A Classification-Regression Deep Learning Model for People Counting

Bolei Xu*, Wenbin Zou*, Jonathan Garibaldi†, and Guoping Qiu*†
*College of Information Engineering, Shenzhen University
xubolei@gmail.com,{wzou,qiu}@szu.edu.cn,
†School of Computer Scinece, University of Nottingham, United Kingdom
jon.garibaldi@nottingham.ac.uk

***Abstract*—In this paper, we construct a multi-task deep learning model to simultaneously predict people number and the level of crowd density. Motivated by the success of applying "ambiguous labelling" to age estimation problem, we also manage to employ this strategy to the people counting problem. We show that it is a reasonable strategy since people counting problem is similar to the age estimation problem. Also, by applying "ambiguous labelling", we are able to augment the size of training dataset, which is a desirable property when applying to deep learning model. In a series of experiment, we show that the "ambiguous labelling" strategy can not only improve the performance of deep learning but also enhance the prediction ability of traditional computer vision methods such as Random Projection Forest with hand-crafted features.**

***Keywords*—People counting; Deep learning; Ambiguous labelling**

I. Introduction

In many application scenarios, there is a need to count the number of people at a scene. For example, in public spaces such as airports and railway stations, knowing the number of people present at the scene can help better manage the space and to ensure public security. With the wide spread installation of visual surveillance cameras almost everywhere in such public space, it is possible to perform automatic people counting through analyzing surveillance videos. Using computer vision and machine learning techniques for people counting has therefore attracted a lot of interest in the literature. However, like many computer vision applications, people counting in video is also a very challenging problem.

In recent years, deep learning neural networks have emerged as a powerful technique for many computer vision problems. In this paper, we are inspired by the significant performance of deep learning on various vision tasks [1], [2], [3] and apply the deep learning method to extract deep feature for the crowd counting problem. In previous work, several kinds of deep learning models have been proposed to address the people counting problem. Zhang *et al.* [4] and Wang *et al.* [5] construct deep networks to directly output people number. Some later works [6], [7] apply deep learning network to produce density map instead of people number to achieve better performance. The density map presents the position of human heads and thus is able to provide people number. However, such method requires to label human position when constructing training datasets, which limits their scalability to the real world application. On the other hand, occlusion is a severe problem in crowd counting. In the case of high density crowd, it is difficult for human to label accurate head positions and provide reliable people numbers for the training datasets.

Inspired by the success of multi-task deep learning method [8], we propose a classification-regression deep learning model which treats the whole surveillance image as the input image, and the deep learning model not only outputs the person number but also estimates the level of crowd density. We show that such multi-task network structure is able to learn more discriminative feature representation than a network solely outputs people number, because the task of estimating the density level could provide a coarse counting number which is less affected by the variation of image scale. In the work of [8], they simultaneously produce density map and 10-way crowd count classification. We differ from their method by predicting people number instead of producing density map. Apart from the aforementioned reasons, directly predicting people number requires less computational resources, since producing density map is usually based on a convolutional layer with filter size of 1×1 to map feature map to the density map. In contrast, the performance of our method is comparable to that of [8] through using only one fully-connected layer after the base network.

In order to address occlusion problem, we also adopt a strategy called "ambiguous labelling" method. The "ambiguous labelling" was first applied to solve the age estimation problem [9], [10], [11], since the faces of neighbouring ages usually present similar image features. Thus, in the previous work of age estimation, authors could assign ambiguous labels to input face images and take the problem as a classification task. We reason that it is also possible to apply "ambiguous labelling" strategy to the crowd counting problem. One reason is that people counting problem is similar to the problem of age estimation, for instance, the image of 500 people is similar to the image of 510 people. On the other hand, the size of people counting dataset is usually small, which is not sufficient to train a deep learning model with large number of parameters. To solve this problem in the deep learning context, "ambiguous labelling" method enables us to create various people number labels for the input image that can augment training dataset for the deep learning model. We provide detail analysis in section III. In the experiment, we show that this method is effective not only for the deep learning model, but also for the traditional computer vision methods such as random projection forest model [12].

