Leveraging Heterogeneous Auxiliary Tasks to Assist Crowd Counting

Muming Zhao [1,2], Jian Zhang [2,4], Chongyang Zhang [1,3]*, Wenjun Zhang [1]
[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
[2]University of Technology, Sydney
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[4]Peng Cheng Laboratory, Shenzhen, China
muming.zhao@student.uts.edu.au, jian.zhang@uts.edu.au, {sunny_zhang, zhangwenjun}@sjtu.edu.cn

Abstract

Crowd counting is a challenging task in the presence of drastic scale variations, the clutter background, and severe occlusions, etc. Existing CNN-based counting methods tackle these challenges mainly by fusing either multi-scale or multi-context features to generate robust representations. In this paper, we propose to address these issues by leveraging the heterogeneous attributes compounded in the density map. We identify three geometric/semantic/numeric attributes essentially important to the density estimation, and demonstrate how to effectively utilize these heterogeneous attributes to assist the crowd counting by formulating them into multiple auxiliary tasks. With the multi-fold regularization effects induced by the auxiliary tasks, the backbone CNN model is driven to embed desired properties explicitly and thus gains robust representations towards more accurate density estimation. Extensive experiments on three challenging crowd counting datasets have demonstrated the effectiveness of the proposed approach.

1. Introduction

Crowd counting and density estimation are of great importance in computer vision due to its essential role in a wide range of surveillance applications including physical security, public space management, and retail space design [11, 38]. However, the presence of drastic scale variations, the clutter background, and severe occlusions make it challenging to generate high-quality crowd density maps.

Various CNN-based counting methods [35, 36, 29, 12] have been proposed to handle the challenging situations mainly by fusing multi-scale or multi-context information to improve the feature representations. For example, Zhang *et al.* [36] generate multi-scale features with the multi-column network to handle scale variations. Sindagi *et al.* [29] fuse local and global features for density estimation. Their successes demonstrate the effectiveness to incorporate information from various sources (i.e., different sub-models). Motivated by these methods, we propose to leverage heterogeneous attributes of the density map as guidances to fully exploit the potential of the underlying representation, without explicit modifications to the features.

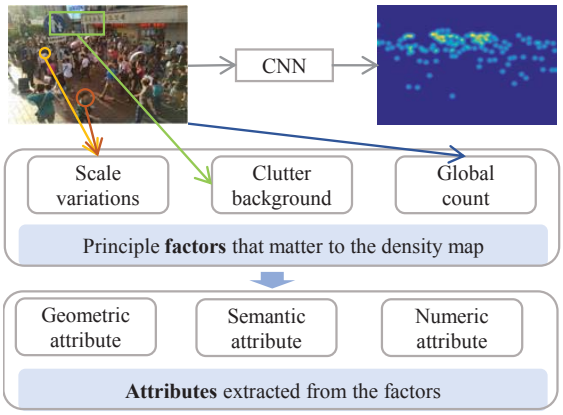

Figure 1. Illustration of essential factors influencing the crowd density estimation.

Figure 1 illustrates our motivation with the observation of three factors of the density estimation. Considering the formulation of density-estimation-based counting paradigm [11] which sums the density values over any region to report the final count, it is desired the estimated densities vary along with object scales given the factor of intra-image scale variations of crowd images. Specifically, the nearer, larger objects should have smaller density values compared to farther objects with smaller scales. We term this as the geometric attribute of the density map. Besides, the clutter background is another factor that should not be neglected. For more accurate density estimation, the density distribution is also desired to conform with the spatial distributions of the crowd to avoid the background clutter, which can be viewed as the semantic attribute of the density esti-

*This is the corresponding author.

mation. Additionally, the global count is also an important indicator measuring the overall density level of one certain image, which can be termed as the numeric attribute of the density estimation. These attributes are heterogeneous and cater for different aspects of crowd images, which should be beneficial to the quality of the density map predictions.

Inspired by these observations, in this paper we propose to leverage the heterogeneous attributes compounded in the density map to improve crowd counting. Specifically, we formulate each attribute as an auxiliary task. For the geometric attribute, we propose the monocular depth prediction to emphasize the relative depth variations of the crowd image, considering that generally the scale variation of one certain object across the scene is inversely proportional to the depth. For the semantic attribute, we introduce the crowd segmentation to highlight the foreground over the background. For the numeric attribute, we introduce the direct count estimation to take care of the overall count accuracy while optimizing per-pixel density. Learning of the auxiliary tasks will drive the intermediate features of the backbone CNN to embed desired information on geometry, semantics and the overall density level, which benefits the generation of robust features against the scale variations and clutter background. Although more objectives are involved, they are readily available either with external models or can be inferred directly from the original density map, which do not need any additional annotations. Furthermore, the formulation of the essential attributes as auxiliary tasks can benefit any backbone CNN model for crowd counting without increasing additional computations at inference, which further introduces flexibility to the proposed approach.

We highlight the main contributions of this work as follows:

- We propose to improve crowd counting by leveraging three heterogeneous attributes compounded in the density map, which influences the quality of the density estimation.
- We formulate each attribute as an auxiliary task, which together provide joint regularization effects to the backbone CNN for more robust representations and density estimation.
- We demonstrate the effectiveness of the proposed method on three challenging datasets, which outperforms the state-of-the-art methods on the ShanghaiTech dataset [36] and the worldExpo'2010 dataset [35], and also achieves very competitive performance on the Mall dataset.

