MULTISCALE CROWD COUNTING AND LOCALIZATION BY MULTITASK POINT SUPERVISION

Mohsen Zand, Haleh Damirchi, Andrew Farley, Mahdiyar Molahasani, Michael Greenspan, Ali Etemad

Dept. ECE & Ingenuity Labs Research Institute, Queen's University, Kingston, Canada

ABSTRACT

We propose a multitask approach for crowd counting and person localization in a unified framework. As the detection and localization tasks are well-correlated and can be jointly tackled, our model benefits from a multitask solution by learning multiscale representations of encoded crowd images, and subsequently fusing them. In contrast to the relatively more popular density-based methods, our model uses point supervision to allow for crowd locations to be accurately identified. We test our model on two popular crowd counting datasets, ShanghaiTech A and B, and demonstrate that our method achieves strong results on both counting and localization tasks, with MSE measures of 110.7 and 15.0 for crowd counting and AP measures of 0.71 and 0.75 for localization, on ShanghaiTech A and B respectively. Our detailed ablation experiments show the impact of our multiscale approach as well as the effectiveness of the fusion module embedded in our network. Our code is available at: https://github.com/RCVLab-AiimLab/crowd_counting

***Index Terms*—** crowd counting, localization, multitask, multiscale, point supervision.

1. INTRODUCTION

Crowd counting is central to many real-world computer vision applications, including crowd management [1, 2], surveillance systems [3], security and planning [4], traffic monitoring [5], animal crowd estimation [6], and cell counting [7]. In general, crowd counting is a challenging problem due to scale variations, occlusions, complex and noisy backgrounds, and variations in perspective and illumination. This problem has been addressed most prominently by estimated density maps [8], wherein annotated head locations are converted to a density map through convolution with a Gaussian kernel, following which integration over the map generates the people count [9]. Utilizing density maps is often considered to be the standard approach towards crowd counting [10, 11].

Recently, promising performances have been achieved by employing convolutional neural networks (CNNs) for crowd counting via density map estimation [10, 12]. Density estimators are however highly sensitive to the choice of the kernel and the kernel size used to generate them. More importantly, the use of density maps results in inconsistent performance with varying crowd sparsities, which in the past has been addressed by CSRNet [10] through expanding the receptive fields of the network. In addition, density-based methods only provide an estimate of people count, and thus fail to capture individual information such as person location and size [1, 13]. Such attributes may be important in many other downstream applications, for instance multi-object tracking, person re-identification, and face recognition.

In this paper, we propose the use of multiscale point supervision [2] to improve the performance of crowd counting in both densely and sparsely populated crowd scenes. Moreover, we employ a multitask approach, which can simultaneously perform localization through exploiting the scene representations learned in the intermediate layers. Our method does not require density maps to be generated and is thus better equipped to deal with varying sparsities in crowd scenes. Our experiments on two large public datasets, ShanghaiTech A and ShanghaiTech B, demonstrate the effectiveness of our proposed method and robustness against varying sparsities.

Our contributions in this paper include the following two aspects: (**1**) We propose a novel multiscale and multitask architecture based on point supervision to effectively estimates both count and location under large variations in the number of people per image. (**2**) Our method achieves strong results and approaches the state-of-the-art on two datasets for both counting and localization. Detailed ablation experiments demonstrate the impact of each component in our network.

2. RELATED WORK

Crowd counting methods can be divided into the two major categories of density-based and point-based [14]. In the following we review the related works in these two categories.

Density-based. Most solutions rely on probability maps to regress the density of the crowd [10, 15]. These approaches utilize binary head locations that have been blurred by a Gaussian kernel to form the ground truth density maps as the estimate of the number of people in the crowd. As an example of such approaches, in [10], dilated convolutions were used to take advantage of a large receptive field to obtain strong crowd image representations.