II. Related Work

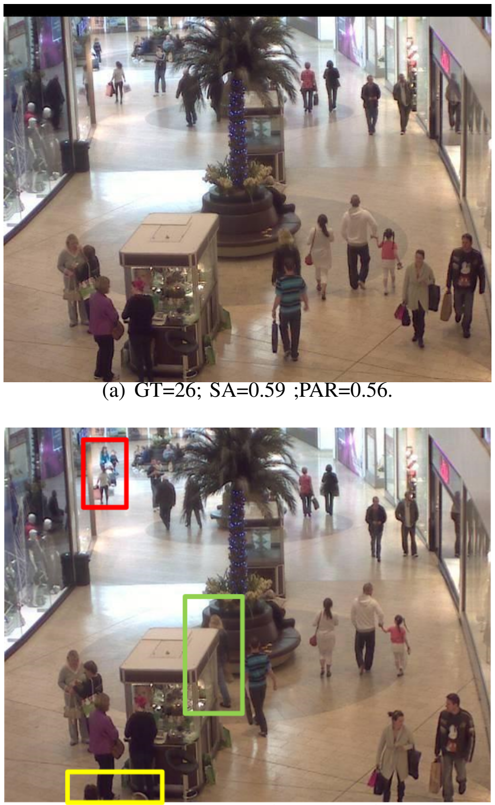
The crowd counting task was initially solved by the detection method. Different kinds of features are used to detect the body of pedestrians including motion features [13], histogram-of-gradients [14] or Bayesian model-based segmentation [15]. However, occlusion becomes a serious problem when applying to estimate high density crowd. Then the part-based detection methods are developed to solve this problem [16], [17]. These methods usually take a long time to count people since they have to exhaustively scan each frame of the video with the trained detector. Another approach is to cluster the trajectories which have coherent motion and then the number of clusters is used to estimate the moving pedestrians [18], [19]. One problem of the clustering method is that it can only provide accurate result when reliable trajectories can be extracted. Thus, this approach is not able to handle the occlusion problem and low video frame rates due to the broken feature tracks. Foroughi *et al.* [20] take the people counting task as a classification problem. They apply sparse representation to capture the hidden structure and semantic information in the image data, and the feature dimension is further reduced by random projection. However, one serious problem with the classification method is if any label information (i.e. the number of people) in the testing set is not included by the training set, this method cannot achieve high accuracy result, which means their algorithm requires large training set to cover almost all the possible situation in the testing set.

A more suitable approach to solving the aforementioned problems is to count by regression. Low-level features are firstly extracted and then mapped to the people number by the regression model. As this kind of approach does not require to detect and track individual person, it has relatively low computational cost and demonstrates promising results on solving the occlusion problem. A variety of features have been used by previous works to estimate the crowd density, such as total area [21], [22], edge count [23], [24] and texture features [25]. Chan *et al.* [26] take the perspective distortion into account and experiment with additional features such as Minkowski fractal dimension to estimate the irregularity of edges.

The traditional approaches are suffering from two main problems. Firstly, they heavily rely on the background segmentation techniques to remove noise. Secondly, an unavoidable step in the traditional approaches is to extract hand-crafted features. However, designing hand-crafted features is not an easy step and it is usually difficult to find out optimal hand-crafted feature representation. The deep learning approach can well-solve both problems. It does not have to apply background segmentation method to pre-process images and it is able to count people number from different perspectives [6]. Another advantage is that deep learning can be constructed as an end-to-end model, which takes whole image as input and outputs people number or the head position. It means feature designing is not a necessary step when applying deep learning.

