Inverse Attention Guided Deep Crowd Counting Network

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

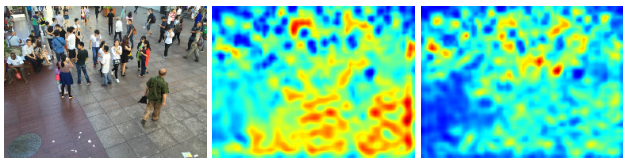
Abstract

In this paper, we address the challenging problem of crowd counting in congested scenes. Specifically, we present Inverse Attention Guided Deep Crowd Counting Network (IA-DCCN) that efficiently infuses segmentation information through an inverse attention mechanism into the counting network, resulting in significant improvements. The proposed method, which is based on VGG-16, is a single-step training framework and is simple to implement. The use of segmentation information results in minimal computational overhead and does not require any additional annotations. We demonstrate the significance of segmentation guided inverse attention through a detailed analysis and ablation study. Furthermore, the proposed method is evaluated on three challenging crowd counting datasets and is shown to achieve significant improvements over several recent methods.

1. Introduction

Crowd counting [15, 44, 11, 45, 46, 36, 31, 5, 28, 47, 16, 23, 24, 34, 30] has attracted a lot of interest in the recent years. With growing population and occurrence of numerous crowded events such as political rallies, protests, marathons, *etc*., computer vision-based crowd analysis is becoming an increasingly important task.

Crowd counting suffers from several challenges such as scale changes, heavy occlusion, illumination changes, clutter, non-uniform distribution of people, etc., making crowd counting and density estimation a very challenging problem, especially in highly congested scenes. Different techniques have been developed to address these issues. The issue of scale variations has received the most interest, with several works proposing different approaches such as multi-column networks [46], switching-cnns [31], use of context information [36], *etc*. While these methods provide significant improvements over recent techniques, the error rates of most of these methods are still far from optimal [25, 46]. A probable reason is that most of these methods train their networks from scratch and since the datasets have limited samples, they are unable to use high-capacity networks. For the few methods [36, 31] that achieve very low error rate, the training process is increasingly complex and requires multiple stages. For instance, Switching-CNN [31] involves different stages such as pre-training, differential training, switch training and coupled training. Similarly, CP-CNN [36] requires that their local and global estimators to be trained separately, followed by end-to-end training of their density map estimator. Although these methods achieve low error, their complex training process makes them hard to use.



(a) (b) (c)

Figure 1. Feature map visualization: (a) Input image, (b) Feature map before refinement, (c) Feature map after refinement using inverse attention. By infusing segmentation information via inverse attention into the counting network, we are able to suppress background regions, thus making the counting task much easier.

Considering these drawbacks, our aim in the paper is to design a simple solution that is easy to train and achieves low count error. Given this objective, we start by presenting a VGG-16 based crowd counting network, which alone is able to achieve results that are comparable to recent state-of-the-art methods. While this baseline network achieves comparable performance with respect to recent methods, there is considerable room for further improvement. We present a simple, yet powerful technique that uses multi-task learning to further boost the counting performance. Specifically, we aim to efficiently infuse foreground/background segmentation mask into the counting network by simultaneously learning to count and segment. This use of related tasks for improving the performance is inspired by the success of recent multi-task approaches such as Hyperface [27], instance aware semantic segmentation

[9] and use of semantic segmentation for improving object detection.

Although this approach of infusing segmentation information through simple multi-task learning achieves considerable improvements in performance, it is limited by the fact that VGG16 is pre-trained on image net dataset and it will concentrate on regions with high response values during learning. To address this issue, we draw inspiration from the success of attention learning in various tasks such as action recognition, object recognition, image captioning, visual question answering *etc*. [7, 38, 32, 2, 18, 43, 22, 41, 8], *etc*. Specifically, we propose an inverse attention module that captures important regions in the feature maps to focus on during learning. The inverted attention map enforces the network to focus specifically on relevant regions, thereby increasing the effectiveness of the learning mechanism.

