IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 26, 2020, accepted May 18, 2020, date of publication May 29, 2020, date of current version June 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2998678

Cascaded Multi-Task Learning of Head Segmentation and Density Regression for RGBD Crowd Counting

DESEN ZHOU AND QIAN HE

School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China
Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China
University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Desen Zhou (zhouds@shanghaitech.edu.cn)

This work was supported in part by the Shanghai NSF under Grant 18ZR1425100, and in part by NSFC under Grant 61703195.

ABSTRACT In this paper we propose a novel regression based RGBD crowd counting method. Compared with previous RGBD crowd counting methods which mainly exploit depth cue to facilitate person/head detection, our approach adopts density map regression and is more robust to severe occlusion under dense crowded scenarios. We develop a cascaded depth-aware counting network that jointly performs head segmentation and density map regression. Our network explicitly feeds depth map at each stage so that depth cues are sufficiently exploited. The multi-task strategy allows the network to explicitly attent to foreground regions of a crowd scene and improve density regression. To generate the ground truth of head segmentation and density map, we propose a head scale estimation method according to the basic geometric assumption and camera projection function. Experiments on two public RGBD crowd counting benchmarks, ShanghaiTechRGBD dataset and MICC dataset show that the proposed method achieves new state-of-the-art on both datasets. Further, our method can be easily extended to RGB datasets and achieves comparable performances on WorldExpo'10 dataset and UCF-QNRF dataset.

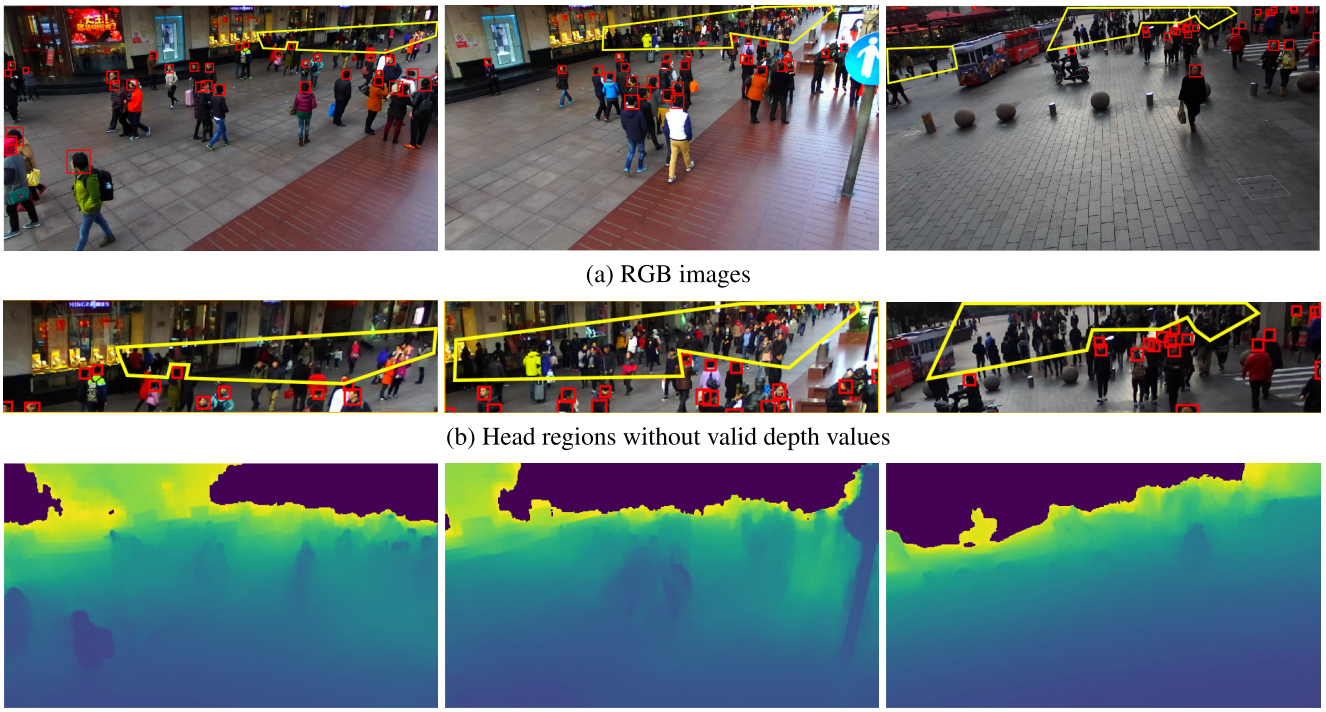
INDEX TERMS Crowd counting, depth map, density estimation, head segmentation.

I. INTRODUCTION

Single image crowd counting aims to estimate the overall person number in a crowded image. It has attracted significant attention in computer vision community during the past years [3]–[6]. Accurately estimating crowd counts of a scene has many applications in real-world scenarios [7]–[9]. For example, the statistics of passenger flow in subway stations are important for scheduling subway trains. The crowd count in a busy street plays an important role in public safety and pedestrian management.

Conventional crowd counting methods [6], [10]–[12] usually estimate crowd counts from RGB images or videos. The RGB crowd counting methods can be categorized into detection based methods [13]–[16] and regression based methods [3], [4], [17]–[19]. The first one treats each human instance in the crowd as an individual object and exploits object detection framework to tackle this problem; while the second one usually extracts low-level features of the scene and applies regressors to regress overall crowd counts or density maps. Although recent progress [11], [20], [21] shows significant improvement in RGB crowd counting, the problem itself is still challenging due to severe occlusion, perspective distortion and complex scene backgrounds. During the past years, depth sensors are becoming increasingly popular. Many people propose to exploit depth information to improve crowd counting [2], [22]–[25]. Most existing RGBD crowd counting methods utilize depth information to facilitate detection. However, detection based crowd counting methods are less robust to dense crowded scenarios with severe occlusion, and usually lead to underestimation when the people's heads are tiny/small [2]. To tackle this problem, Lian *et al.* [2] proposed a density map regression guided detection method. They first utilize a regressor to estimate a density map, which is used as a probability prior to facilitate detection. Although their results show that density map indeed improves detection, the proposed method suffers from several drawbacks. First, the depth cues are not explicitly fed into the regression network, which restricts the performance of density regression, and further affects the detection performance. Second, despite

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenbao Liu.



(a) RGB images

(b) Head regions without valid depth values

(c) Depth maps

FIGURE 1. RGB images (a) and their corresponding depth maps(c) in ShanghaiTechRGBD dataset. Depth maps are not perfect and contain invalid regions. We generate a bounding box for a head using depth map if a head has valid depth value, as suggested by [1], [2]. As shown in (a), many heads are not assigned with valid depth values(yellow boxes). These heads usually locate at the far regions of an image and hence they are small and dense, so the overall number is large. We enlarge those head regions with null depth values in (b).

the guidance of density map, detecting crowd instances is still challenging. The results of [2] show that crowd counting performance of its detection network is worse than its simple regression network.