2. Related Work

Numerous methods have been proposed for crowd counting. Detection-based [26, 31] approaches are usually limited by challenging situations of severe occlusions in extremely crowded scenes. As a result, regression-based methods [6, 7] are proposed, which learns a mapping function from holistic crowd features to the global count. However, these early methods mainly use hand-crafted features and have been surpassed by the deep features extracted from CNNs [9].

Recently, deep CNNs have brought a new era for the computer vision society. As one of the earliest CNN-based methods, Zhang *et al.* [35] train a deep model to estimate the crowd density map and count in a switchable learning process. To handle scale variations in the crowd images, Zhang *et al.* [36] introduce the multi-column CNN with different receptive field sizes in each column for multi-scale feature fusion for density estimation. Similarly, a pyramid of input patches is used for the network in [16] to generate multi-resolution features. Recently, Li *et al.* [12] adapt the VGG model [27] with dilation processing and achieve state-of-the-art performance on several benchmark datasets. Introducing novel deep architectures has benefited the learning of more robust features and thus boosts the counting performances.

Other researchers dedicated to incorporate various modules conveying contextual/scale information to improve the base CNN. These work include the multi-context fusion in [29] where global and local contextual information is additionally learned and combined with features from the base model for density estimation. Another typical work is [22] where a switch network is built to relay each input patch into different sub-network [36] for density estimation other than aggregating features from all the sub-models. The switch module in this method is considered to convey the information on intra-image density variations. To handle influences induced by scale variations, in [24] an adversarial learning framework is proposed to pursue cross scale consistency. Recently, a top-down feedback gating module is proposed in [20], which introduces multiplicative feedback to original feature of the base model. The feedback module can be viewed to learn the correction signal towards a good density map estimation.

Instead of augmenting the base CNNs with additional modules, we enhance the features by mining the potential of a model itself with the formulation of auxiliary tasks. From this perspective, our work is also related to multi-task learning [2], where a shared representation is learned for multiple tasks. The effectiveness of multi-task learning in deep CNNs has been validated in various tasks [37, 18] and inspire us to explore its benefits for crowd counting. A similar work is proposed in [28] where the count group classification is learned as a high-level prior and cascaded with the features in the density estimation branch. Although both are using the tool of multi-task learning for counting, our approach differs in the analysis and the disentanglement of the heterogeneous attributes emerging from the density es-

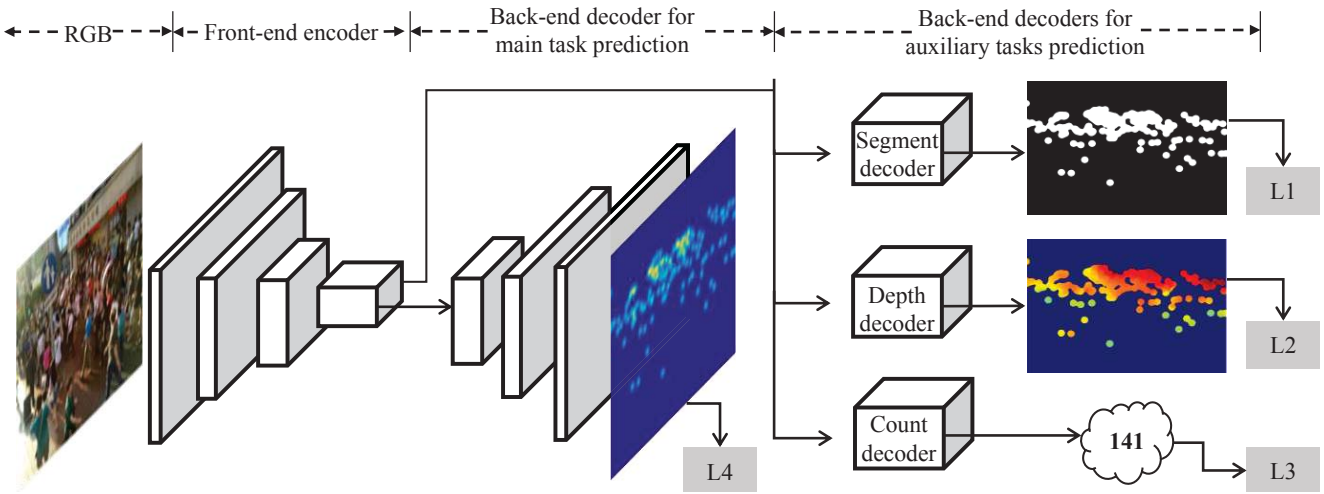

Figure 2. Overview of the proposed approach with the learning of three auxiliary tasks in CNNs (AT-CNN). The symbols of L1 to L3 denote the losses to optimize the auxiliary tasks of crowd segmentation, depth prediction and count regression. The symbols of L4 is the loss for the main task of density estimation.

timation, especially for the handling of the scale variation and the clutter background, which has not fully exploited in existing methods.

3. Methodology

As discussed in Section 1, we propose to leverage heterogeneous attributes to assist crowd counting, which mainly aims to improve the feature representations of the backbone CNN with the learning with auxiliary tasks (AT-CNN). Generally, the crowd density estimation can be viewed as an encoding-decoding process with a front-end CNN (encoder) mapping the input image to high-dimensional feature maps and a back-end CNN (decoder) interpreting the features from the encoder into pixel-wise density values. Denoting the front-end CNN as a function g^e parameterized with $\mathbf{w}^e$, then the features F from the encoder can be represented as $F = g^e(\mathbf{X}; \mathbf{w}^e)$ for an input image $\mathbf{X}$. For any given backbone CNN model, our method constructs the auxiliary tasks prediction (AT) module which uses the deep features F from the front-end CNN to optimize the auxiliary predictions and inversely improve the intermediate representations itself. The framework of our method is shown in Figure. 2. During training, ground-truth labels for the density estimation and the three auxiliary tasks, *i.e.*, depth prediction, crowd segmentation and count estimation are used. Although four different kinds of supervision signals are involved, we do not require any extra annotation effort. Specifically, we exploit modern CNN-based depth prediction models to derive the ground-truth labels for the auxiliary depth prediction. Ground-truth information for crowd segment and count can be directly inferred from the density map labels, respectively.