Through a detailed ablation study, we demonstrate that the infusion of segmentation information via inverse attention results in enrichment of feature maps there by providing considerable gains in the count error. Fig. 1 visualizes the feature maps from intermediate layer of the base network before and after segmentation infusion via inverse attention. It can be easily observed that by incorporating segmentation information, we are able to suppress the background regions easily. More visualization results are provided in the results section. Furthermore, since the infusion requires minimal additional parameters, hence it resulting in minimal computational overhead during inference.

To summarize, the following are our key contributions:

- We propose a crowd counting network that efficiently infuses segmentation information into the counting network.
- An inverse attention mechanism is introduced to further improve the efficacy of the learning mechanism.

In the following sections, we discuss related work (Section 2) and the proposed method in detail (Section 3). Details of experiments and results of the proposed method along with comparison on different datasets are provided in Section 4, followed by conclusions in Section 5.

2. Related work

Crowd Counting. Some of the early methods for crowd counting were based on detection techniques [14, 21]. However, these methods were not robust to occlusions in crowded scenes. To overcome this, several works [29, 6, 11] extracted hand designed features from image patches and employed them in different regression frameworks. Since these approaches mapped image patches to count directly, they tend to loose spatial information in the images. This was overcome by the density estimation techniques [13, 26, 42], where the counting problem is posed as a pixel-to-pixel translation. A more comprehensive survey of different crowd counting methods can be found in [6, 15]. These methods relied on hand-designed features and hence have limited abilities to achieve low count error.

Advancements in deep learning and convolutional neural networks (CNN) have boosted the accuracy of counting techniques. For example several works like [40, 45, 31, 1, 39, 25, 46, 31, 36, 4] have demonstrated significant improvements over the traditional methods. Interested readers are referred to [37] for a comprehensive survey of existing methods. Most of the CNN-based methods [46, 25] addressed the issue of scale variation using different architectures. Zhang *et al*. [46] used multi-column architecture, with each column having different receptive field sizes. Sam *et al*. [31] built upon the multi-column networks, where they trained a Switching-CNN network to automatically choose the most optimal regressor among several independent regressors for a particular input patch. Sindagi *et al*. [36] proposed Contextual Pyramid CNN (CP-CNN), where they demonstrated significant improvements by fusing local and global context through classification networks. Babu *et al*. [3] proposed a mechanism which involved automatically growing CNN to incrementally increase the network capacity. In another interesting approach, Liu *et al*. [19] proposed to leverage unlabeled data for counting by introducing a learning to rank framework. Recently, Li *et al*. [17] proposed CSR-Net, that consists of two components: a front end CNN-based feature extractor and a dilated CNN for the back-end.

3. Inverse Attention Guided Deep Crowd Counting Network (IA-DCCN)

Fig. 2 provides an overview of the proposed Inverse Attention Guided Deep Crowd Counting Network (IA-DCCN) which is based on the VGG-16 network. We include an inverse attention block (IAB) with the objective of enriching the feature maps from VGG-16, thereby resulting in substantial improvements in the performance. This inverse attention block aims to encode segmentation masks into the feature maps due to which the counting task becomes considerably easier. Additionally, we employ a simple hard-mining technique, that effectively samples the training data due to which appreciable gains are observed. Details of the proposed method and its various components are described in the following sub-sections.

3.1. Base network

As illustrated in Fig. 2, the base network consists of three parts: (i) first five convolutional blocks (conv1-conv5) from the VGG-16 architecture, (ii) dimensionality reduction and upsampling (DRU) module that reduces the dimensionality of feature maps from VGG-16 along the depth to 64 channels and upsamples them, and (iii) density module (DM) a