In this work, we propose a simple but effective method for regression based RGBD crowd counting. Inspired by the pose estimation methods [26], [27], we propose a cascaded network for RGBD crowd counting. To sufficiently exploit depth information, we explicitly feed the depth map into the network multiple times. In addition to density map regression at each stage, we also predict a segmentation mask of crowd heads. The segmentation mask indicates foreground regions that the density regressor should attent to. Finally we utilize the depth map to generate the ground truth of density and segmentation. In this way, depth information is used in the input of the network and also the ground truth generation, and hence is sufficiently exploited.

Specially, we develop a cascaded depth-aware counting network (*Cascaded-DCNet*) to jointly estimate the segmentation mask and density map. Our network consists of two stages. The first stage takes image and depth map to estimate an initial segmentation and density map, and the second stage fuses the features, depth map and initial predictions to conduct refinement process. By explicitly predicting foreground mask, our network is able to focus on head regions and estimate density better.

To conduct multi-task learning on head segmentation and density map regression, we need to generate ground truth. An ideal ground truth of head segmentation may need the accurate label of each pixel indicating whether it belongs to a head. However, labeling each pixel for head segmentation is labor-intensive and not practical since many small/tiny heads only consist of few pixels. As the original annotation of the crowd counting tasks consists of pixel position of each head's center, we may want to estimate the scale of each annotated head, and then put a circle-like mask at the head center to generate head segmentation. In a pinhole camera system, each person's head radius is roughly inverse proportioned to its depth [1], [2]. This means if we can get the depth of each annotated head, then we are able to estimate the corresponding scale. However, we notice that this is usually not practical in real world scenarios since depth maps are not perfect. As shown in Fig. 1, many pixels around heads have invalid depth values. In ShanghaiTechRGBD dataset, the statistics show that there are 38.9% of the annotated heads have no valid depth values. To address this issue, we propose a head depth refinement method that takes the geometric assumption that all the heads of a crowd are on a 3D plane and leverages camera projection function to refine the depth values at those annotated head pixels, which are further used to estimate head scales.

We utilize the estimated head scales to generate a segmentation mask of each person' head. Then the segmentation masks are fused to generate a union segmentation of heads, which indicates the foreground regions of a crowd image. We also utilize the estimated head scale to generate

a scale-aware density map, in which each head's location is convolved with a scale-aware gaussian kernel. The density map encodes perspective information which captures scale variance of heads of the crowd scene.

We evaluate our approach on two public RGBD crowd counting benchmarks, ShanghaiTechRGBD dataset [2] and MICC dataset [22]. The results show that our method achieves new state-of-the-art on both datasets and validate the effectiveness of our proposed method. We further extend our method to RGB datasets. Results on WorldExpo'10 dataset [28] and UCF-QNRF dataset [29] show our method achieves comparable performances. We summarize our contributions as follows:

- We propose a new cascaded depth-aware counting network for regression based RGBD crowd counting. The depth map is explicitly fed into the network multiple times to extract depth information sufficiently.
- We propose a multi-task learning strategy for head segmentation and density map regression. Our network first estimates a head segmentation and then regresses density map based on the estimated segmentation. In this way our network is able to focus on foreground regions and estimate density better.
- We propose a novel ground truth generation method for head segmentation and density map. We first refine the depth values of the annotated heads according to camera projection function and basic geometric assumption, and then utilize the refined head depth map to estimate head scales, which are further used to generate head segmentation and scale-aware density map.
- Our method achieves new state-of-the-art on two public RGBD crowd counting benchmarks.

II. RELATED WORK

A. RGB CROWD COUNTING

Existing RGB crowd counting methods are mainly divided into detection based crowd counting and regression based crowd counting.

1) DETECTION BASED CROWD COUNTING

Detection based methods assume that a crowd is composed of some individual objects and treat crowd counting as an object/person detection problem. Early works [13]–[16] design hand-crafted features to perform person detection, but they are not robust to the severe occlusion or large scale variation on clustered environments or dense crowded scenes. Although recent deep network based object detectors [30], [31] show impressed performance on object detection, they still perform worse then regression based method on crowd counting [2].

2) REGRESSION BASED CROWD COUNTING

Starting from pre-deep learning era, regression based methods [17]–[19], [32]–[34], usually first 6segment foreground regions and extract various low-level features, and utilize a regression model, such as ridge regression [18], Gaussian process regression(GPR) [17] to estimate crowd count. In deep learning era, people formulate the crowd counting problem as a density map regression problem [11], [21], [35]–[39]. Zhang *et al.* [28] proposed to utilize a patch based crowd counting method by CNN. Zhang *et al.* [4] first proposed a multi-column CNNs, in which different column CNNs tackle heads with different sizes. Sam *et al.* [3] improved MCNN and propose a switchable module to classify the crowd density of each patch and assign it to corresponding regressor. Sindagi and Patel [12] proposed a top-down and bottom-up multi-level fusion mechanism to fuse features for crowd counting. CSRNet [21] stacks dilated convolutions after VGGNet [40]. Yan *et al.* [41] proposed a novel convolution operator that based on estimated perspective map. Our method follows density map regression methods. Previous density regression based works [12], [21], [39], [41] usually first extract image/patch features using a backbone network(e.g. VGG16 [40]), and then perform density regression. Our model has similar structure. However, the input of our network has two sources: RGB image and depth map. The cascaded architecture and multi-task strategy also make our method different from most regression based methods [12], [21], [36], [41].

B. RGBD CROWD COUNTING

To better estimate crowd counts, several works have explored the RGBD crowd counting. Most of these works focus on exploiting depth information to improve person/head detection of a crowd scene. Bondi *et al.* [22] utilized the depth information to estimate a crowd segment and further localize head candidates. However, the system is not end to end. Song *et al.* [42] proposed a detection proposal network for depth image based on Faster RCNN [30]. Zhang *et al.* [23] proposed an unsupervised method to estimate locations of heads based on depth image with vertical view. However, the method assumes the head regions are always closest to the camera compared with other body parts, and hence cannot be generalized to general crowd scenarios. Fu *et al.* [24] proposed to detect head-shoulder jointly based on template matching to improve robustness. However, in a dense crowd scenario, a person's shoulder is usually occluded. Xu *et al.* [25] utilized depth map to segment the image into two regions: a far-view region and a near-view region. A density regression module is used to tackle far-view crowd counting and an object detection module is used to tackle near-view crowd counting. However, depth information is not explicitly used for each region's estimation. Lian *et al.* [2] proposed a density map guided detection network for joint crowd head detection and density map regression. However, the performance of their detection module does not surpass the regression module. Meanwhile, the depth map is not explicitly fed into the regression module and hence it is not sufficiently used. In contrast, our model leverages cascaded architecture and explicitly fuses depth map twice. We further adopt multi-task learning strategy on head segmentation and density regression, and both of them