Some previous work apply deep learning method to address the problem of people counting. At the initial stage, the deep learning framework is usually employed to directly output people number. Zhang *et al.* [4] propose a Convolutional Neural Network (CNN) based framework to extract deep features of crowd scene and use a data-driven method to fine-tune the CNN model to the target scene. Wang *et al.* [5] also construct a deep network in order to estimate extremely dense crowds. Marsden *et al.* [27] apply a scale aware deep learning model with a single column fully convolutional network that takes multiple scales of image as the input in the prediction stage. Each scale of image produces a people number and the final counting number is to take the average of these estimates.

Apart from directly predicting people number, another way to apply deep learning is to generate density map and then count people number from density map. Zhang *et al.* [6] first develop this method to count people number from density map. They use a Gaussian kernel to convolve a labelled image and then compute people number by summarizing pixel value. There are also some following work to produce density map based on deep learning approach. Boominathan *et al.* [28] combine one deep network and one shallow network to predict a density map for a given crowd image. Sindagi *et al.* [8] propose a cascaded deep network structure to simultaneously classify crowd into different levels and produce density map. However, the approach based on density map has to label the head positions for the whole dataset, which is a time-consuming process when applying to the high density crowd or the large scale datasets.


(a) GT=26; SA=0.59 ;PAR=0.56.

(b) GT=31; SA=0.57; PAR=0.57.

Fig. 1. Both images are captured from Mall dataset. 'GT' refers to ground-truth, 'SA' refers to segment area and 'PAR' refers to perimeter-area ratio. All values are provided by the original dataset author.

III. Application of Ambiguous Labels to People Counting

We here illustrate the rationales that we apply "ambiguous labelling" strategy for the people counting problem.

Firstly, we show that people counting problem is similar to the age estimation problem. Fig. 1 presents a typical case in the people counting problem. The ground-truth number for Fig. 1(a) is 26 persons while the person number in Fig. 1(b) is 31. Although the people numbers are totally different, the major contents of both images are very close. It is confirmed by the traditional features extracted from both images. Two main features (segment area and perimeter-area) employed by the previous work [26] are almost the same. If we look into the details of both images, there are three minor differences leading to different person numbers: (1) In the red bounding box, a woman is pushing a stroller for a baby but the size of baby body is small in the image. (2) In the green bounding box, a walking woman's body is occluded by an obstruction while only part of woman body is shown in image. (3) In the yellow bounding box, three persons' heads appear on the image. However, only piece of their heads can be seen in the image. Thus, we can see that similar image features do not always refer to the same person number. It is the same as the age estimation problem that neighbouring age might present similar image features. This is the main reason that we could assign various labels to each input image as done in the previous age estimation work.

Secondly, "ambiguous labelling" strategy enables us to create augmented training dataset for deep learning model. As insufficient training data could lead to over-fitting problem, a desirable training dataset should have multiple images for each image label. However, the mainstream people counting datasets (UCSD and Mall datasets) usually contain limited number of images for each people number. Consequently, we could improve the predicting ability of model by enlarging the size of training dataset. By assigning various labels to the images in the training dataset, we can obtain a much larger size of training datasets than that of the original one. It means for each specific people number (training label), we can find a variety of crowd scenes (training image) in the training dataset. The deep learning model can thus learn more discriminative features with sufficient number of training images.

IV. Label Ambiguity Construction

In this section, we introduce our method to model the randomness of people number and thus to create ambiguous labels for each input surveillance image. For each scalar-valued people number label $l \in \mathbb{R}$ of the input image, we seek a label distribution that should satisfy two criteria: (1) the ground truth value should have the highest possibility of being assigned to the image; and (2) when the labels are farther from the ground truth, they should be assigned to the image with lower probabilities. In this paper, we adopt the Gaussian distribution in the experiment to model the ambiguous labels for each surveillance image as shown in Fig.2, whose mean value μ is equal to the ground-truth value. The corresponding standard deviation σ for the Gaussian distribution is usually an unknown factor but can work well when it is carefully chosen [11]. We thus empirically set σ to 2 in the experiment. By constructing a Gaussian distribution, we can randomly sample M labels for each input image. As the problem of occlusion usually appears in the relatively high density crowd, we only apply the "ambiguous labelling" strategy to the images of people number over 15.