3.1. Auxiliary Tasks Prediction

Based on the deep features from the front-end CNN, we build the three auxiliary tasks, *i.e.*, crowd segmentation, depth prediction and the count estimation. These three tasks, with each in charge of different characteristics of the density map, can provide multi-fold regularization effects to optimize the front-end CNN. We describe the details for each auxiliary task in the following article.

Attentive Crowd Segmentation Due to the complex situations such as the extremely limited pixels of pedestrians occupied in the image as well as the clutter background, the crowd density map is usually noisy. Towards this problem, we introduce the crowd segmentation as an auxiliary task, which will help the front-end CNN generate more discriminative representations and thus purify the output prediction.

A segmentation decoder network g^{seg} parameterized with $\mathbf{w}^{seg}$ is built as the back-end CNN for crowd segmentation. Performing a two-way classification task, the decoder accepts feature F from the front-end encoder and outputs a crowd segment $\widehat{\mathbf{S}}$ with values indicating the probability of pixels belonging to the targets: $\widehat{\mathbf{S}} = g^{seg}(F; \mathbf{w}^{seg})$. Ground-truth labels for crowd segmentation can be inferred from the dotted annotations of pedestrians provided in counting dataset [36, 35] by simple binarization as shown in Figure 3. We dubbed the result an as *attentive* crowd segment, since it conveys important information clarifying the attentive areas occupied by the targeted objects. Strictly speaking the derived segment map is not the same as the ones in semantic segmentation [5] where detailed boundaries of objects are depicted, however we show in experiments that this simple strategy can yield effective improvements for density estimation.

Given a pair of input image and the ground-truth atten-

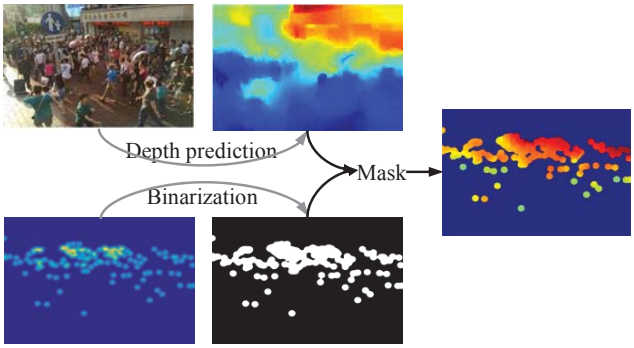

Figure 3. Label generation for auxiliary tasks. Given a pair of crowd image and its ground truth density map (the first column), the depth map can be estimated using external depth prediction algorithms [13] and the crowd segment is inferred through binarization of the density map (the second column). The distilled depth map (the third column) used to supervise the auxiliary task is obtained by masking the originally estimated depth map with the crowd segment map.

tive crowd segmentation map $\{\mathbf{X}, \mathbf{S}\}$, loss function for the segmentation decoder is the binary cross-entropy between the predicted and the ground-truth probability of each pixel:

$$L_1 = \frac{1}{|\mathbf{X}|} \sum_{(i,j)\in\mathbf{X}} t_{ij} log o_{ij} + (1 - t_{ij}) log(1 - o_{ij}), \quad (1)$$

where $t_{ij} \in \{0, 1\}$ is the actual classes of pixels in $\mathbf{S}$ with 1 for the target area and 0 for the background, and o_{ij} denotes the pixel-wise probability in the prediction $\widehat{\mathbf{S}}$.

Distilled Depth Prediction To handle the perspective distortion in surveillance scenes [30], we introduce the single-image monocular depth prediction as an auxiliary task. Informally speaking, for a given object category (*e.g.* pedestrians) the size of an object in the image is inversely proportional to the distance from the camera [8]. In the regions with larger depth values, the objects have smaller sizes and should be adversely assigned with larger density values to guarantee their summation gives accurate counts. By inferring the depth maps, the front-end CNN is imposed to take care of the scene geometry and hence gains the awareness of the intra-image scale variations, which will help generate more discriminative features for scale-aware density estimation.

Similar to the task of crowd segmentation, a depth decoder network g^{dep} parameterized with $\mathbf{w}^{dep}$ is built for depth prediction. The input to the decoder is the features F from the front-end CNN and the output is the depth map with values indicating the distances of each pixel to the camera: $\widehat{\mathbf{D}} = g^{dep}(F; \mathbf{w}^{dep})$.

Towards this task, we resort to depth maps derived from a CNN-based single-image depth prediction model [13] (DCNF) for monocular depth prediction. The DCNF model can estimate depths for general scenes with no geometric priors nor any extra information injected, and hence is suitable in our situation to help illustration of geometry in crowded scenes. Given the input crowd image $\mathbf{X}$, we use the pre-trained DCFN model [13] to generate a *raw* measurement of depth $\mathbf{D}_{raw}$. As observed in Figure 3, it is capable of depicting depth disparities between pedestrians at different positions. However, due to the DCNF model has not been specifically adapted to the target scenes in the crowd counting tasks and hence the depth predictions contain clutter that degrades the efficiency, especially for background areas. Towards this problem, we further calculate a *distilled* depth map $\mathbf{D}$ which only preserve the depth information of the attentive target areas. This is derived using both the raw depth map and the attentive crowd segment: $\mathbf{D} = \mathbf{S} \odot \mathbf{D}_{raw}$, where $\odot$ denotes the Hadamard matrix multiplication. With the distilled depth as the supervision for depth prediction, the front-end CNN is desired to be especially aware of the depth relationships/scale variation between those attentive areas with target objects.