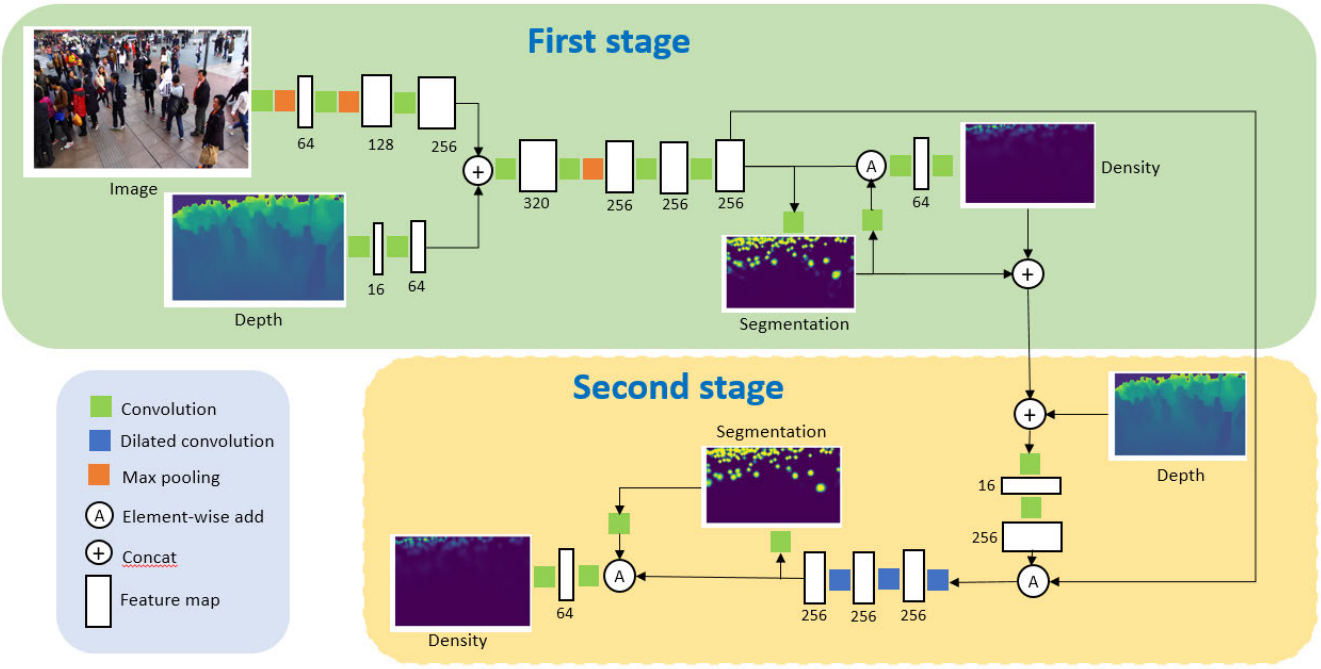


FIGURE 2. Overview of our proposed cascaded depth-aware counting network. The first stage takes the RGB image and depth map to generate an initial segmentation probability and density map, and the second stage combines the depth map, estimated predictions and their features to refine the estimations. We feed the depth map at each stage to sufficiently exploit depth cues. The multi-task learning on head segmentation and density map regression allows the network to attent to foreground regions of heads and improve density regression. Each convolution has a kernel size 3 × 3, a dilated convolution has dilation rate = 2. Each max pooling has a kernel size = 2 × 2 with stride = 2. Density maps and segmentation masks are 1/8 of original resolution due to pooling operations.

are supervised by depth guided ground-truths. By exploiting depth information sufficiently, our method is more robust to large scale variation and heavy occlusion under dense crowd scenarios.

III. OUR METHOD

A. FORMULATION

Given an image $I \in \mathbb{R}^{H\times W\times 3}$ with N heads annotated at $x = \{x_1, \ldots, x_N\}$, where $x_i \in \mathbb{R}^2$ denotes pixel location of i-*th* head. We denote the depth map as $D \in \mathbb{R}^{H\times W}$, and aim to design a network $\mathcal{F}$ that does the mapping $\{I, D\} \xrightarrow{\mathcal{F}} N$. As directly predicting N is highly non-linear, following prior regression based methods [3], [4], we first predict a density map $\mathbf{d} \in \mathbb{R}^{h\times w}$ indicating person densities at each pixel and then do the integration over the image/RoI, where h and w are downscaled height and width due to downsampling operations. In addition to density map, we also leverage our network to estimate a head segmentation which indicates foreground mask of a crowd image. The overall problem formulation becomes:

$$\{\mathbf{d}, \mathbf{s}\} = \mathcal{F}_\Theta(I, D), \tag{1}$$

where $\mathbf{s} \in [0, 1]^{h\times w}$ denotes estimated head segmentation probability, $\mathcal{F}$ is our network and Θ denotes the parameters. In the following subsections we will first describe our network architecture, and then introduce the ground truth generation of density map and segmentation mask. Finally we describe our loss function. In the following equations, the '+' denotes the element-wise addition operation, and '*' denotes the convolution operation, '·' denotes scalar product.

B. NETWORK ARCHITECTURE

Fig. 2 shows an overview of our proposed cascaded network which consists of two stages. The first stage takes original image I and its corresponding depth map D to generate initial segmentation probability and density map; the second stage combines the depth map, estimated predictions and their feature to refine the estimations. Both stages feed depth cue as their inputs and exploit depth information sufficiently through learned convolution filters. Below we will describe each stage in detail:

1) FIRST STAGE

We utilize three convolution layers and two max-pooling layers to extract image features from RGB image, and utilize another two convolution layers to extract depth features from depth map. The image features and depth features are fused by a concatenate operation. Such two stream strategy allows the network to extract image cue and depth cue independently in the shallow layers, hence it can avoid the confliction caused by domain gap between depth distribution and RGB distribution, and hence can be more efficient. Then we use several convolution operations and pooling operation to

generate initial predicting feature Γ_0, which is used to predict a segmentation probability $\mathbf{s}_0$:

$$\mathbf{s}_0 = g_0(\Gamma_0), \tag{2}$$

where g_0 indicates a convolution operation. The segmentation probability indicates the pixels of head regions, to which a density map regressor should attent. Hence it can be used as an attention to enhance features. We thus fuse $\mathbf{s}_0$ with Γ_0 to generate initial density map $\mathbf{d}_0$.

$$\mathbf{d}_0 = f_0(g_0'(\mathbf{s}_0) + \Gamma_0), \tag{3}$$

where g_0' is a convolution layer that embed $\mathbf{s}_0$ to a feature space with the same dimension as Γ_0. f_0 is a simple two convolutional neural network.