V. Deep Classification-regression Learning Model

In this paper, we do not apply Resnet [1] or VGG deep learning model [2] as the base convolutional network to address the problem. The reason is that the size of crowd counting datasets is relatively small (usually around 2000 images), which is not sufficient to train the Resnet or VGG network with large number of parameters. For this crowd counting problem, we construct the convolutional network based on a custom network structure as shown in Fig.3. We construct the multi-task deep learning model by connecting two parallel sub-networks to the base convolutional network. One sub-network is used to predict people number and another sub-network is used to estimate the crowd density level.

The people counting network is consisting of one fully-connected layer with 256 neurons and Rectified Linear Unit (ReLU) is taken as the activation function. This branch finally produces people number $\hat{l_k}$ for the input image x_k with label l_k, and we use Mean Squared Error (MSE) as the objective function for this branch:

$$\mathcal{L}_{MSE} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{2}\|l_k - \hat{l_k}\|^2. \tag{1}$$

The classification layer aims to classify input image to one of the density levels. We create classification labels for each dataset with an interval of 10 people. For instance, if the maximum people number in the training dataset is 100, then we can create 11 labels for the dataset. The level-1 density refers to the people number of 0 to 10, and level-2 refers to people number of 11 to 20. The rest can be done in the same manner where level-11 refers to the people number above 100. The classification layer also contains a fully-connected layer that has 256 neurons with ReLU activation function. We use softmax function as classifier and use the cross-entropy error as the loss function:

$$\mathcal{L}_{level}(p,q) = -\sum_{x} p(x)\log q(x), \tag{2}$$

where p is the ground-truth distribution of density level, and q is the estimated class probabilities produced by the softmax classifier. Then the total loss for the whole deep learning model can be written as:

$$\mathcal{L}_{total} = \lambda\mathcal{L}_{MSE} + \mathcal{L}_{level}, \tag{3}$$

where λ is a weighting factor.

VI. Experiment

A. Experiment Setup

For the parameter settings, we initialize the whole deep network with Gaussian distribution of zero mean and set its standard deviation to 0.01, and bias to zeros. We empirically set $\lambda = 2$ in Eq.3. We then optimize the network by Stochastic Gradient Descent (SGD) with a learning rate of 0.01 and the size of mini-batches is 128. In the experiment, the network usually convergences around 30 epochs. We conduct all the experiments over the UCSD pedestrian dataset and Mall dataset. When creating ambiguous labels for each dataset, we randomly sample $M = 5$ labels from Gaussian distribution for

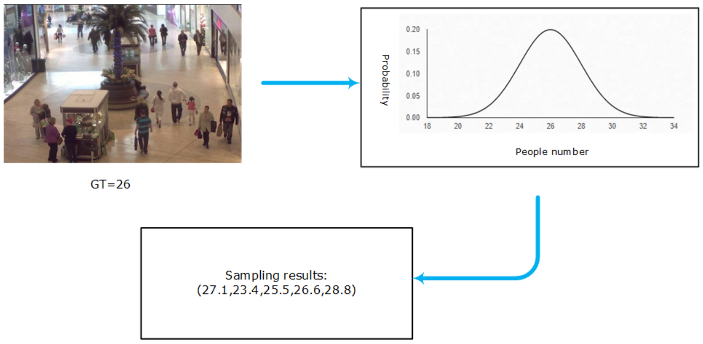


Fig. 2. The process of how to assign ambiguous labels to an input image. The ground-truth value of input image is regarded as the mean value μ for the Gaussian distribution. We randomly sample $M = 5$ labels for each image.