With the training pairs of $\{\mathbf{X}, \mathbf{D}\}$, the depth decoder can be trained using a simple Euclidean loss for the predicted depth map $\widehat{\mathbf{D}}$:

$$L_2 = \frac{1}{|\mathbf{D}|} \sum_{(i,j)\in\mathbf{D}} \left\| \widehat{\mathbf{D}}_{ij} - \mathbf{D}_{ij} \right\|_2^2 \quad (2)$$

Crowd Count Regression Most density estimation based counting algorithms optimize their counting model by measuring the per-pixel errors between the predicted and the ground-truth density maps [35, 36, 22, 16, 29]. However, one problem is this supervision is not directly related to the evaluation metric of MAE/MSE [15] which measures global counting errors of input images. To this end, we introduce another auxiliary task of crowd count regression which directly estimates the crowd count from the encoded features. Empowered with this auxiliary task, the front-end encoder will generate features adapted to the overall density level of the input image, which helps produce more accurate density values.

A count decoder g^{num} parameterized with $\mathbf{w}^{num}$ is built to map the features F from the front-end encoder to the crowd count $\widehat{C}$: $\widehat{C} = g^{cnt}(F; \mathbf{w}^{cnt})$. The ground-truth count C can be directly derived by summing up all the dotted annotations in an input image $\mathbf{X}$. The L_2 norm is used to train the count decoder:

$$L_3 = \left\| \widehat{C} - C \right\|_2^2 \quad (3)$$

3.2. Main Tasks Prediction

The density estimation decoder g is built on the features F emitted from the front-end encoder to perform the main task of density estimation. To generate the ground-truth density maps, we follow [11] to apply 2D Gaussian kernels

on each dotted annotations, where the same-spread (sigma Σ) Gaussian kernels are simply adopted at different positions. The decoder for the main task is trained using the Euclidean loss for the density map $\widehat{\mathbf{Y}}$:

$$L_4 = \frac{1}{|\mathbf{Y}|} \sum_{(i,j)\in\mathbf{Y}} \left\| \widehat{\mathbf{Y}}_{ij} - \mathbf{Y}_{ij} \right\|_2^2 \quad (4)$$

3.3. Optimization

The final learning objective function utilizes multiple losses weighted by hyper-parameters:

$$L_{mt} = \sum_{i=1}^{4} \lambda_i L_i \quad (5)$$

We employ a stage-wise procedure to train the network with auxiliary tasks, by varying the hyper-parameters as detailed in Section 4.

4. Implementation

We implemented the network using the publicly available Matconvnet toolbox [32] with an Nvidia GTX Titan X GPU. Stochastic gradient descent (SGD) is used to optimize the parameters. We set the momentum and weight decay to 0.9 and 0.0005, respectively. We used the initial learning rate of 10^{-6} and divided it by 10 when the validation loss plateaus. Parameters of all the deconvolution layers are fixed as the bilinear up-sampling kernels for training and inference. During training, random flipping is applied to augment the input image patches.

Training of the proposed model proceeds in three stages. First, we train the feed-forward baseline model for density estimation. Based on the base model, the segment decoder, the depth decoder and the count decoder are successively trained. In the third stage, the four decoders are jointly optimized and the model is trained end-to-end using the objective function of Eq. 5.

Once the model has been trained, the auxiliary tasks prediction module can be detached and the original model with more powerful capacity is used at inference.

5. Experiments

In this section, we evaluate the proposed crowd counting method on three benchmark datasets of the shanghaiTech-B [36], the worldExpo'2010 [35] and the Mall [3] dataset. Following the convention of existing work [35, 36], metrics of the mean absolute error (MAE) and the mean square error (MSE) are computed for evaluation.

5.1. Datasets

ShanghaiTech part_B It is the largest dataset for crowd counting in terms of the number of annotated people. It contains 716 images with a fixed size of 768×1024 taken from busy streets. Compared to the Mall dataset [3], it poses more challenging situations with severe perspective distortion and diverse scenes. Following the public splits, 400 images are for training and the remaining 316 are for testing. We crop image patches with a size of 224×224 for training.

WorldExpo'2010 It is a large-scale dataset including 3980 annotated video frames captured by 108 surveillance cameras from Shanghai 2010 worldExpo, with a fixed size of 576×720. Compared to the ShanghaiTech part_B [36], it covers a large variety of scenes. Following the public splits, 3380 frames from 103 scenes are treated as training and validation sets. The left 600 frames, with 120 from each of the 5 test scenes, are set for testing. It provides Region of Interest (ROI) for each scene, and hence only the pedestrians within the ROI are considered in evaluation following previous methods [35, 36]. Image patches in a size of 256×256 are cropped from the original image for training.

Mall It contains 2000 frames with a fixed size of 320×240 recorded from a surveillance camera in a shopping mall. We use the public splits for training and testing, i.e., the first 800 frames for training, and the rest 1200 frames for testing. $1/6$ of the training images are randomly selected as validation, which is the same for all the evaluation datasets. To augment training data, we crop image patches with a size of 160×160 from the original image.

Table 1. Different encoder-decoder architectures evaluated in the experiment.