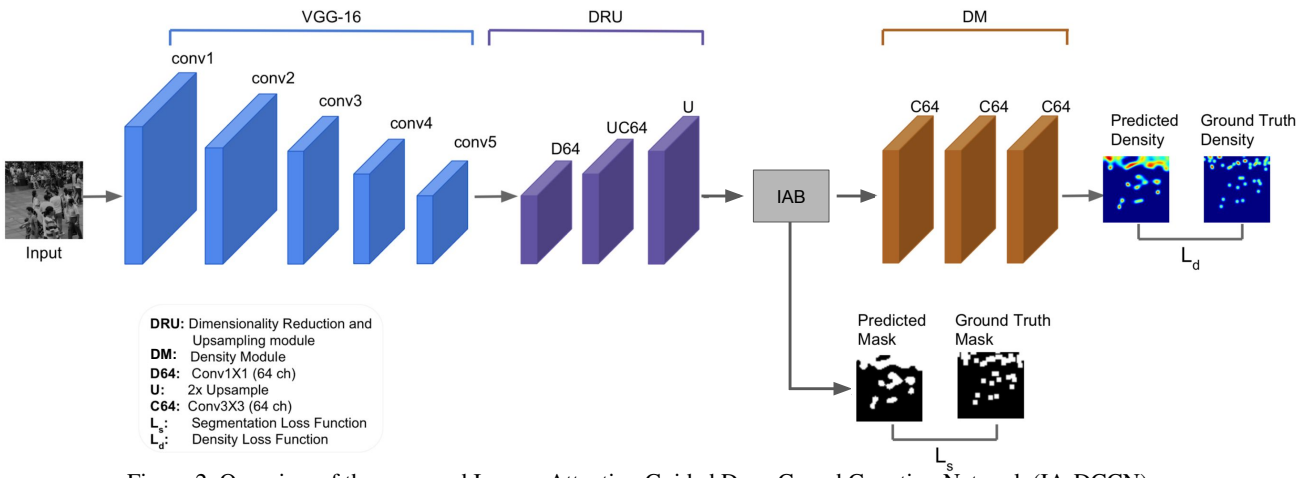


Figure 2. Overview of the proposed Inverse Attention Guided Deep Crowd Counting Network (IA-DCCN).

set of three conv layers (with 64 channels and 3×3 filters) to perform the density estimation. Note that the network is fully convolutional and hence, it can be used on images of any size. The entire network regresses on the input image to produce a density map which indicates the number of people per pixel. This density map, when summed over all the pixels, provides an estimate of the number of people in the input image. The conv layers belonging to VGG-16 architecture are initialized with pre-trained weights, where as the conv layers in the density estimation network are randomly initialized with $\mu = 0$ and $std = 0.01$. The network is trained by minimizing the following loss function:

$$L_d = \frac{1}{N}\sum_{i=1}^{N} \|F_d(X_i, \Theta) - D_i\|_2, \tag{1}$$

where, N is number of training samples, X_i is the i^{th} input image, $F_d(X_i, \Theta)$ is the estimated density, D_i is the i^{th} ground-truth density and it is calculated by summing a 2D Gaussian kernel centered at every person's location x_g as follows:

$$D_i(x) = \sum_{x_g \in S} \mathcal{N}(x - x_g, \sigma), \tag{2}$$

where σ is scale parameter of 2D Gaussian kernel and S is the set of all points at which people are located. Following [46], the density map generated by the network is $1/4^{th}$ of the input image resolution.

3.2. Segmentation infusion via Inverse Attention

The base-network, although very simple, achieves significantly low count errors and the results are better/comparable with respect to recent state-of-the-art methods [31, 36]. In order to further boost the performance, we propose to incorporate segmentation information into the counting network. A naive idea would be add to segmentation loss layer after an intermediate block in the network and train the network in a multi-task fashion. This would be similar to recent works like [27, 10, 10, 9] that learned different tasks simultaneously.

While this method results in a better performance as compared to the base network, we propose a more sophisticated method that uses inverse attention to incorporate segmentation information. For this, we draw inspiration from the recent work on tasks like image captioning, super-resolution, classification, visual question answering [7, 38] that use different forms of attention mechanisms to learn the features more effectively. Specifically, we introduce an inverse attention block (IAB) on top of the DRU module in the counting network as shown in Fig. 2.