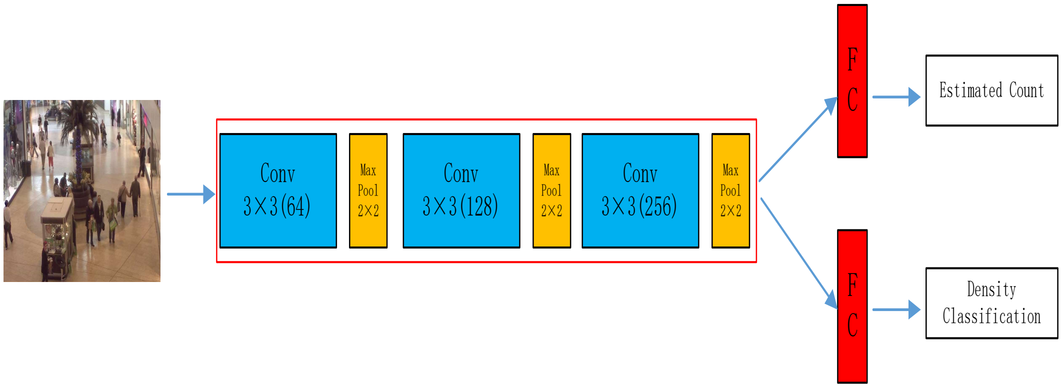


Fig. 3. The network architecture of our classification-regression deep learning model. The regression branch outputs the accurate crowd density, while the classification branch predicts the coarse people number.

each image. The input image is resized to 256×256 for the deep learning model.

We test our proposed algorithm on the UCSD pedestrian database [26] and Mall dataset [29], which are two well-known datasets on the evaluation of people counting algorithms. Both datasets contain 2000 frames that are captured by a stationary camcorder from outdoor and indoor scene respectively. The example images from two datasets are shown in Fig. 4.

We separate the datasets as previous work: in UCSD dataset, frames 601-1400 are employed for training; in Mall dataset, the first 800 frames are used. The rest frames in each dataset are applied for testing. Two evaluation metrics are applied for numerical testing and comparison with the-state-of-art algorithms. The first one is called *mean absolute error* (MAE) to estimate the average absolute error of each testing frames:

$$\epsilon_{mae} = \frac{1}{N}\sum_{i=1}^{N}|m_i - \widetilde{m}_i|, \qquad (4)$$

where N is the total number of test images, m_i is the ground truth for ith test image, and $\widetilde{m}_i$ is the corresponding prediction result. The second one is *mean squared error* (MSE) which assesses the average mean squared error:

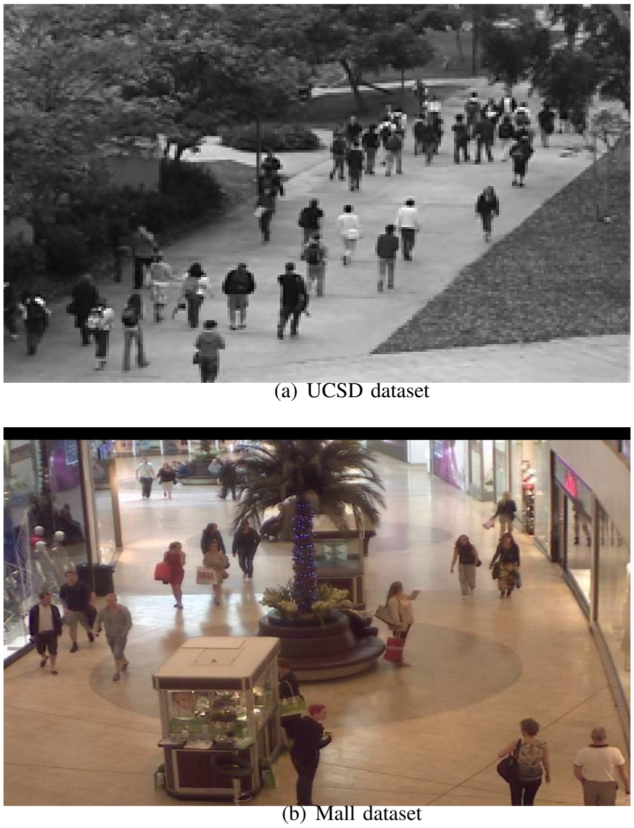
$$\epsilon_{mse} = \frac{1}{N}\sum_{i=1}^{N}(m_i - \widetilde{m}_i)^2. \qquad (5)$$

TABLE I. PERFORMANCE COMPARISON WITH THE TRADITIONAL COMPUTER VISION METHODS ON TWO DATASETS.