Architecture	AT-CFCN	AT-CSRNet
Encoder	7×7×32 conv, stride 2 7×7×64 conv, stride 2 5×5×128 conv	(3×3×64 conv)×2, stride 2 (3×3×128 conv)×2, stride 2 (3×3×256 conv)×2, stride 2 (3×3×512 conv)×2, stride 2
Decoder (for density, depth and segment prediction)	5×5×64 conv 7×7×32 deconv, upsample 2 7×7×1 deconv, upsample 2	(3×3×512 conv, dilate 2)×3 3×3×256 conv, dilate 2 3×3×128 conv, dilate 2 3×3×64 conv, dilate 2 3×3×1 conv
Decoder (for count regression)	N× N× 64 conv 1× 1×32 conv 1× 1× 1 conv	N× N× 512 conv, dropout 0.5 1× 1× 256 conv 1× 1× 128 conv 1× 1× 64 conv 1× 1× 1 conv

5.2. Diagnostics Experiments

To deeply analyze the proposed approach and demonstrate its effectiveness, we conduct diagnostics experiments on two evaluation datasets: the ShanghaiTech-B [36] and the Mall [3]. For the backbone CNN, we experiment with two models with various capacity to adapt to various dataset sizes and also to study the performance gains grounded on different models. A lightweight counting FCN model (CFCN) with three convolution layers for both the encoder and decoder is chosen for the Mall dataset [3]. Another one is a much deeper model (CSRNet [12]) which adapts

VGG network [27] for crowd counting with dilation processing. Detailed architectures of the AT-CFCN and AT-CSRNet which integrate the auxiliary tasks prediction module are shown in Table. 1. The convolution kernel N in the decoder for count regression depends on the input image size and the downsample factors in the front-end encoder, which transforms the feature maps into 1×1 vectors for count estimation. In both two baseline models, each convolutional layer is followed by a rectified linear unit (RELU) and is accordingly padded to keep the spatial resolution.

From the base backbone model of CFCN/CSRNet, we compare several different variants, including those with only one auxiliary task (i) base CNN + DE: performing the depth prediction (DE) task with the front-end CNN; (ii) base CNN + SE: performing the crowd segmentation (SE) task with the front-end CNN; (iii) base CNN + CT: performing the count estimation (CT) task with the front-end CNN. The variants with two auxiliary task include (iv) base CNN + DE + SE: performing the depth prediction and crowd segmentation task at the same time; (v) base CNN + DE + CT and (vi) base CNN + SE + CT which are similar to (iv) with learning of two auxiliary tasks. Finally, we compare with the variant where all the three auxiliary tasks are integrated: (vii) of base CNN + DE + SE + CT.

Several conclusions could be drawn from Table 2. i). The three auxiliary tasks all take effects on decreasing the counting errors in terms of the MAE and MSE (compare $b\sim d$ vs a). This demonstrates that the auxiliary tasks carry the key information that influences the accuracy of the density estimation and jointly optimize the main task. ii). Including any two of the three auxiliary tasks will further decrease the counting errors (compare e vs b, e vs c, f vs b, *etc*.), and leveraging all of them achieves the best performance. This result is in alignment with our hypothesis that the auxiliary tasks each focus on heterogeneous attributes of the density map and their collaboration will further improve the representations for more accurate density estimation. iii). The proposed approach not only improves the simpler model (CFCN), and also significantly improves the deep model (CSRNet) which are naturally armed with stronger representation ability. This further validates the necessity and effectiveness of the proposed approach to explicitly leverage the heterogeneous attributes existing in the density map. Similar situations can be observed from Table 6 for the Mall dataset [3].

5.3. Comparison with State-of-the-art

The proposed method is compared with several state-of-the-art methods on three challenging benchmarks. The comparison results are shown in Table 4, 6 and 5. As demonstrated in Table 4 and 5, our method outperforms previous state-of-the-art methods on both the ShanghaiTech-B dataset [36] and the WorldExpo'2010 dataset [35]. The images in both of these two datasets are collected from outdoor scenes with significant perspective variations and complex background clutter, which easily incurs the geometric and the semantic inconsistency problems. The superior performance of the proposed method demonstrates the effectiveness to leverage the auxiliary attributes during the training process to help pursue the geometric and semantic consistency of the density estimation. Our method is also validated on the Mall dataset [3] for sparse crowds in indoor scenes. Due to the perspective distortion is not very obvious in the indoor scenes, the effectiveness of our approach against the scale variations is limited in this dataset. However in Table 6 we still achieve competitive results compared with prior art, showing our approach is not only effective to in dense scenarios but also generalizes well to the images with sparse pedestrians.

Table 2. Diagnostic experiments of AT-CFCN and AT-CSRNet on the ShanghaiTech-B dataset [36].

Item	Method	AT-CFCN		AT-CSRNet	
		MAE	MSE	MAE	MSE
a	base CNN	12.89	22.3	10.6	16.0
b	base CNN + DE	11.72	19.76	8.73	13.63
c	base CNN + SE	12.31	20.66	9.20	14.14
d	base CNN + CT	12.24	21.49	9.11	14.39
e	base CNN + DE + SE	11.52	19.78	8.28	13.97
f	base CNN + DE + CT	11.58	19.73	8.32	13.57
g	base CNN + SE + CT	11.88	20.42	8.51	13.66
h	base CNN + DE + SE + CT	**11.05**	**19.66**	**8.11**	**13.53**

Table 3. Diagnostic experiments of AT-CFCN on the Mall dataset [3].