2) SECOND STAGE

The second stage performs refinement process based on the results of first stage. We first concatenate $\mathbf{s}_0$, $\mathbf{d}_0$ and $D_{1/8}$, where $D_{1/8}$ indicates depth map at $1/8$ of image resolution, and use two convolution layers to embed the output predictions, and then add the initial predicting features Γ_0 to generate refinement feature.

$$\Gamma_{ref} = f_0'(\mathbf{s}_0 \oplus \mathbf{d}_0 \oplus D_{1/8}) + \Gamma_0, \tag{4}$$

where f_0' indicates two convolution layers, $\oplus$ indicates concatenate operation, Γ_{ref} is the refinement feature. The refinement feature is fed though three dilated convolution layers to extract second predicting feature Γ_1. The dilation is used to enlarge receptive field. Similar to the first stage, Γ_1 is first used to generate a segmentation mask:

$$\mathbf{s}_1 = g_1(\Gamma_1), \tag{5}$$

where g_1 indicates a convolution operation. Then we fuse $\mathbf{s}_1$ and Γ_1 to generate second density map $\mathbf{d}_1$:

$$\mathbf{d}_1 = f_1(g_1'(\mathbf{s}_1) + \Gamma_1), \tag{6}$$

where g_1' is a convolution layer to embed $\mathbf{s}_1$ to a feature space, and f_1 is two layer convolutional neural network.

C. GROUND TRUTH GENERATION

To facilitate multi-task learning using our cascaded counting network, we need to generate ground truth for density map and head segmentation. As the annotation of the crowd counting task only consists of locations of each head's center, we need to first estimate the scale of each head. For segmentation mask generation, we label a pixel to foreground if its distance to a head annotation is smaller to that head's radius. For density map generation, we encode the head scale into the density map. Below we will first describe head scale estimation method, and then introduce the segmentation generation and density map generation.

1) HEAD SCALE ESTIMATION

As suggested in [1] and [2], for a fixed object with fixed physical size(e.g. head), its image size is usually inverse proportioned to the depth of the object due to theorem of similar triangles. The ratio is determined by the focal length of the camera, and we assume it is fixed across images of an existing RGBD dataset. Hence, if we get the depth of an annotated head, then we will get its scale. However, although depth map is provided, it does not always have valid/accurate values across all image pixels. For example, in ShanghaiTechRGBD dataset, the depth map is generated based on stereo matching, which is not very robust to simple textures such as heads/hairs. Further more, its depth map has a valid range of 0 to 20 meters, which does not cover the common crowd area in an image under outdoor scenes. This motivated us to find a way to estimate/refine the depth of heads without valid/accurate values.

Assume there is a set of heads located at $\mathbf{X} = \{X_1, \ldots, X_N\}$, where $X_i \in \mathbb{R}^3$ indicates physical 3D coordinate of i-*th* head under camera coordinate system. Since there is always enough people in a crowd scene, we can simply assume that each person has the same height and those heads lie on a plane. We denote the height of camera to the head plane as $H \in \mathbb{R}$ and the unit normal vector of the plane as $\mathbf{n} \in \mathbb{R}^3$, then we have:

$$\mathbf{n}^T X_i = H, \quad \forall i \in \{1, \ldots, N\}. \tag{7}$$

For a standard pinhole camera, we have the projection function:

$$D(x_i) \cdot x_i = K X_i, \tag{8}$$

where $D \in \mathbb{R}^{H \times W}$ represents the depth map of the entire image, $x_i \in \mathbb{R}^3$ is the projected pixel position on the image, denoted by homogeneous representation, $K \in \mathbb{R}^{3\times3}$ is the intrinsic parameter of the camera, and $D(x_i) \in \mathbb{R}$ is the normalization term indicating the depth of X_i.[1] From Eq. 8 we have $X_i = K^{-1}D(x_i)x_i$, and thus from Eq. 7 we have

$$\frac{1}{H}\mathbf{n}^T K^{-1} D(x_i) \cdot x_i = 1. \tag{9}$$

We denote $W = \frac{1}{H}\mathbf{n}^T K^{-1} \in \mathbb{R}^{1\times3}$ which is a fixed vector across the image representing the relation between x_i and $D(x_i)$. We can use those $\{x_i\}$ with valid depth values to estimate W and use the infered W to estimate those $\{x_i\}$ without valid depth available. For a head located at x_i with valid depth available, we denote $q_i = D(x_i) \cdot x_i \in \mathbb{R}^3$. Then for N' heads with valid depth available, we have

$$\begin{cases} Wq_1 = 1, \\ Wq_2 = 1, \\ \vdots \\ Wq_{N'} = 1. \end{cases} \tag{10}$$

We denote $Q = [q_1, \ldots, q_{N'}] \in \mathbb{R}^{3\times N'}$ and $E = [1, \ldots, 1] \in \mathbb{R}^{1\times N'}$, then we can find the best W that

[1] Here we ignore the effect of lens distortion.

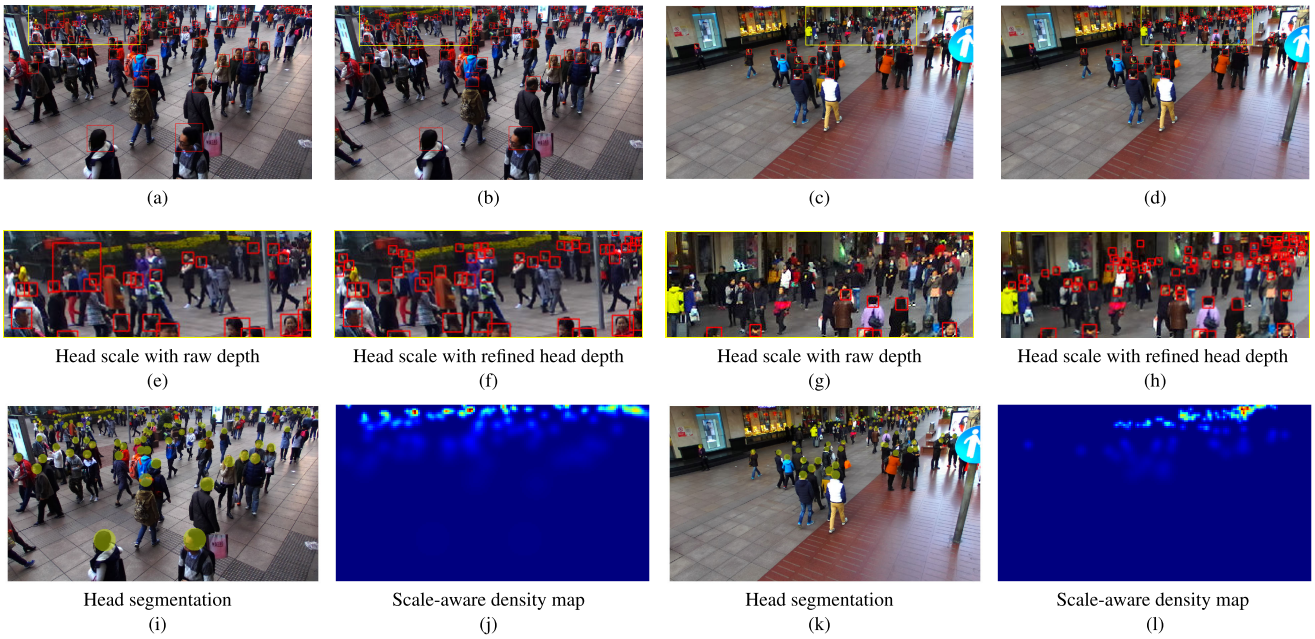


FIGURE 3. Two examples of our estimated head scales, head segmentations and density maps. First row denotes the original image and the estimated head scales using raw depth ((a), (c)) and refined depth ((b), (d)). We enlarge the yellow boxes in the first row and show them in (e), (f), (g), (h). We can observe that our depth refinement method refines those depth outliers and estimate depth of those heads without providing valid depth values. (i), (k) are the generated head segmentations, covered with original image for better visualization. (j), (l) are the scale-aware density maps.