Model	UCSD		Mall	
	MAE	MSE	MAE	MSE
GPR [26]	2.24	7.97	3.72	20.10
MORR [29]	2.25	7.82	3.59	19.00
CA-RR [30]	2.07	6.86	3.43	17.70
RPF (hf)	1.90	6.01	3.22	15.52
RPF (fc1)	1.78	5.46	3.02	13.80
RPF (fc2)	1.62	4.84	2.86	11.44
Ours	**1.48**	**3.24**	**2.73**	**10.20**

B. Comparing with Hand-crafted Features

In the first experiment, we compare our deep learning method with the traditional computer vision methods including our random projection forest that employ hand-crafted features. Table.I presents the results of this experiment. It can be seen that our deep learning method significantly outperforms other traditional methods. We also conducted an experiment on the Random Projection Forest (RPF) [12], which employs different kinds of feature. One is the same hand-crafted features (hf) as [26], and another one is the deep feature from the FC layer in the regression branch (fc1), and the FC layer in the classification branch (fc2). It can be seen that the deep features from deep learning model are more discriminative than the hand-crafted features, and the features from fc1 is better than that from fc2, which is caused by the regression branch is able to predict more detail people density scenario than the classification branch.



(a) UCSD dataset

(b) Mall dataset

Fig. 4. Crowd scenes of UCSD dataset and Mall dataset. The UCSD dataset captures outside scene while the Mall dataset captures indoor scene.

C. *Comparing with CNN-based Approaches*

TABLE II. PERFORMANCE COMPARISON WITH THE CNN-BASED METHODS ON TWO DATASETS, WHERE '-' INDICATES NO RESULT REPORTED.

Model	UCSD		Mall	
	MAE	MSE	MAE	MSE
Zhang *et al.* [4]	1.60	3.31	-	-
Kumagai *et al.* [31]	-	-	2.75	13.40
Sam *et al.* [32]	1.62	**2.10**	-	-
Sheng *et al.* [33]	2.86	13.0	**2.41**	**9.12**
Ours	**1.48**	3.24	2.73	10.20

We also compare our method with the CNN-based approaches. These approaches include Zhang *et al.* [4], Kumagai *et al.* [31], Sam *et al.* [32], and Sheng *et al.* [33].

Form Table. II we can see that our CNN method achieves the best performance on the UCSD dataset with MAE as the evaluation criteria, and slightly worse performance than Sam *et al.* [32] on the MSE evaluation. On the Mall dataset, our deep learning approach provides comparable performance when comparing to other CNN methods. Comparing with other approaches, the classification branch in our model can provide a coarse estimation to the people density, which is less influenced by the variation of perspectives and image scale.