Item	Method	MAE	MSE
a	base CNN	3.14	3.90
b	base CNN + DE	2.79	3.51
c	base CNN + SE	2.68	3.37
d	base CNN + CT	2.83	3.55
e	base CNN + DE + SE	2.36	3.02
f	base CNN + DE + CT	2.48	3.18
g	base CNN + SE + CT	2.34	2.99
h	base CNN + DE + SE + CT	**2.28**	**2.90**

Table 4. Comparison with other state-of-the-art crowd counting methods on the ShanghaiTech-B dataset [36].

Method	MAE	MSE
LBP + RR [23]	59.1	81.7
Crowd-CNN [35]	32.0	49.8
MCNN [36]	26.4	41.3
Cascade-CNN [28]	20.0	31.1
Switch-CNN [22]	21.6	33.4
CP-CNN [29]	20.1	30.1
DecideNet [14]	20.75	29.42
ACSCP [24]	17.2	27.4
IG-CNN [21]	13.6	21.1
CSRNet [12]	10.6	16.0
AT-CSRNet	**8.11**	**13.53**

To gain further understanding of the proposed approach, we conduct detailed comparison experiments with the recent state-of-the-art CSRNet [12] on ShanghaiTech part-B.

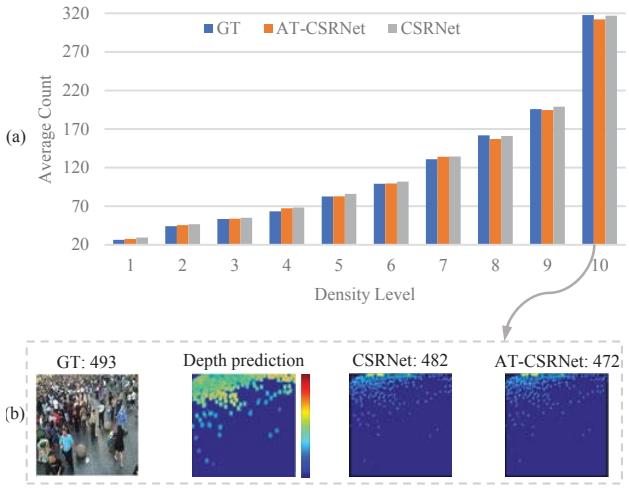

Figure 4. (a) Histogram: comparison of average count estimation on 10 splits of ShanghaiTech-B dataset according to the increasing number of people in each image. (b) Visualization of a failure case from the last split.

Table 5. Comparison with other state-of-the-art crowd counting methods on the WorldExpo'2010 dataset [35].

Method	S1	S2	S3	S4	S5	Average
LBP + RR [23]	13.6	59.8	37.1	21.8	23.4	31.0
Cascade-CNN [28]	4.8	32.5	10.8	13.3	4.5	13.2
Crowd-CNN [35]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [36]	3.4	20.6	12.9	13.0	8.1	11.6
Switch-CNN [22]	4.4	15.7	10.0	11.0	5.9	9.4
CP-CNN [29]	2.9	14.7	10.5	10.4	5.8	8.86
IG-CNN [21]	2.6	16.1	10.15	20.2	7.6	11.3
DecideNet [14]	2.0	13.14	8.9	17.40	4.75	9.23
CSRNet [12]	2.9	**11.5**	**8.6**	16.6	**3.4**	8.6
AT-CSRNet	**1.8**	13.7	9.2	**10.4**	3.7	**7.8**

Table 6. Comparison with other state-of-the-art crowd counting methods on the Mall dataset [3].

Method	MAE	MSE
SquareChn Detector [1]	20.55	439.1
R-FCN [4]	6.02	5.46
Faster R-CNN [19]	5.91	6.60
Ridge Regression [23]	3.59	19.0
MORR [3]	3.15	15.7
Count Forest [17]	4.40	2.40
Cascade-CNN [28]	3.02	3.81
Weighted VLAD [25]	2.41	9.12
Exemplary Density [34]	1.82	2.74
Boosting CNN [33]	2.01	N/A
MoCNN [10]	2.75	13.4
DecideNet [14]	**1.52**	**1.90**
AT-CFCN	2.28	2.90

Test images are divided into ten groups according to the increasing number of people in each image. It can be observed from Figure. 4 (a) that our method outperforms the CSRNet across most data splits, demonstrating the robustness and the effectiveness of the proposed approach. We further visualize a failure case from the last data split in Figure 4 (b). We keep the depth decoder at testing and save the depth predictions. As shown in the second column of Figure 4 (b), we found that the depth map for the sample image failed to properly depict the depth relationships especially for the farthest crowd in the left upper corner, which may lead to inaccuracy of the density estimation and hence the count result. This indicates the insufficient ability of the trained depth decoder. Considering the fact that ground truth depth maps currently used to train our model are generated by existing depth algorithms which have not been specifically adapted to crowd scenes, we guess with more accurate depth ground truth provided, the depth decoder could be better optimized and inversely benefit the base model for better results on such kind of examples.

Figure. 6 visualize and compares the predicted density maps and counts of our method (AT-CSRNet) and the CSRNet. Overall we achieve more accurate count estimations and reserve more consistency with the crowd distributions.