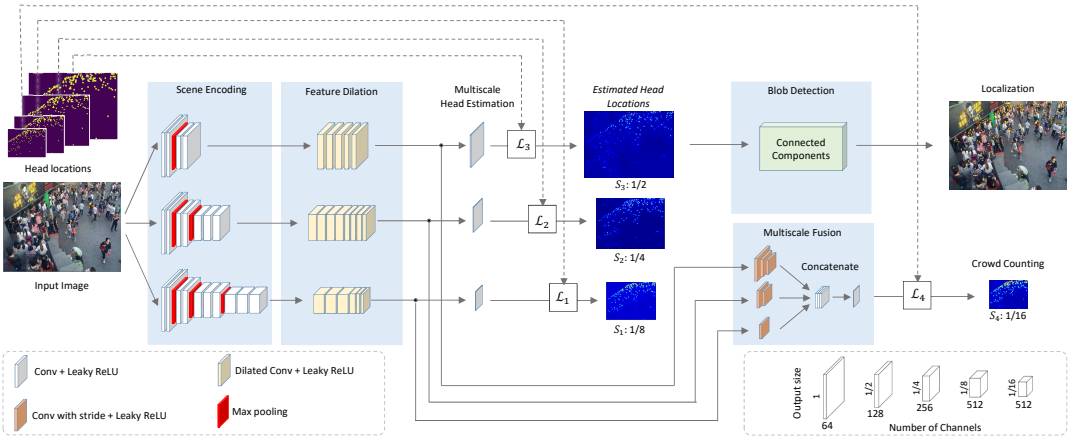

Fig. 1: The architecture of the proposed model. Dashed lines are used solely in the training stage.

To solve the problem of head size variations in crowd images, the use of several branches of deep network has been proposed. For instance, in [15], a switching network was used alongside three regressor networks. Each regressor had a different architecture for different rates of scale and density. The switching network learned to direct the images with high density towards the regressor with a smaller receptive field. In [16], an architecture with two parallel branches of CNNs, one shallow and one deep, was proposed to tackle the problem of scale and perspective variance in crowd images. In [17], an inception-like network was used to generate feature maps of different sizes, while in [9], a CNN with three branches was proposed, each with a different receptive field. The problem of scale-variant head sizes was also addressed and tackled in [8] in which the training images were divided into patches, and categorized by density. After training the CNN, fine-tuning for each test image was carried out using the patches of training images with the same density rate. A stacked pooling approach was later proposed in [18] to solve the problem of counting variable head sizes in crowd images.

Point-based. These methods use crowd head annotations either directly as ground truth or as a means to generate bounding boxes around heads. This approach has the advantage of being able to predict the location of heads in the crowd in addition to counting the number of people present in the image. The method presented in [2] proposed an extension of the semantic segmentation loss in [19] for point supervision based localization. In [1], bounding boxes around point annotations were initialized using their distance to the nearest neighbour head and while training a CNN to output box annotations, updated bounding boxes around the heads to obtain the most suitable size in the anchor box set for the corresponding head. In [20], two scales of the input image feature maps (one-fourth and one-eighth) were used depending on the sparsity of the crowd to output a point map. Crowd counting was also cast as a classification task of dot prediction with point supervision, dropping the prevalent density regression. An adaptive scale fusion module combined the multiscale confidence maps into a single map, where each value indicated the confidence of person detection. A threshold was then applied on this map to generate the final accurate dot predictions. Therefore in [20], the fusion module was not learnable and the count highly depended on the accuracy of the threshold value which might be different across different images.

3. PROPOSED METHOD

Let $P = \{p_i\}_{i=1}^{M}$ be the head coordinates of each person i, where $p_i = (x_i, y_i)$, and M is the total number of people in the image. Our goal is to estimate M and P through a point-supervised multiscale unified neural framework. Our proposed model is described as follows (see Fig. 1).

Multiscale Scene Encoder. An arbitrary-sized image I is fed to three VGG16-based scene encoders, pre-trained on ImageNet [21], to obtain representations in three scales, *i.e.* $S_1 = 1/8$, $S_2 = 1/4$, and $S_3 = 1/2$ of the original image size. We thus obtain three separate embeddings corresponding to the three respective scales, which ultimately improves our point-supervised network's ability to estimate M and P accurately for a large variation of the number of people in a scene (i.e. crowd spatial distribution).

Feature Dilation. The multiscale scene representations are then fed to dilated convolutional layers which are proven useful in previous works [10, 13]. They serve to extend the layers' receptive fields and capture higher level features. Specifically, the dilated CNNs have the benefit of exploiting various receptive field sizes from the original image, without degrading the resolution as may occur when downsampling by increasing convolutional kernel size. The number of layers however are different in these networks. To achieve consistent convergence between the different scales of our pipeline, we adjust the number of layers in different branches.