approximates Eq. 10:

$$W' = \arg\min_{W} ||WQ - E||. \tag{11}$$

We have a closed form solution: $W' = (QQ^T)^{-1}Q^TE$. As N' is sufficiently large for a crowd, QQ^T is to be invertible. We solve W' in this way and for any pixel x in the image, we have $D(x) = 1/(W'x)$ denoting the depth value if x is a center of a head, and the head radius

$$r = \alpha/D(x) = \alpha W'x. \tag{12}$$

Fig. 3 shows two examples of our estimated head scales.

2) HEAD SEGMENTATION

As we have estimated head scale for each annotated head, we can utilize it to generate a head segmentation mask by masking a circle around the head centers. The segmentation mask is not perfect compared with human annotated segmentation but it provides the foreground regions that a network should focus on.

We use a uniform kernel $u_r(x)$ which indicates a kernel with all pixels equal to 1 inside a circle with radius r:

$$u_r(x) = \begin{cases} 1, & |x| \le r, \\ 0, & |x| > r. \end{cases} \tag{13}$$

Then, our head segmentation mask has the form:

$$\mathbf{m}(x) = \min(\sum_{i=1}^{N} \delta(x - x_i) * u_{r_i}(x), 1) \tag{14}$$

where $r_i = \alpha W'x_i$ indicates head radius for i-*th* head. $u_{r_i}(x)$ indicates scale-aware uniform kernel. Examples of our head segmentation mask are shown in Fig. 3.

3) DENSITY MAP WITH HEAD SCALE ENCODING

For an image with N head annotated at $x = \{x_1, \ldots, x_N\}$ where $x_i \in \mathbb{R}^2$, we may first convolve a gaussian kernel at each head to generate density map:

$$\mathbf{d}(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma(x). \tag{15}$$

Eq. 15 is the most commonly used density map generation for existing regression based methods. However, this density map generation method assumes that each person/head is individual on the image, and does not consider the scale variance of heads caused by perspective distortion. A better density map may consider the scale of heads in the gaussian kernel to encode the head scales in the density map, so that a network may easily capture the head regions from RGBD image and aligns to the density map without doing scale normalization. For a head at x_i, we have estimated its head scale $r_i = \alpha W'x_i$ by head scale estimation. Thus we may encode the head scale to density map, Eq. 15 changes to:

$$\mathbf{d}(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma_i}(x), \tag{16}$$

where $\sigma_i = \beta \cdot r_i = \beta \cdot \alpha W'x_i$.

D. LOSS FUNCTION

Consider a set of training samples $\{(I^k, D^k, \mathbf{d}^k, \mathbf{m}^k)\}$ where I^k is the RGB image, D^k is the depth map, $\mathbf{d}^k$ is our generated scale-aware density map, $\mathbf{m}^k$ is the segmentation mask, and the training set has M samples. For each sample, our network estimated two density maps $\mathbf{d}_0$, $\mathbf{d}_1$ and two segmentation probabilities $\mathbf{s}_0$, $\mathbf{s}_1$.

We utilize Euclidean loss for density map estimation:

$$L_d = \frac{1}{2M}\sum_{k=1}^{M}(||\mathbf{d}_0^k - \mathbf{d}^k||_2^2 + ||\mathbf{d}_1^k - \mathbf{d}^k||_2^2) \tag{17}$$

For segmentation, we use the average Binary Cross Entropy loss for each pixel, we indicate

$$L_{BCE}(a, b) = -\frac{1}{h \cdot w}\sum_{\forall p} b(p)\log(a(p)) + (1 - b(p))\log(1 - a(p)) \tag{18}$$

as the standard BCE loss for input a and target b with spatial resolution $h \times w$, and p indicates pixel location. Our segmentation loss is:

$$L_m = \frac{1}{2M}\sum_{k=1}^{M}(L_{BCE}(\mathbf{s}_0^k, \mathbf{m}^k) + L_{BCE}(\mathbf{s}_1^k, \mathbf{m}^k)) \tag{19}$$

The overall loss function is given by:

$$L = L_d + \mu L_m \tag{20}$$

where μ is the weight for segmentation loss to balance the gradient of segmentation and density map estimation.

IV. EXPERIMENTS

In this section, we perform experiments to evaluate our proposed method. We first describe the evaluation datasets and evaluation method. We then report the quantitative comparison on two RGBD benchmarks. We also perform ablation studies to validate the effectiveness of our proposed components or strategies. We then report results on RGB datasets. We finally show some qualitative results to demonstrate the efficacy of our framework.

A. DATASETS

1) RGBD DATASETS

a: ShanghaiTechRGBD DATASET

ShanghaiTechRGBD dataset [2] is a large-scale RGB-D dataset which consists of crowd scenes of metropolitan streets. The dataset consists of 1193 training images and 1000 test images. Each image has a fixed resolution-1920 × 1080. Each person in this dataset is annotated, and the overall crowd counts is 144,512. The person number of each image varies from 10 to over 200, and is 65.9 on average. Its depth map is generated using stereo matching algorithm, and has a range from 0 to 20 meters. The regions outside the range have no depth values. Readers are encouraged to refer [2] for more information of this dataset.

b: MICC DATASE

MICC dataset [22] is a dataset of indoor surveillance video frames. This dataset consists of three video sequences that represent for different crowd behaviors: in 'FLOW' sequence, people are walking from one point to another of the room, there are overall 1260 image frames and 3,542 crowd counts in this sequence; in 'QUEUE' sequence, people are acting as waiting in a line, and there are overall 918 frames and 5,031 crowd counts in this sequence; in 'GROUPS' sequence, people are talking in a controlled area, there are 1180 images and 9,057 crowd counts in this sequence. This dataset is a small dataset in terms of crowd counts compared with ShanghaiTechRGDB. The average person is 5.32 in each image. Following [2], we choose the 20% of each video sequence as training set and the remained are used as test set. The split is the same as [2].