Fig 3 illustrates the mechanism of the inverse attention block. Specifically, the IAB takes feature maps (F) from the DRU as input and forwards them through a conv block CB_A to estimate background regions (which we call as inverse attention map - A^{-1}) in the input image. CB_A is defined by $\{\text{conv}_{512,32,1}\text{-relu-conv}_{32,32,3}\text{-relu-conv}_{32,1,3}\}$[1]. Feature maps F weighted by the inverse attention map are then subtracted from F to suppress the background regions *i.e.*,

$$F' = F - F \odot A^{-1},$$

where F' is the attended feature map which is then forwarded through the density map module.

While the existing attention-based work learn the attention maps in a self-supervised manner, we instead use the ground-truth density maps to generate ground-truth inverse attention maps for supervising the inverse attention block. To generate the ground-truth, the density maps are thresholded and inverted. Note that by learning to estimate the

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

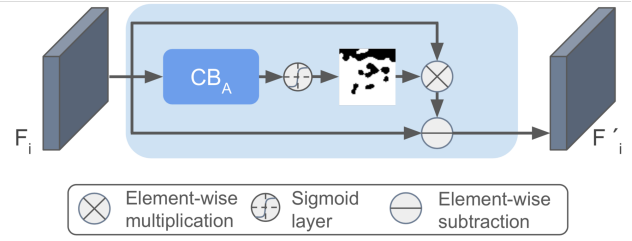


Figure 3. Inverse attention block.

background regions we are automatically suppressing the background information from the feature maps of the DRU, hence, making it easier for the density module (DM) to learn the features more effectively. Fig. 1 illustrates the feature maps before and after enrichment using IAB. It can be clearly observed that the use of inverse attention block aids in better feature learning.

The entire network is trained in a multi-task fashion by simultaneously minimizing the density loss and the segmentation loss. Formally, the overall loss (L) is defined as follows:$L = L_d + \lambda_s L_s$

$$L = L_d + \lambda_s L_s, \tag{3}$$

where, L_d is the density loss (Eq. 1), L_s is segmentation loss that is used for training the features of the inverse attention block, and λ_s is weighting factor for the segmentation loss. L_s is pixel-wise cross entropy error between estimated mask and ground-truth mask. Note that, the method does not require any additional labeling and it uses weakly annotated head/person regions based on the existing labels to compute the segmentation loss. The ground-truth mask is generated by thresholding the ground-truth density map. Bascially, the pixels that contain head regions are labeled as 1 (foreground), and otherwise as 0 (background). In spite of these annotations being noisy, the use of segmentation information results in considerable gains.

3.3. Hard sample mining (HSM)

Recent methods such as [33, 20] have demonstrated that effective sampling of data by selecting harder samples improves the classification performance of the network. Similar to these work, we employ an offline hard mining technique to train the network. This process, used to select samples from the training set every 5 epochs, involves the following steps: (i) compute the histogram of error on the entire training data, (ii) find the mode (T) of this error distribution (iii) training samples with $error > T$ are considered as hard samples and selected for training. This sample selection technique is effective in lowering the count error by an appreciable margin.

4. Experiments and results

In this section, we first describe the training and implementation specifics followed by a detailed ablation study to understand the effects of different components in the proposed network. We chose the ShanghaiTech dataset [46] to perform the ablation study as it contains significant variations in count and scale. Finally, we compare the results of the proposed method against several recent approaches on three publicly available datasets containing congested scenes. (ShanghaiTech, UCF_CROWD_50 [11], UCF-QNRF [12]).

4.1. Training and implementation details

The network is trained end-to-end using the Adam optimizer with a learning rate of 0.00005 and a momentum of 0.9 on a single NVIDIA GPU Titan Xp. 10 % of the training set is set aside for validation purpose. The final training dataset is formed by cropping patches of size 224×224 from 9 random locations from each image. Furthermore, data augmentation is performed by randomly flipping the images (horizontally) and adding random noise. Since the network is fully convolutional, image of any arbitrary size or resolution can be input to the network. Similar to earlier work, the count performance is measured using the following metrics:

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

where, MAE is mean absolute error, MSE is mean squared error, N is the number of test samples, y_i is the ground-truth count and y_i' is the estimated count corresponding to the i^{th} sample.