Multiscale Head Estimation. Each embedding extracted by the feature dilation is fed to a single layer to generate an es-

timate for the head locations in the scene. This is achieved using a one-channel convolution network composed of a convolution layer with a kernel size of 1×1. This structure serves to improve the accuracy of the overall localization results, as well as generating inputs to the multitasking network to support accurate crowd counting. The heatmaps of these embeddings are visualized in Fig. 1 for each scale.

To obtain the estimated head locations, we need the ground-truth head points, which are given in the dataset. Nevertheless, these locations must be normalized to be used at different scales. For each point, we extract the head coordinates from the head location and normalize the x- and y-values in the range of $[0, 1]$ by dividing them by the image width and height, respectively. We then multiply the normalized head locations by the three scale factors to obtain the resized ground-truth head locations.

Multiscale Fusion. This module generates the final density map for crowd counting. It consists of three networks with different number of convolutional layers with stride of 2 and kernel size of 2, each of which works on one of the dilated embeddings. The first network comprises one layer and works on the S_1 scale. Similarly, the second and third networks respectively work on the S_2 and S_3 scales, and include two and three layers. These networks specifically downsample the dilated embeddings at scales S_1, S_2, and S_3, respectively, and generate the same-sized outputs as they use different number of convolutional layers. These outputs are then concatenated channel-wise and fed to a convolutional layer to generate the final density map of $S_4 = 1/16$.

Localization. Given that the resolution of S_3 is higher than the other two branches, the head *locations* can be extracted from this branch more precisely. Accordingly, we use this embedding to perform localization by utilizing a connected components algorithm [22] to obtain the blobs in the scenes. The center of the blobs represent the head locations in the crowd image. Fig. 1 illustrates the detected blobs.

Total Loss. We use four MSE loss terms $\mathcal{L}_j = \|\widehat{D}_j - D_j\|_2^2$ for the three multiscale branches and the multiscale fusion network, where $\widehat{D}_j$ and D_j denote the estimated map at scale S_j (consisting of head locations at this scale) and its ground-truths, respectively. Specifically, D_j represents the embedding at scale S_j. Each location $(x_k^{S_j}, y_k^{S_j})$ in D_j shows the integrated number of people which their original coordinates (x_i, y_i) map to. For example, each 16×16 pixel block in the original image corresponds to one location on the embedding of scale S_4, and the total number of people p_i with $0 < x_i \leq 15$ and $0 < y_i \leq 15$ assign to $(0, 0)$ on D_j. Ideally, all outputs generated from different branches would correspond to the same number of people. This however does not hold in practice, as the different branches of the network will have varying accuracies at detecting heads of different sizes (likely due to varying distances from the camera) [23, 4, 24]. To address this problem, we use task-specific weights w_j in the final loss function, and obtain $\mathcal{L}_{total} = \sum_{j=1}^{4} w_j \mathcal{L}_j$. This approach also helps obtain a more consistent convergence when training our end-to-end model as different branches could have the tendency to learn at different rates.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

Datasets. We use two popular datasets, ShanghaiTech A and ShanghaiTech B [9], which are the most commonly used datasets in the area of crowd counting [8, 10, 2, 11, 20, 18]. ShanghaiTech A contains 482 images with 241,677 total annotated heads. ShanghaiTech B has 716 images with 88,488 total annotated heads [9]. Part A is representing more crowded scenes while part B includes more sparse images. We compensate for the relatively small number of images in the standard ShanghaiTech datasets by augmenting the images similar to [10]. We also first pretrain the scene encoders in our network with the much larger ImageNet dataset [21].

Implementation Details. Adam optimizer was used with a momentum of 0.934 and an initial learning rate of 1E-6. The model was trained for 200 epochs with early stopping, and the task-specific weights were $w_1 = 0.1$, $w_2 = 0.2$, $w_3 = 0.3$, $w_4 = 0.1$ for both datasets. Training was performed with an Nvidia Titan RTX GPU. The architectural details are provided in Fig. 1, where kernel sizes of 3, 1, and 2 have been used for the scene encoding, head estimation, and fusion, respectively. A stride of 1 is used throughout all the convolutional layers, except for the fusion network which uses a stride of 2.