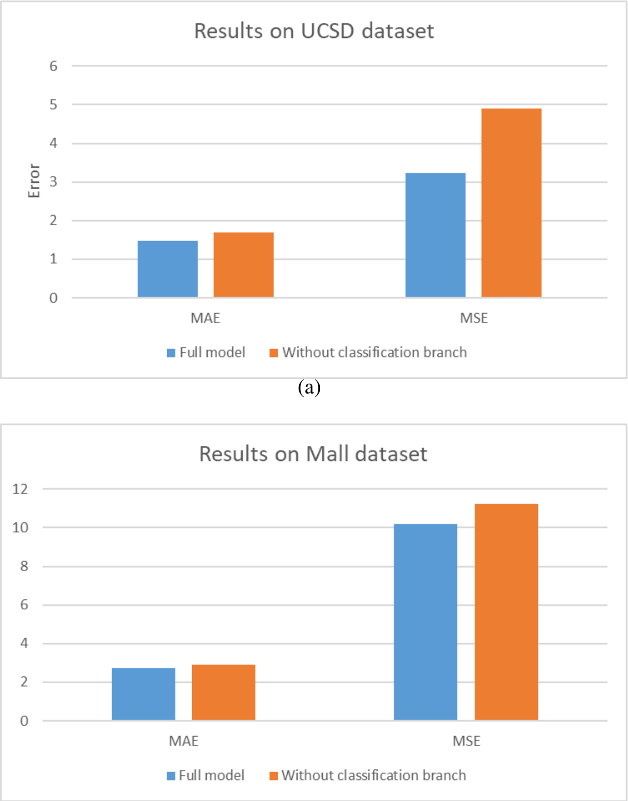
D. *Evaluation of Ambiguous Labelling*

Then we conduct an experiment to evaluate the effectiveness of ambiguous labelling strategy. We apply ambiguous labelling method to both deep learning model and also the random projection forest model. From Table. III we can seen that by employing ambiguous labelling method can increase the performances of both deep learning model and the random projection forest model with larger size of training dataset. It confirms the effectiveness of ambiguous labelling method that it is not only effective on age estimation problem in previous work but also helpful on the crowd estimation problem.

TABLE III. EVALUATION OF AMBIGUOUS LABEL ON RANDOM PROJECTION FOREST AND DEEP LEARNING MODEL.

Model	UCSD		Mall	
	MAE	MSE	MAE	MSE
RPF without ambiguous labels	1.90	6.01	3.22	15.52
RPF with ambiguous labels	1.78	5.10	3.04	13.82
Deep learning without ambiguous labels	1.62	4.82	2.92	12.61
Deep learning with ambiguous labels	1.48	3.24	2.73	10.20

E. *Evaluation of Necessity of Classification*



(a)

(b)

Fig. 5. Comparison results on UCSD and Mall datasets with and without classification branch. Figure (a) is the results on UCSD dataet, and figure (b) is the results on Mall dataset.

As we propose a multi-task deep learning model, it is also necessary to evaluate the necessity of the classification branch in the deep learning model. We compare two models: one is the full model and another one is the model without the classification branch. From Fig.5 we can see that it is necessary to include the classification branch to the model. The classification branch provides a coarse counting number that is less influenced by the image scale and the variation of perspectives. Thus, from the experiment result, we can see that the full model with two branches shows much better performance than the model without the classification branch on both datasets.

F. Evaluation of the Influence of Dataset Size

TABLE IV. We construct four different models to evaluate how the size of training dataset affect the prediction performance.

Model	MAE	MSE
Model 1	1.32	2.89
Model 2	1.48	3.24
Model 3	2.52	9.00
Model 4	2.73	10.20

One inevitable problem when applying deep learning model is the size of dataset. Insufficient training dataset size would lead to the over-fitting problem and reduce the generalization ability of the model.

In this experiment, we modify the training dataset to evaluate the influence of dataset size. When testing on the UCSD dataset, we also include the whole Mall dataset into the training dataset. When testing on the Mall dataset, we add the whole UCSD dataset to the training dataset. It results in four kinds of model:

1) Model 1: Training on the training dataset of UCSD and whole dataset of Mall, and testing on the testing dataset of UCSD.
2) Model 2: Training on the training dataset of UCSD, and testing on the testing dataset of UCSD.
3) Model 3: Training on the training dataset of Mall and whole dataset of UCSD, and testing on the testing dataset of Mall.
4) Model 4: Training on the training dataset of Mall, and testing on the testing dataset of Mall.

From Table.IV we can see that when the training data grows, the performance produced by the deep learning increases as well. It verifies the assumption that the larger dataset would lead to the better performance when applying deep learning model.

VII. Conclusion

In this paper, we have constructed a multi-task deep learning model for the crowd estimation problem. We show that the deep learning method is able to outperform previous computer vision methods based on hand-crafted features. Apart from employing deep feature, we propose an ambiguous labelling method to create various label for each input image. The experiment result confirms the effectiveness of the ambiguous labelling method, which is able increase the performance of both deep learning method and also our previous random projection forest method.

