsaftaris, "Learning to count leaves in rosette plants," 2016.

[10] S. Aich and I. Stavness, "Leaf counting with deep convolutional and deconvolutional networks," in *ICCV Workshop, Venice, Italy*, 2017, pp. 22–29.

[11] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *CVPR*. IEEE, 2003, p. 459.

[12] L. Dong, V. Parameswaran, V. Ramesh, and I. Zoghlami, "Fast crowd segmentation using shape indexing," in *ICCV*. IEEE, 2007, pp. 1–8.

[13] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance*. IEEE, 2012, pp. 470–475.

[14] D. Kong, D. Gray, and H. Tao, "A viewpoint invariant approach for crowd counting," in *ICPR*, vol. 3. IEEE, 2006, pp. 1187–1190.

[15] P. Siva, M. Javad Shafiee, M. Jamieson, and A. Wong, "Real-time, embedded scene invariant crowd counting using scale-normalized histogram of moving gradients (homg)," in *CVPR Workshop*, 2016, pp. 67–74.

[16] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *NIPS*, 2010, pp. 1324–1332.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[18] 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*, 2016, pp. 589–597.

[19] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 532–546.

[20] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *ICCV*. IEEE, 2017, pp. 1879–1888.

[21] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5197–5206.

[22] Y. Zhang, C. Zhou, F. Chang, and A. C. Kot, "Attention to head locations for crowd counting," *arXiv preprint arXiv:1806.10287*, 2018.

[23] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016, pp. 2818–2826.

[24] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 734–750.

[25] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao, "Crowd counting and density estimation by trellis encoder-decoder networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6133–6142.

[26] D. Guo, K. Li, Z.-J. Zha, and W. Meng, "Dadnet: Dilated-attention-deformable convnet for crowd counting," in *Proceedings of the ACM International Conference on Multimedia*, 2019, pp. 1823–1832.

[27] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[28] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8198–8207.

[29] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 4031–4039.

[30] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, J.-Y. He, and A. G. Hauptmann, "Improving the learning of multi-column convolutional neural network for crowd counting," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19. New York, NY, USA: ACM, 2019, pp. 1897–1906. [Online]. Available: http://doi.acm.org/10.1145/3343031.3350898

[31] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1774–1783.

[32] V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 270–285.

[33] V. A. Sindagi and V. M. Patel, "Ha-ccn: Hierarchical attention-based crowd counting network," *IEEE Transactions on Image Processing*, vol. 29, pp. 323–335, 2019.

[34] D. B. Sam and R. V. Babu, "Top-down feedback for crowd counting convolutional neural network," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[35] V. A. Sindagi and V. M. Patel, "Multi-level bottom-top and top-bottom feature fusion for crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1002–1012.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015, pp. 1–9.

[37] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, "Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3225–3234.

[38] 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*, 2015, pp. 833–841.

[39] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 615–629.

[40] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7279–7288.

[41] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding, "Perspective-guided convolution networks for crowd counting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 952–961.

[42] M. Zhao, J. Zhang, C. Zhang, and W. Zhang, "Leveraging heterogeneous auxiliary tasks to assist crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 736–12 745.

[43] V. A. Sindagi and V. M. Patel, "Inverse attention guided deep crowd counting network," in *AVSS*, 2019.

[44] Z. Shi, P. Mettes, and C. G. M. Snoek, "Counting with focus for free," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019.

[45] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 41–48.

[46] J. L. Elman, "Learning and development in neural networks: The importance of starting small," *Cognition*, vol. 48, no. 1, pp. 71–99, 1993.

[47] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann, "Self-paced curriculum learning," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[48] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Advances in Neural Information Processing Systems*, 2010, pp. 1189–1197.

[49] Y. Liu, M. Shi, Q. Zhao, and X. Wang, "Point in, box out: Beyond counting persons in crowds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6469–6478.

[50] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2547–2554.

[51] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS Workshop*, 2017.

[52] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[53] A. Zhang, J. Shen, Z. Xiao, F. Zhu, X. Zhen, X. Cao, and L. Shao, "Relational attention network for crowd counting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 6788–6797.

[54] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1091–1100.

[55] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5099–5108.

[56] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1130–1139.

[57] A. Zhang, L. Yue, J. Shen, F. Zhu, X. Zhen, X. Cao, and L. Shao, "Attentional neural fields for crowd counting," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 5714–5723.

[58] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. Hauptmann, "Learning spatial awareness to improve crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6152–6161.

[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

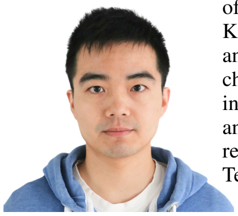
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[61] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[62] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “Shufflenet v2: Practical guidelines for efficient cnn architecture design,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[63] Q. Wang and T. P. Breckon, “Unsupervised domain adaptation via structured prediction based selective pseudo-labeling,” in *AAAI Conference on Artificial Intelligence*, 2020.

[64] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Conference on Neural Information Processing Systems*, 2017.



Qian Wang is a research associate with department of computer science at Durham University, United Kingdom. His researches focus on deep learning and computer vision. He received his PhD in machine learning from The University of Manchester in 2017, Master’s degree in Biomedical Engineering and Bsc in Electronic Engineering in 2013 and 2010 respectively, both from University of Science and Technology of China (Hefei).

Toby P. Breckon is currently a Professor within the Departments of Engineering and Computer Science, Durham University (UK). His key research interests lie in the domain of computer vision and image processing and he leads a range of research activity in this area.

Prof. Breckon holds a PhD in informatics (computer vision) from the University of Edinburgh (UK). He has been a visiting member of faculty at the Ecole Supérieure des Technologies Industrielles Avancées (France), Northwestern Polytechnical University (China), Shanghai Jiao Tong University (China) and Waseda University (Japan).

Prof. Breckon is a Chartered Engineer, Chartered Scientist and a Fellow of the British Computer Society. In addition, he is an Accredited Senior Imaging Scientist and Fellow of the Royal Photographic Society. He led the development of image-based automatic threat detection for the 2008 UK MoD Grand Challenge winners [R.J. Mitchell Trophy, (2008), IET Innovation Award (2009)]. His work is recognised as recipient of the Royal Photographic Society Selwyn Award for early-career contribution to imaging science (2011). http://www.durham.ac.uk/toby.breckon/