Evaluation. The summation of the output of the multiscale fusion corresponds to the total number of people in the scene. The mean absolute error (MAE) and mean square error (MSE) of the predicted count with respect to the ground truths are calculated. To evaluate the localization results we use average precision (AP) which is the area under the precision-recall curve. If a detected head is within 5 pixels of the ground truth head location, that detection is denoted as a true positive, and is deleted from the ground truth points so that it would not be matched with any other prediction in the future assessments. If a detection it is not within that distance of a head detection, it is counted as a false positive, and finally, if a head detection in ground truth is not matched with any detection, then it is counted as a false negative.

4.2. Results

Crowd Counting. The performance of the proposed model for crowd counting is compared with similar works in Table 1. We observe that for both datasets, training on auxiliary localization information in the multitask approach helps with learning more effective and precise scene representations, and thus our method achieves better or competitive results to the related works. Notably, in ShanghaiTech B which contains more sparse scenes, our point-supervised model outperforms recent density-based approaches such as [9] and [10]. This

Table 1: MAE, MSE, and AP scores on the ShanghaiTech datasets.

Dataset	ShanghaiTech A			ShanghaiTech B		
	Counting		*Loc.*	*Counting*		*Loc.*
	MAE↓	*MSE*↓	*AP*↑	*MAE*↓	*MSE*↓	*AP*↑
Cross-scene [8]	181.1	277.7	-	32.0	49.8	-
MCNN [9]	110.2	173.2	-	26.4	41.3	-
LC-ResFCN [2]	-	-	-	25.89	-	-
LC-PSPNet [2]	-	-	-	21.61	-	-
RAZNet [11]	75.2	133.0	0.69	13.5	25.4	0.69
RAZNet+ [11]	71.6	120.1	0.69	9.9	15.6	0.71
DD-CNN [20]	71.9	111.2	0.65	-	-	-
Deep-Stacked [18]	94.0	150.6	-	18.7	31.9	-
CSRNet [10]	**68.2**	115.0	-	10.6	16	-
Ours	71.4	**110.7**	**0.71**	**9.6**	**15.0**	**0.75**

Fig. 2: Visualized samples of detection and localization. Yellow and red points denote detected and ground truth head locations, respectively.

can be attributed to the fact that density-based method in which the individual heads are not distinctly identified are generally more robust when dealing with denser scenes.

Localization Results. The evaluation results for localization of the detected heads are depicted in Table 1. We observe that our point-supervised model delivers the best results in comparison to prior work on both datasets. It should be noted that the values for [11] and [20] reported in our table are taken directly from the respective papers, which may have used slightly different definitions for true-positive detections given a lack of standard definition in the field. Fig. 2 shows three sample images where the crowd and their locations have been identified, along with the corresponding ground truths.

Ablation Study. We aim to validate the multiscale aspect of our approach by systematically removing each scale branch through ablation experiments. We specifically remove S_i by excluding their corresponding loss terms from our total loss. Each row in Table 2 shows an experiment where crosses indicate the exclusion of the loss term for a particular scale. We observe that for both datasets, the performance degrades when any of the scales are removed.

In order to show that the multiscale fusion module can effectively combine the information extracted from each branch in different densities we partition the images in the dataset into 5 different crowd density groups based on the number of people in each image, as in [18]. Consequently, Group 1

Table 2: Ablation studies on different configurations for crowd counting on the ShanghaiTech datasets.

$\mathcal{L}_1$	$\mathcal{L}_2$	$\mathcal{L}_3$	ShanghaiTech A		ShanghaiTech B	
			MAE ↓	*MSE* ↓	*MAE* ↓	*MSE* ↓
✗	✓	✓	84.7	125.6	10.5	17.3
✓	✗	✓	85.1	129.1	10.4	17.5
✓	✓	✗	85.2	128.6	9.7	15.4
✓	✓	✓	**71.4**	**110.7**	**9.6**	**15.0**

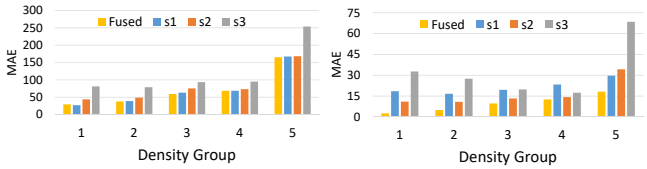

Fig. 3: The MAE for different crowd density groups in ShanghaiTech A (left) and ShanghaiTech B (right).

comprises the first 20$^{\text{th}}$ percentile of density, Group 2 the second 20$^{\text{th}}$ percentile, etc. Next, we calculate the MAE for each branch in order to investigate the impact of our multiscale fusion module in different density groups. The results are illustrated in Fig. 3, where we observe that the fusion network successfully combines outputs from different scales in different densities. These results shows that the fusion network outperforms any single scale for almost every density group, which indicates that the fusion approach performs better than simply averaging the results of each individual scale. In particular, our multiscale fusion approach would yield equal or better performance than switching scales based on prior knowledge of these density groups.