2) RGB DATASETS

a: WorldExpo'10 DATASET

WorldExpo'10 dataset [28] is a standard dataset for crowd counting. It consists of 1,132 video sequences captured by 108 surveillance cameras with different viewpoints. 3,380 images from 103 scenes(viewpoints) are used for training and 600 images from 5 scenes(viewpoints) are used for testing. Each image has an average count of 56. The perspective maps and Region of Interest(RoI) masks are provided for each scene. During evaluation, only those crowd counts within RoI will be evaluated.

b: UCF-QNRF DATASET

UCF-QNRF
dataset [29] is a large dataset consists of 1,535 images in which 1201 images are used for training and 334 images are used for testing. It consists of 1.25 million persons annotated in total. The person counts, scales, backgrounds, viewpoints and image resolutions are varying significantly across different images, which cause this dataset very challenging.

B. EVALUATION METHOD

Following prior work of crowd counting [4], [12], we use Mean Absolute Error (MAE) and Mean Squared Error(MSE) for evaluation:

$$MAE = \frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}|\hat{N}_i - N_i| \tag{21}$$

$$MSE = \sqrt{\frac{1}{\mathcal{M}}\sum_{i=1}^{\mathcal{M}}|\hat{N}_i - N_i|^2} \tag{22}$$

where N_i represents the ground truth head counts, $\hat{N}_i = \sum_{\forall x}\mathbf{d}(x)$ is estimated head counts generated by integration on estimated density map $\mathbf{d}$, and $\mathcal{M}$ is the number of testing images.

C. IMPLEMENTATION DETAILS

We set gaussian parameter $\beta = 0.25$ and head radius parameter $\alpha = 5$, the loss weight $\mu = 5 \times 10^{-4}$. For shanghaiTechRGBD dataset, we first resize the images and depth maps to 1280×720 to decrease computation complexity. Depth maps are normalized to 0 to 255 before feed into the network. All the segmentation masks and density maps are generated at $1/8$ of original image resolutions. Each image is randomly flipped for data augmentation. During training process, we use Adam optimizer [43]. We set batch size to 4 and initial learning rate to 2×10^{-4}. We drop the learning rate to 2×10^{-5} and 2×10^{-6} at epoch 50 and 100, and stop training at epoch 150.

D. RESULTS ON ShanghaiTechRGBD DATASET

We first evaluate our method on ShanghaiTechRGBD dataset, and the results are shown in Tab. 1. Our final model achieves a MAE of 4.26 and a MSE of 6.27, which is a significant improvement compared with current state-of-the-art method [2]. RDNet [2] is a joint detection and regression network and its regression module utilizes a VGG backbone which requires pre-training on ImageNet. Instead, our proposed network does not need pre-training. We also report a result of CSRNet which only utilizes depth map as input instead of RGB image, the performance are not satisfactory. This is because the depth maps are very noisy and many head regions have no valid depth values. Hence, only utilizing depth maps as input is not applicable for those cluttered outdoor scenarios. Our method combines both depth information and RGB image. It is rather simple but effective on density map regression by utilizing depth information sufficiently.

TABLE 1. Results on ShanghaiTechRGBD dataset.

Model	MAE	MSE
MCNN [4]	7.56	10.92
MCNN-adaptive [2]	7.14	9.99
CSRNet [21]	5.11	7.34
CSRNet-depth [21]	13.76	20.49
CSRNet-adaptive [2]	4.91	7.11
RetinaNet [31]	10.25	14.56
DecideNet(DetNet) [10]	9.74	13.14
Idrees et al. [29]	7.32	10.48
RDNet [2]	4.96	7.22
Cascaded-DCNet (ours)	**4.26**	**6.27**

E. EXPERIMENTS ON MICC DATASET

On MICC dataset, the depth map is generated by Kinect. As the dataset contains indoor scenes, its depth range is much smaller than outdoor scenes (ShanghaiTechRGBD). However, we notice that the output depth map of kinect sensor still has invalid regions, as shown in Fig. 4. So the head depth refinement process is still required. As the head count of an image in MICC dataset is much less than ShanghaiTechRGBD, we only estimate the head plane parameters for those images with at least 5 heads annotated. For the rest images, we refine head depth by adopting max pooling at each head's local region. Our final result on MICC dataset is shown in Tab. 2. Our network achieves a MAE of 0.836 and a MSE of 1.031. Please note that the original MICC dataset contains the head box annotations, and we only use the center of each box as its dot annotation, while the current state-of-the-art method [2] utilizes box annotations for head detection.

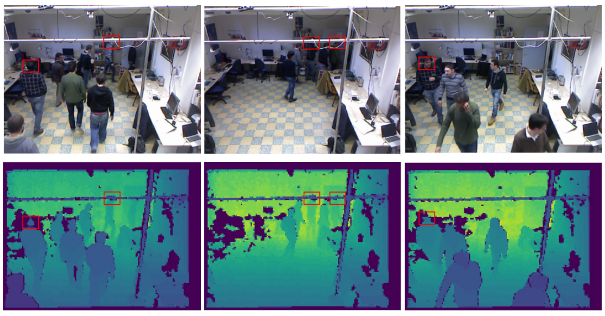


FIGURE 4. Depth maps are noisy in MICC dataset and some annotated head centers have no valid depth values (pointed out with red boxes).

TABLE 2. Results on MICC dataset.

Model	MAE	MSE
MCNN [4]	1.500	2.259
MCNN-adaptive [2]	1.489	2.114
CSRNet [21]	1.359	2.125
CSRNet-adaptive [2]	1.343	2.007
RetinaNet [31]	1.641	2.554
DecideNet(DetNet) [10]	1.541	2.382
Idrees et al. [29]	1.396	2.642
RDNet [2]	1.380	2.551
Cascaded-DCNet (ours)	**0.836**	**1.031**

F. ABLATION STUDY

We perform ablation study on our method to evaluate the effecviteness of our proposed architecture or strategies on ShanghaiTechRGBD dataset. We first perform ablation study on the cascaded depth-aware architecture to validate the performance of cascaded strategy. We then performance ablation study on our proposed scale-aware density map generation. We finally performance ablation study on the joint segmentaton and density map regression task.

1) CASCADED DEPTH-AWARE ARCHITECTURE

We perform ablation study for the cascaded architecture, and the results are shown in Tab. 3. Note that all the experiments are performed using fixed gaussian kernel for density map generation. Comparing the first row and second row, third row and fourth row, we can observe that using depth decreases the MAE by 0.33 and 0.19 respectively, demonstrating that our depth fusion mechanism is helpful for crowd counting. Comparing the first row and third row, second row and fourth row, we can notice that cascaded architecture improves the performance a lot.

TABLE 3. Ablation study on Depth-aware Cascaded Architecture, 'CNet' denotes basic counting network, 'DCNet' denotes depth-aware counting network, 'Cascaded' denotes the corresponding model is a cascaded architecture.