4.2. Architecture ablation

To understand the effectiveness of the various modules present in the network, we perform experiments with the different settings using the ShanghaiTech dataset (Part A and Part B). This dataset consists of 2 parts with Part A containing 482 images and Part B containing 716 images and a total of 330,165 head annotations. Both parts have training and test subsets. Due to various challenges such as high density crowds, large variations in scales, presence of occlusion, etc, we chose to perform the ablation study on this dataset.

The results of these experiments are tabulated in Table 1. It can be observed that the base network, consisting of VGG-16 conv layers along with DRU module and density module (described in Section 3.1), does not provide the optimal performance. With the addition of the segmentation loss layer (Base network + S), we can observe

Table 1. Results of the ablation study on the ShanghaiTech Part A and Part B datasets. Figures in braces indicate the percentage improvement in error over previous configuration.

	Part A		Part B	
Configuration	MAE	MSE	MAE	MSE
Base network	76.7	119.1	17.3	22.5
Base network+S	71.2 (7.1%)	117.5 (1.3%)	15.0 (13.3%)	21.0 (6.7%)
Base network+IAB	68.1 (4.3%)	114.5 (2.5%)	13.6 (9.3%)	19.6 (6.7%)
Base network+IAB+HSM	66.9 (1.8%)	108.5 (5.2%)	10.2 (25.0%)	16.0 (18.3%)

Table 2. Comparison of results on the ShanghaiTech dataset.

	Part A		Part B	
Method	MAE	MSE	MAE	MSE
Cascaded-MTL [35] (AVSS '17)	101.3	152.4	20.0	31.1
Switching-CNN [31] (CVPR '17)	90.4	135.0	21.6	33.4
TopDownFeedack [30] (AAAI '18)	97.5	145.1	20.7	32.8
CP-CNN [36] (ICCV '17)	73.6	**106.4**	20.1	30.1
IG-CNN [3] (CVPR '18)	72.5	118.2	13.6	21.1
CSR-Net [17] (CVPR '18)	68.1	115.0	10.6	16.0
IA-DCCN (ours)	**66.9**	108.4	**10.2**	**16.0**

an improvement of ∼7.1%/1.3% in MAE/MSE on Part-A and ∼15.0%/6.7% in MAE/MSE on Part-B over the base network. While this naive method of infusing segmentation network results in considerable improvements in the error, we show that there is still room for further improvements with the experiment where we incorporated the inverse attention block after the DRU module (Base network + IAB). The IAB module results in an improvement of ∼4.3%/2.5% in MAE/MSE on Part-A and ∼9.3%/6.7% in MAE/MSE on Part-B over the naive method.

Fig. 1 visualizes the feature maps from the DRU module network with/without feature enrichment via segmentation infusion using inverse attention. Effects of incorporating segmentation information into the counting network can be clearly observed. This considerable reduction in the count error confirms our intuition that segmentation guided inverse can be used to aid the counting task, by introducing high-level foreground background knowledge into the feature maps of the VGG-16 network. λ_s in Eq. 3 is set equal to 0.1 based on cross-validation.

Finally, we trained the entire network with hard sample mining (Base network + IAB + HSM as described in Section 3.3), where samples for training were selected based on the error at every fifth epoch. For this configuration, we observed an improvement of ∼1.8%/5.2% in MAE/MSE on Part A and ∼25.0%/18.3% in MAE/MSE on Part B over the base network with inverse attention. Hence, it can be concluded that hard sample mining is effective and provides appreciable gains on both parts of the dataset.

4.3. Comparison with recent methods

For comparison with various methods on different datasets, the entire network (IA-DCCN) is trained with hard sample mining.