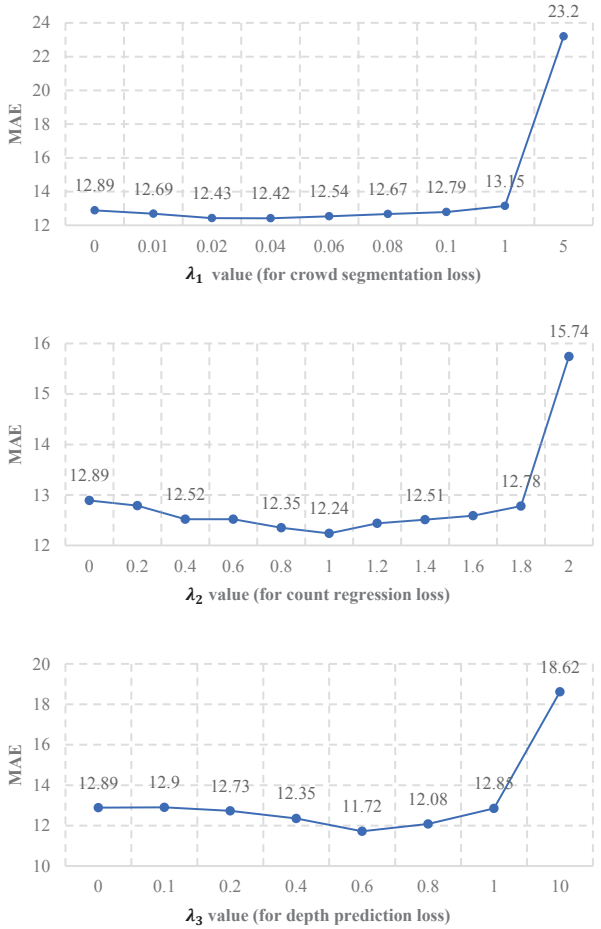

Figure 5. Comparison of MAE with different weight of the loss for the three auxiliary tasks on ShanghaiTech-B dataset [36].

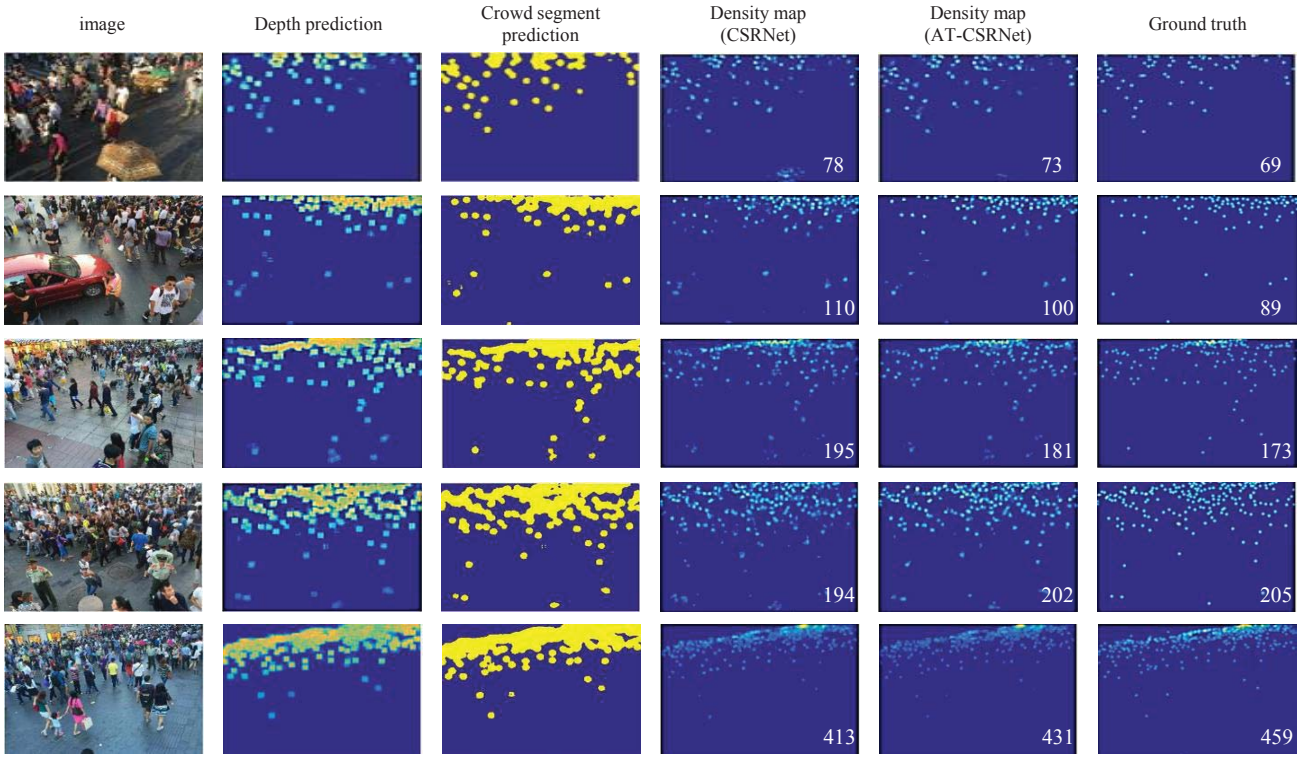

Figure 6. Visualization and comparison. First column: test image. The second and third columns show the predicted depth map and the crowd segment from the corresponding decoder, respectively. The last three columns are the estimated density map by CSRNet [12], by our method (AT-CSRNet) and the ground-truth, respectively. Count estimation are labeled at the right corner of each density prediction.

For instance, for the first image, the estimation of CSRNet shows inaccuracy in the umbrella area, however with the learning of auxiliary segmentation task which inversely help refine the intermediate features and avoid such falsely activated density estimations in our prediction. Similar situations can be observed for other sample images.