References

[1] 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.

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

[3] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li, "Deep reconstruction-classification networks for unsupervised domain adaptation," in *European Conference on Computer Vision*. Springer, 2016, pp. 597–613.

[4] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. CVPR*, 2015.

[5] 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*. ACM, 2015, pp. 1299–1302.

[6] 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.

[7] 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.

[8] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," *arXiv preprint arXiv:1707.09605*, 2017.

[9] K. Chen and J.-K. Kämäräinen, "Learning with ambiguous label distribution for apparent age estimation," in *Asian Conference on Computer Vision*. Springer, 2016, pp. 330–343.

[10] X. Geng, Q. Wang, and Y. Xia, "Facial age estimation by adaptive label distribution learning," in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 4465–4470.

[11] B.-B. Gao, C. Xing, C.-W. Xie, J. Wu, and X. Geng, "Deep label distribution learning with label ambiguity," *IEEE Transactions on Image Processing*, vol. 26, no. 6, pp. 2825–2838, 2017.

[12] B. Xu and G. Qiu, "Crowd density estimation based on rich features and random projection forest," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–8.

[13] P. Viola, M. J. Jones, and D. Snow, "Detecting pedestrians using patterns of motion and appearance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*. IEEE, 2003, pp. 734–741.

[14] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[15] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," in *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, vol. 2. IEEE, 2003, pp. II–459.

[16] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1. IEEE, 2005, pp. 90–97.

[17] S.-F. Lin, J.-Y. Chen, and H.-X. Chao, "Estimation of number of people in crowded scenes using perspective transformation," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 31, no. 6, pp. 645–654, 2001.

[18] G. J. Brostow and R. Cipolla, "Unsupervised bayesian detection of independent motion in crowds," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 594–601.

[19] V. Rabaud and S. Belongie, "Counting crowded moving objects," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 1. IEEE, 2006, pp. 705–711.

[20] H. Foroughi, N. Ray, and H. Zhang, "Robust people counting using sparse representation and random projection," *Pattern Recognition*, 2015.

[21] N. Paragios and V. Ramesh, "A mrf-based approach for real-time subway monitoring," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–1034.

[22] S.-Y. Cho, T. W. Chow, and C.-T. Leung, "A neural-based crowd estimation by hybrid global learning algorithm," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 29, no. 4, pp. 535–541, 1999.

[23] A. C. Davies, J. H. Yin, and S. A. Velastin, "Crowd monitoring using image processing," *Electronics & Communication Engineering Journal*, vol. 7, no. 1, pp. 37–47, 1995.

[24] C. S. Regazzoni and A. Tesei, "Distributed data fusion for real-time crowding estimation," *Signal Processing*, vol. 53, no. 1, pp. 47–63, 1996.

[25] A. Marana, L. d. F. Costa, R. Lotufo, and S. Velastin, "On the efficacy of texture analysis for crowd monitoring," in *Computer Graphics, Image Processing, and Vision, 1998. Proceedings. SIBGRAPI'98. International Symposium on*. IEEE, 1998, pp. 354–361.

[26] A. B. Chan, Z.-S. 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*. IEEE, 2008, pp. 1–7.

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

[28] 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*. ACM, 2016, pp. 640–644.

[29] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting." in *BMVC*, vol. 1, no. 2, 2012, p. 3.

[30] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 2467–2474.

[31] S. Kumagai, K. Hotta, and T. Kurita, "Mixture of counting cnns: Adaptive integration of cnns specialized to specific appearance for crowd counting," *arXiv preprint arXiv:1703.09393*, 2017.

[32] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," *arXiv preprint arXiv:1708.00199*, 2017.

[33] B. Sheng, C. Shen, G. Lin, J. Li, W. Yang, and C. Sun, "Crowd counting via weighted vlad on dense attribute feature maps," *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.