ShanghaiTech. The proposed method is compared with four recent approaches (Cascaded-MTL [35], Switching-CNN [31], CP-CNN [36], Top-down feedback [30], IG-CNN [3] and CSR-Net [17]) on Part A and Part B of the ShanghaiTech dataset and the results are presented in Table 2. The proposed IA-DCCN method achieves the lowest error rate in terms of MAE/MSE as compared to all recent methods on both parts of the dataset. Sample density estimation results are shown in Fig. 4. From these results, it can be noted that the proposed method is able to achieve encouraging results while being simple to train as compared to existing approaches.

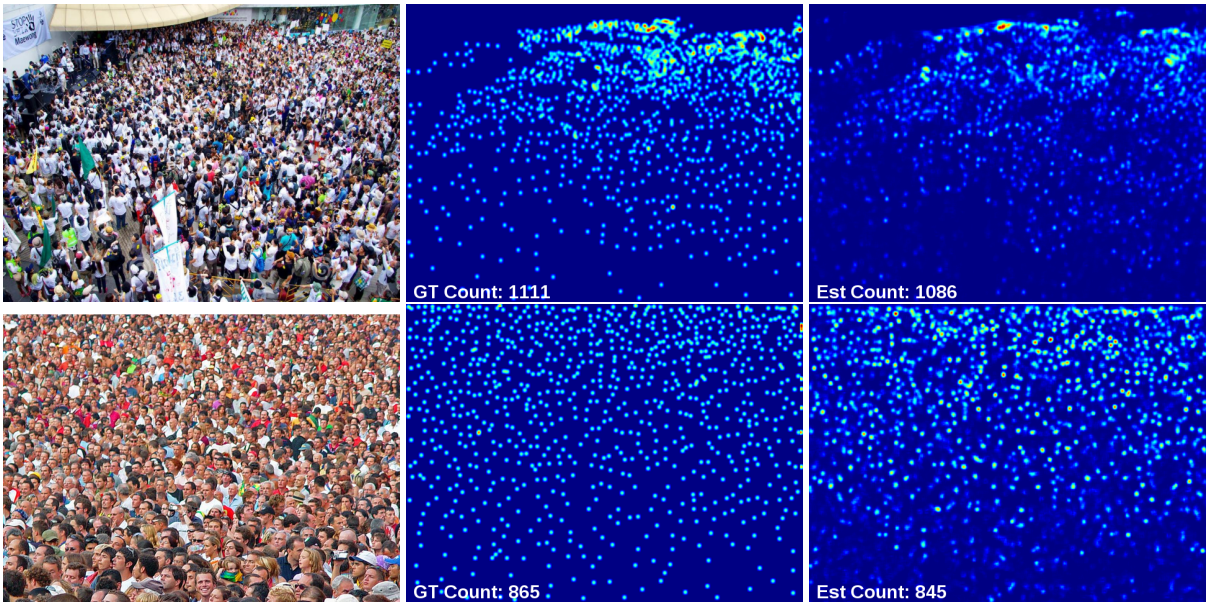
Table 3. Comparison of results on the UCF_CROWD_50 dataset.

Method	MAE	MSE
Cascaded-MTL [35] (AVSS '17)	322.8	397.9
Switching-CNN [31] (CVPR '17)	318.1	439.2
CP-CNN [36] (ICCV '17)	295.8	**320.9**
TopDownFeedback [30] (AAAI '18)	354.7	425.3
IG-CNN [3] (CVPR '18)	291.4	349.4
CSR-Net [17] (CVPR '18)	266.1	397.5
IA-DCCN (ours)	**264.2**	394.4

Table 4. Comparison of results on the UCF-QNRF datastet.