5. CONCLUSION

In this paper, we propose a novel multitasking deep neural network capable of performing both crowd counting and localization simultaneously using point supervision. Our model uses a multiscale architecture and an effective fusion module to deal with the different crowd densities that can occur in images. Our rigorous experiments demonstrate that our proposed model can achieve strong results in comparison to other works in the area on two popular datasets, ShanghaiTech A and B, for both crowd counting and localization tasks. Moreover, our ablation experiments showed the positive impact of the multiscale and fusion elements of our model.

For future work, we may explore avenues for the different terms in our loss function to be automatically weighted through learning. To this end, different attention mechanisms may be explored and integrated in our model. Additionally, using density maps alongside our point-based approach may be explored through ensemble or fusion approaches.

Acknowledgements. Thanks to Geotab Inc., the City of Kingston, and NSERC for their support of this work.

6. REFERENCES

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

[2] I. H. Laradji, N. Rostamzadeh, P. O. Pinheiro, D. Vazquez, and M. Schmidt, "Where are the blobs: Counting by localization with point supervision," in *European Conference on Computer Vision*, 2018, pp. 547–562.

[3] Y. Wang, J. Hou, and L. P. Chau, "Object counting in video surveillance using multi-scale density map regression," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 2422–2426.

[4] A. B. Chan, Z. S. John Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2008, pp. 1–7.

[5] Y. L. Chen, B.F. Wu, H. Y. Huang, and C. J. Fan, "A real-time vision system for nighttime vehicle detection and traffic surveillance," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 5, pp. 2030–2044, 2010.

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8070–8079.

[7] Y. Wang and Y. Zou, "Fast visual object counting via example-based density estimation," in *IEEE International Conference on Image Processing*, 2016, pp. 3653–3657.

[8] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 833–841.

[9] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589–597.

[10] Y. Li, X. Zhang, and D. Chen, "Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1091–1100.

[11] C. Liu, X. Weng, and Y. Mu, "Recurrent attentive zooming for joint crowd counting and precise localization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1217–1226.

[12] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "Cnn-based density estimation and crowd counting: A survey," *arXiv preprint arXiv:2003.12783*, 2020.

[13] Y. Wang, J. Hou, X. Hou, and L. P. Chau, "A self-training approach for point-supervised object detection and counting in crowds," *IEEE Transactions on Image Processing*, vol. 30, pp. 2876–2887, 2021.

[14] 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*, pp. 347–382. 2013.

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

[16] L. Boominathan, Srinivas S.S. K., and R. . Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *24th ACM International Conference on Multimedia*, 2016, pp. 640–644.

[17] L. Zeng, X. Xu, B. Cai, S. Qiu, and T. Zhang, "Multi-scale convolutional neural networks for crowd counting," in *IEEE International Conference on Image Processing*, 2017, pp. 465–469.

[18] S. Huang, X. Li, Z. Q. Cheng, Z. Zhang, and A. Hauptmann, "Stacked pooling for boosting scale invariance of crowd counting," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 2578–2582.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[20] D. B. Sam, S. V. Peri, N. Mukuntha, and R. V. Babu, "Going beyond the regression paradigm with accurate dot prediction for dense crowds," in *IEEE Winter Conference on Applications of Computer Vision*, 2020, pp. 2853–2861.

[21] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[22] K. Wu, E. Otoo, and A. Shoshani, "Optimizing connected component labeling algorithms," in *Medical Imaging: Image Processing*. International Society for Optics and Photonics, 2005, vol. 5747, pp. 1965–1976.

[23] D. B. Sam, S. V. Peri, M. N. Sundararaman, A. Kamath, and V. B. Radhakrishnan, "Locate, size and count: Accurately resolving people in dense crowds via detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[24] Y. Yao, X. Zhang, Y. Liang, X. Zhang, F. Shen, and J. Zhao, "A real-time pedestrian counting system based on rgb-d," in *12th International Conference on Advanced Computational Intelligence*, 2020, pp. 110–117.