5.4. Parameter Study of the Weights for Auxiliary Tasks

The weights λ_i in Equation 5 determines the influence of each auxiliary task on the main task, which is a key parameters in our approach. To optimize the selection of λ_i, we conduct comparative experiments with the AT-CFCN model on the ShanghaiTech-B dataset. Figure. 5 shows the influences on density estimation when λ for each auxiliary task varies (parameters for other auxiliary tasks are set to be 0). As observed, for the depth prediction task, the MAE error decreases when the weights lie in a certain range of values. Too small weights are hard to contribute to the main tasks while too large weights will drift the feature representations and deteriorate the performances. Similar situations can be be observed for the crowd segmentation loss and the count regression loss. In our experiment, we select the weights for depth prediction loss, crowd segmentation loss and the count regression loss as 0.6, 0.04 and 1, respectively.

6. Conclusion

In this paper, we propose to leverage the heterogeneous attributes compounded in the density map to assist the crowd counting task. Specifically, we formulate the observed attributes as three auxiliary tasks to regularize the learning of the intermediate features for the main task of density estimation. Learning of the auxiliary tasks drives the embedding the information on geometry, semantics and the overall density level, which helps the feature to be more robust against the scale variations and clutter background. The proposed method does not incur any additional computations at inference, which gains efficiency over the general feature fusion scheme to augment the representations. Extensive experiments on multiple datasets shows our model achieves significant improvements or competitive results compared to recent state-of-the-art methods .

7. Acknowledgments

This work was partially funded by the National Key Research and Development Program No.2017YFB1002401, the National Science Fund of China under Grant No.61571297 and No.61420106008, 111 Program No.B07022 and STCSM No.18DZ2270700 and No.18DZ1112300.

References

[1] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. Ten years of pedestrian detection, what have we learned? In *European Conference on Computer Vision*, pages 613–627. Springer, 2014.

[2] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

[3] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *BMVC*, volume 1, page 3, 2012.

[4] KHJS Jifeng Dai and Yi Li R-fcn. Object detection via region-based fully convolutional networks. NIPS, 2016.

[5] Kai Kang and Xiaogang Wang. Fully convolutional neural networks for crowd segmentation. *arXiv preprint arXiv:1411.4464*, 2014.

[6] Dan Kong, Douglas Gray, and Hai Tao. Counting pedestrians in crowds using viewpoint invariant training. In *BMVC*, pages 1–6, 2005.

[7] Dan Kong, Douglas Gray, and Hai Tao. A viewpoint invariant approach for crowd counting. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 1187–1190. IEEE, 2006.

[8] Shu Kong and Charless Fowlkes. Recurrent scene parsing with perspective understanding in the loop. *arXiv preprint arXiv:1705.07238*, 2017.

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[10] Shohei Kumagai, Kazuhiro Hotta, and Takio Kurita. Mixture of counting cnns. *Machine Vision and Applications*, Jul 2018.

[11] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *Advances in Neural Information Processing Systems*, pages 1324–1332, 2010.

[12] Yuhong Li, Xiaofan Zhang, and Deming 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.

[13] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2016.

[14] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5197–5206, 2018.

[15] Chen Change Loy, Ke Chen, Shaogang Gong, and Tao Xiang. Crowd counting and profiling: Methodology and evaluation. In *Modeling, Simulation and Visual Analysis of Crowds*, pages 347–382. Springer, 2013.

[16] Daniel Onoro-Rubio and Roberto J López-Sastre. Towards perspective-free object counting with deep learning. In *European Conference on Computer Vision*, pages 615–629. Springer, 2016.

[17] Viet-Quoc Pham, Tatsuo Kozakaya, Osamu Yamaguchi, and Ryuzo 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.

[18] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[20] Deepak Babu Sam and R Venkatesh Babu. Top-down feedback for crowd counting convolutional neural network. *arXiv preprint arXiv:1807.08881*, 2018.

[21] Deepak Babu Sam, Neeraj N Sajjan, R Venkatesh Babu, and Mukundhan Srinivasan. Divide and grow: Capturing huge diversity in crowd images with incrementally growing cnn. *arXiv preprint arXiv:1807.09993*, 2018.

[22] Deepak Babu Sam, Shiv Surya, and R. Venkatesh Babu. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4031–4039, 2017.

[23] Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. 1998.

[24] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5245–5254, 2018.

[25] Biyun Sheng, Chunhua Shen, Guosheng Lin, Jun Li, Wankou Yang, and Changyin Sun. Crowd counting via weighted vlad on dense attribute feature maps. *IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[26] Oliver Sidla, Yuriy Lypetskyy, Norbert Brandle, and Stefan Seer. Pedestrian detection and tracking for counting applications in crowded situations. In *2006 IEEE International Conference on Video and Signal Based Surveillance*, pages 70–70. IEEE, 2006.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Vishwanath A Sindagi and Vishal 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 14th IEEE International Conference on*, pages 1–6. IEEE, 2017.

[29] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *IEEE International Conference on Computer Vision*, 2017.

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

[31] Venkatesh Bala Subburaman, Adrien Descamps, and Cyril Carincotte. Counting people in the crowd using a generic head detector. In *Advanced Video and Signal-Based Surveillance (AVSS), 2012 IEEE Ninth International Conference on*, pages 470–475. IEEE, 2012.

[32] Andrea Vedaldi and Karel Lenc. Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 689–692. ACM, 2015.

[33] Elad Walach and Lior Wolf. Learning to count with cnn boosting. In *European Conference on Computer Vision*, pages 660–676. Springer, 2016.

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

[35] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang 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.

[36] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi 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.

[37] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, Cham, 2014.

[38] M. Zhao, J. Zhang, F. Porikli, C. Zhang, and W. Zhang. Learning a perspective-embedded deconvolution network for crowd counting. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 403–408, July 2017.