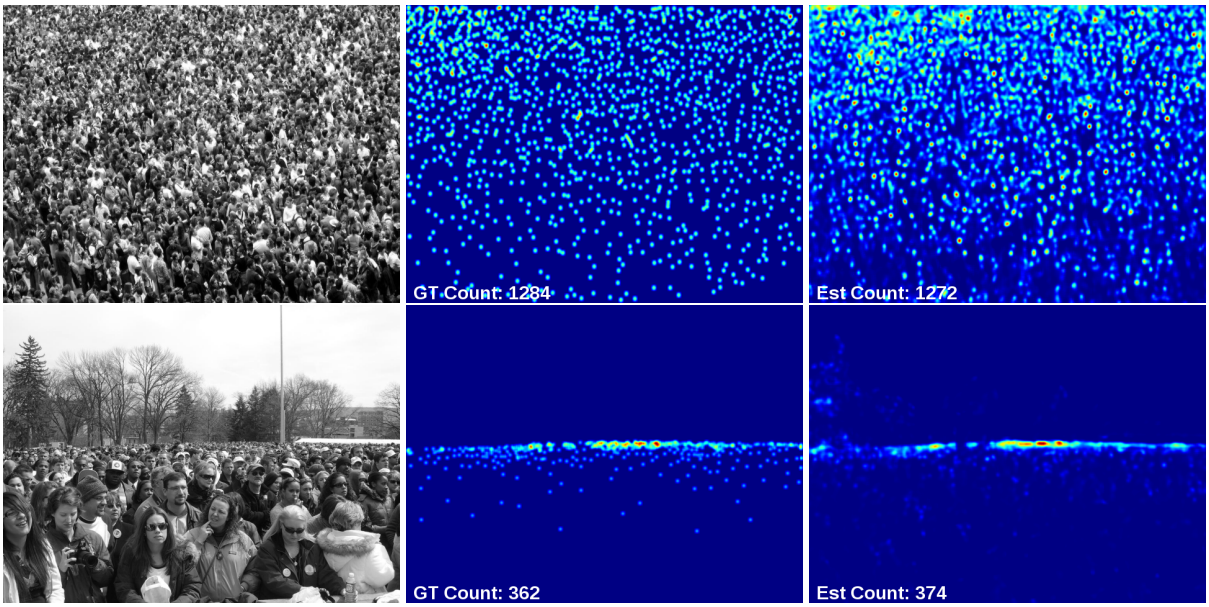
Method	MAE	MSE
Idrees *et al.* [11] (CVPR '13)	315.0	508.0
Zhang *et al.* [45] (CVPR '15)	277.0	426.0
Cascaded-MTL [35] (AVSS '17)	252.0	514.0
Switching-CNN [31] (CVPR '17)	228.0	445.0
Idrees *et al.* [12] (ECCV '18)	132.0	191.0
IA-DCCN (ours)	**125.3**	**185.7**

UCF_CROWD_50. The UCF_CC_50 dataset [11] is a relatively smaller dataset with 50 annotated images of different resolutions and aspect ratios. We used the standard 5-fold cross-validation protocol discussed in [11] to evaluate the proposed method. Results are compared with several recent approaches: Cascaded-MTL [35], Switching-CNN [31], CP-CNN [36], Top-down feedback [30], IG-CNN [3] and CSR-Net [17]. The results are tabulated in Table 4. It can be observed that the proposed method achieves the lowest MAE error as compared to the recent methods. Although the proposed approach performs slightly worse in terms of MSE as compared to CP-CNN [36], it is important to note that the MSE error is comparable to the existing approaches. Additionally, we believe that these results are especially significant considering the simplicity of the



(a) (b) (c)

Figure 4. Sample results of the proposed method on the ShanghaiTech dataset [46]. *(a)* Input. *(b)* Ground truth *(c)* Estimated density map.



(a) (b) (c)

Figure 5. Sample results of the proposed method on the UCF_CROWD_50 dataset [11]. *(a)* Input. *(b)* Ground truth *(c)* Estimated density map.

proposed approach. Sample density estimation results are shown in Fig. 5.

UCF-QNRF: The UCF-QNRF [12] is a recent dataset that contains around 1200 images with approximately 1.2 million annotations. The results of the proposed method on this dataset as compared with recent methods ([11],[46],[35]) are shown in Table 4. The proposed method is compared against five different approaches: [11], [46], [35],[31], and [12]. It can be observed that the proposed method outperforms other methods by a considerable margin.