Model	MAE	MSE
CNet	6.52	9.08
DCNet	6.19	8.32
Cascaded-CNet	5.03	7.28
Cascaded-DCNet	**4.82**	**7.04**

2) DENSITY MAP GENERATION

We perform ablation study on different density map generation methods using the Cascaded-DCNet architecture, and the results are shown in Tab. 4. We first compare the density map with fixed gaussian kernel and depth-adaptive density map as proposed in [2] which utilizes raw depth map to estimate head scales(for pixels with invalid depth values, we pad the head scales using nearest neighbor, as in [2]). We can see that using depth-adaptive density map performs better than fixed kernel. We then utilize our proposed head depth refinement method to refine head depths and further estimate head scales, the proposed scale-aware density map generation further boosts the performance by a MAE of 0.13 and a MSE of 0.20.

TABLE 4. Ablation study on different density map generation, 'Depth-adaptive' denotes the adaptive-kernel using raw depth, 'Scale-aware' denotes the adaptive-kernel using head depth refinement.

Method	MAE	MSE
Fixed kernel	4.82	7.04
Depth-adaptive	4.66	6.97
Scale-aware	**4.53**	**6.77**

3) MULTI-TASK LEARNING ON SEGMENTATION AND DENSITY REGRESSION

We perform ablation study on multi-task learning of segmentation and density regression, and show the results in Tab. 5. We use the Cascaded-DCNet architecture, and the density map is generated by proposed scale-aware kernel. We can see that supervising on segmentation improves the performance by 0.27 MAE and 0.50 MSE, demonstrating that segmentation helps the network to better localize the foreground regions of heads.

TABLE 5. Ablation study on joint segmentation and density map regression.

Model	MAE	MSE
Single task	4.53	6.77
Multi-task	**4.26**	**6.27**

4) COMPARISON OF MODEL PARAMETERS

We also perform ablation study on model parameters to see how the model complexity affects final performance on ShanghaiTech RGBD dataset. The results are shown in Tab. 6. We compare the parameters in two directions: feature dimension and the number of layers. We increase the feature dimension from 256 to 512 and 1024, and observe that the performance improves, but the improvement becomes smaller. In the meantime, model parameters are increased significantly from 5.03M to 21.14M and further to 84.52M. This is because the model becomes overfitting as parameters increase. For the number of layers, we increase the layer by 1, 2, and 4 layers. We observe that the performance are not consistently becoming better as layer grows. We believe this is because the model becomes overfitting easily when it has more layers.

TABLE 6. Ablation study on model parameters on ShanghaiTech RGBD dataset. '−256', '−512', '−1024' denote the feature dimension of the backbone network. '-addLayer1', '-addLayer2', '-addLayer4' indicate the increased number of layers in the second stage.

Model	Params(M)	MAE	MSE
MCNN [4]	0.13	7.56	10.92
CSRNet [21]	16.26	5.11	7.34
Cascaded-DCNet-256	**5.03**	**4.26**	**6.27**
Cascaded-DCNet-512	21.14	4.15	6.33
Cascaded-DCNet-1024	84.52	4.11	6.12
Cascaded-DCNet-addLayer1	6.80	4.17	6.25
Cascaded-DCNet-addLayer2	7.98	4.28	6.33
Cascaded-DCNet-addLayer4	9.16	4.42	6.66

G. EXPERIMENTS ON RGB CROWD COUNTING DATASETS

In this paper, we propose a cascaded depth-aware network for joint head segmentation and density map regression. However, our proposed multi-task learning strategy and cascaded architecture can be easily extended to RGB crowd counting datasets by removing depth input and depth-aware ground truth generation. By estimating a segmentation mask, our network is able to attend to foreground regions and facilitate density regression. We evaluate our method on WorldExpo'10 dataset [28] and UCF-QNRF dataset [29].

1) RESULTS ON WorldExpo'10 DATASET

We compare the results of WorldExpo'10 dataset in Tab. 7. Following previous works [4], [28], we utilize the perspective

TABLE 7. Results on WorldExpo'10 dataset, measured by MAE.

Method	S1	S2	S3	S4	S5	Avg
Zhang et al. [28]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [4]	3.4	20.6	12.9	13.0	8.1	11.6
Switching-CNN [3]	4.4	15.7	10.0	11.0	5.9	9.4
CP-CNN [5]	2.9	14.7	10.5	10.4	5.8	8.9
PACNN [44]	2.6	15.4	10.6	10.0	4.8	8.7
DecideNet [10]	2.0	13.1	8.9	17.4	4.8	9.2
CSRNet [21]	2.9	11.5	8.6	16.6	3.4	8.6
SANet [45]	2.6	13.2	9.0	13.3	3.0	8.2
PGCNet [41]	2.5	12.7	8.4	13.7	3.2	8.1
ANF [46]	2.1	10.6	15.1	9.6	3.1	8.1
DRSAN [47]	2.6	11.8	10.3	10.4	3.7	7.8
ACSCP [48]	2.8	14.1	9.6	8.1	2.9	7.5
DSSINet [11]	1.57	9.51	9.46	10.35	2.49	6.67
Ours	**1.44**	**12.84**	**9.32**	**9.36**	**2.28**	**7.05**