4.3.1 Inference speed

To evaluate the inference speed of the proposed approach, we run IA-DCCN on our machine which is equipped with Intel Xeon E5-2620v4@2.10GHz and an NVIDIA Titan Xp

GPU. The run times are reported in Table 5 for different resolutions ranging from 320×240 to 1600×1200. It can be observed that the proposed method is efficient and is able to run at ∼76 fps while processing high resolution images (1600×1200). Note that the majority of processing time is taken up by the VGG-16 network.

Table 5. Inference time for different resolutions in msec.

Res (W×H)	320×240	640×480	1280×960	1600×1200
IA-DCCN (ours)	2.7	4.8	8.9	13.1

5. Conclusions

We presented a very simple, yet effective crowd counting approach based on the VGG-16 network and inverse attention, referred to as Inverse Attention Guided Deep Crowd Counting Network (IA-DCCN). The proposed approach aims to infuse segmentation information into the counting network via an inverse attention mechanism. This infusion of segmentation maps into the network enriches the feature maps of VGG-16 network due to which the background information in the feature maps get suppressed, making the counting task rather easier. In contrast to existing approaches that employ complex training process, the proposed approach is a single-stage training framework and achieves significant improvements over the recent methods while being computationally fast.

Acknowledgments

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

References

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

[2] J. Ba, V. Mnih, and K. Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014. 2

[3] D. Babu Sam, N. N. Sajjan, R. Venkatesh Babu, and M. Srinivasan. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3626, 2018. 2, 5

[4] 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. 2

[5] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008. 1

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

[7] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6298–6306. IEEE, 2017. 2, 3

[8] S. Chen, X. Tan, B. Wang, and X. Hu. Reverse attention for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018. 2

[9] J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016. 2, 3

[10] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014. 3

[11] 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. 1, 2, 4, 5, 6

[12] 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 *European Conference on Computer Vision*, pages 544–559. Springer, 2018. 4, 5, 6

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

[14] 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

[15] 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, 2

[16] W. Li, V. Mahadevan, and N. Vasconcelos. Anomaly detection and localization in crowded scenes. *IEEE transactions on pattern analysis and machine intelligence*, 36(1):18–32, 2014. 1

[17] 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*, pages 1091–1100, 2018. 2, 5

[18] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan. End-to-end comparative attention networks for person re-identification. *IEEE Transactions on Image Processing*, 26(7):3492–3506, 2017. 2

[19] X. Liu, J. van de Weijer, and A. D. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *The*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018. 2

[20] I. Loshchilov and F. Hutter. Online batch selection for faster training of neural networks. *arXiv preprint arXiv:1511.06343*, 2015. 4

[21] C. C. Loy, K. Chen, S. Gong, and T. Xiang. Crowd counting and profiling: Methodology and evaluation. In *Modeling, Simulation and Visual Analysis of Crowds*, pages 347–382. Springer, 2013. 2

[22] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016. 2

[23] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos. Anomaly detection in crowded scenes. In *CVPR*, volume 249, page 250, 2010. 1

[24] M. Marsden, K. McGuinness, S. Little, C. E. Keogh, and N. E. O'Connor. People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[25] 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, 2

[26] 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

[27] 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. 1, 3

[28] 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. 1

[29] 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

[30] 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. 1, 5

[31] 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, 2, 3, 5, 6

[32] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. *arXiv preprint arXiv:1511.04119*, 2015. 2

[33] A. Shrivastava, A. Gupta, and R. Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 761–769, 2016. 4

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

[35] 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. 5, 6

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

[37] 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. 2

[38] C. Song, Y. Huang, W. Ouyang, and L. Wang. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1179–1188, 2018. 2, 3

[39] E. Walach and L. Wolf. Learning to count with cnn boosting. In *European Conference on Computer Vision*, pages 660–676. Springer, 2016. 2

[40] 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. 2

[41] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015. 2

[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] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[44] 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

[45] 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. 1, 2, 5

[46] 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

[47] F. Zhu, X. Wang, and N. Yu. Crowd tracking with dynamic evolution of group structures. In *European Conference on Computer Vision*, pages 139–154. Springer, 2014. 1