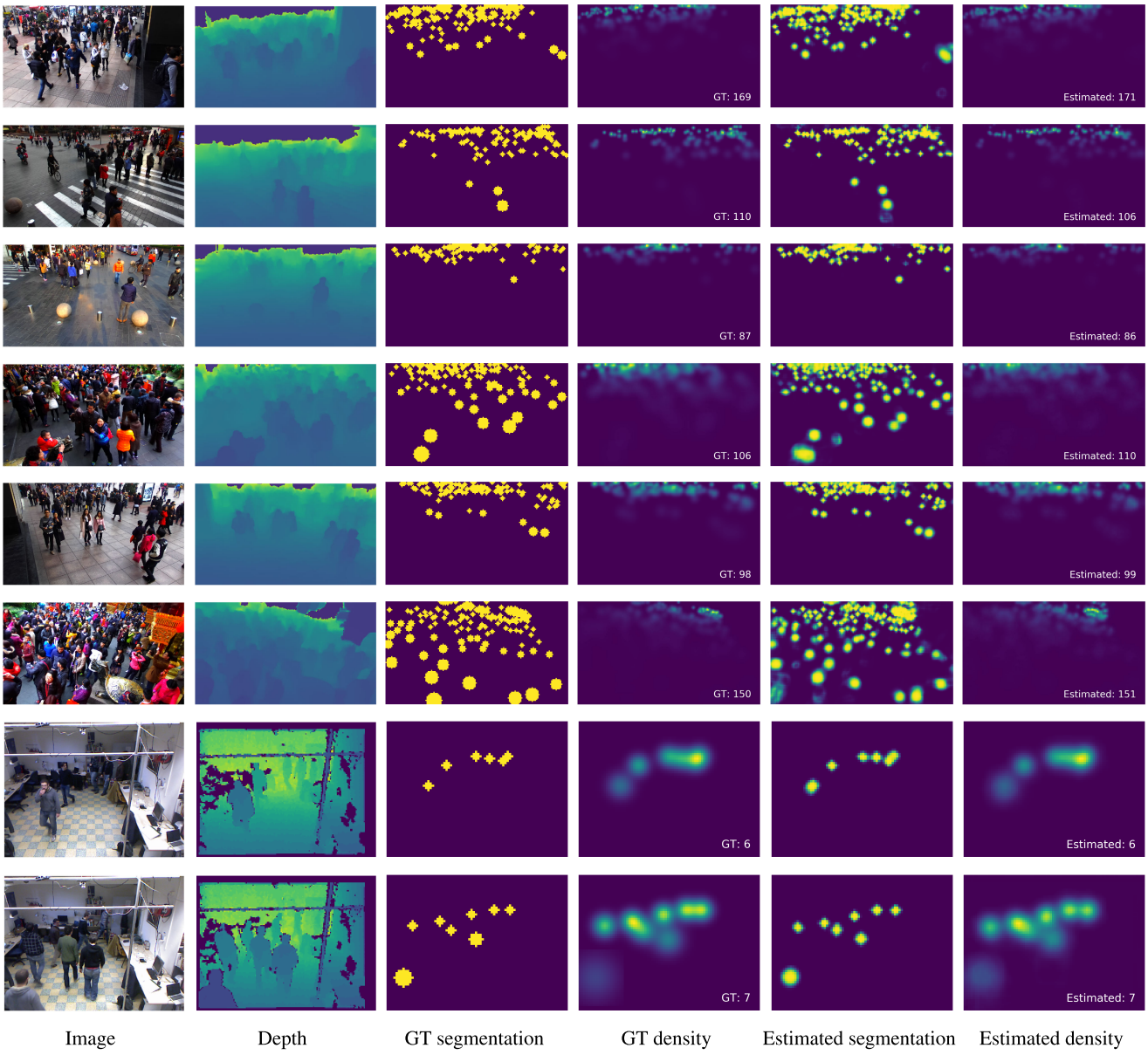


FIGURE 5. Qualitative results on ShanghaiTechRGBD dataset(first six rows) and MICC dataset(last two rows). We show all the inputs, ground truths and the predictions of our model's second stage. Our cascaded model exploits depth information sufficiently by multi-task learning on head segmentation and density regression, and hence it robust to severe occlusion and large scale variance.

maps provided by WorldExpo'10 dataset to generate ground truth density map and head segmentation mask. During testing, we only evaluate the crowd counts within given Region of Interest(RoI). We can see that our method outperforms other methods in two scenes and achieves comparable performance on average with current-state-of-the-art DSSINet [11], which utilizes multi-scale images as inputs and it's based on conditional random fields(CRF). Our method is based on single-scale image and its structure is simple.

2) RESULTS ON UCF-QNRF DATASET

We report the performance of UCF-QNRF dataset in Tab. 8. As the resolutions of images vary significantly, we randomly sample 224 × 224 patches to generate training data. Since UCF-QNRF dataset does not provide depth maps, we utilize k-nearest neighbor to estimate the head scales, which are further utilized to generate ground truth segmentation and density map. We notice that this dataset is much bigger than the ShanghaiTechRGBD dataset, MICC dataset and WorldExpo'10 dataset. Hence we replace the backbone of first stage(i.e. the feature extractor of Γ_0) to the first ten layers of VGG16 and utilize the pre-trained parameters to initialize our model, as many state-of-the-art methods ([11], [53]) utilize VGG16 to extract features. Our method achieves comparable performance with DSSINet, and outperforms other methods. It's worth noting that we only utilize coarse head scales

TABLE 8. Results on UCF-QNRF dataset.

Method	MAE	MSE
Idrees et al. [49]	315	508
MCNN [4]	277	426
Encoder-Decoder [50]	270	478
CMTL [51]	252	514
Switching-CNN [3]	228	445
Densenet-201 [52]	163	226
CL [29]	132	191
ANF [46]	110	174
CAN [53]	107.0	183.0
HA-CCN [54]	118.1	180.4
DSSINet [11]	99.1	159.2
Ours	**101.3**	**169.5**

estimated by k-nearest neighbor due to the lack of depth maps. We can expect the performance to be further improved if depth maps are used for network input and generating more precise head scales.

H. QUALITATIVE RESULTS

We show qualitative results of ShanghaiTechRGBD dataset and MICC dataset in Fig. 5. We can observe that segmentation predictions are quite reasonable. By multi-task learning on segmentation and density regression, our cascaded model is robust to heavy occlusion, large scale variance and variance of crowd counts.

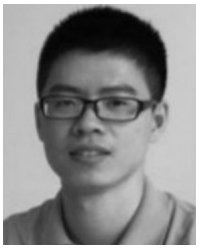
V. CONCLUSION

In this paper, we propose a novel cascaded depth-aware counting network for regression based RGBD crowd counting. The proposed network explicitly feeds depth map at each stage, exploiting depth cues sufficiently. We design a multi-task strategy that jointly estimates head segmentation and density map. Estimating head segmentation allows the network to focus on foreground regions of heads and improves density regression. To generate ground truth of head segmentation and density map, we first estimate the head scales. As in existing RGBD datasets, depth maps usually have invalid/inaccurate regions, we thus propose a head depth refinement approach to estimate/refine head depth at head locations. The refined head depth map is used to estimate head scales, and further generate segmentation mask and density map. Experiments show that our proposed cascaded network outperforms the single-stage network, and depth cue indeed helps density map regression. We also encode head scales to density map and result shows improvement. By conducting multi-task learning, the results show that predicting segmentation helps the network to attent to foreground regions and improve performance. Our method achieves new state-of-the-art on ShanghaiTechRGBD dataset and MICC dataset. We further extend our method to RGB datasets and it achieves comparable performances on WorldExpo'10 dataset and UCF-QNRF dataset.

REFERENCES

[1] E. Cheung, A. Wong, A. Bera, and D. Manocha, "Mixedpeds: Pedestrian detection in unannotated videos using synthetically generated human-agents for training," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 6738–6747.
[2] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for RGB-D crowd counting and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1821–1830.
[3] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4031–4039.
[4] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 589–597.
[5] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid CNNs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1861–1870.
[6] M. Marsden, K. McGuinness, S. Little, and N. E. O'Connor, "Fully convolutional crowd counting on highly congested scenes," 2016, *arXiv:1612.00220*. [Online]. Available: http://arxiv.org/abs/1612.00220
[7] D. Kang, Z. Ma, and A. B. Chan, "Beyond counting: Comparisons of density maps for crowd analysis tasks—Counting, detection, and tracking," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 5, pp. 1408–1422, May 2018.
[8] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 615–629.
[9] H. Lu, Z. Cao, Y. Xiao, B. Zhuang, and C. Shen, "TasselNet: Counting maize tassels in the wild via local counts regression network," *Plant Methods*, vol. 13, no. 1, p. 79, Dec. 2017.
[10] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "DecideNet: Counting varying density crowds through attention guided detection and density estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5197–5206.
[11] L. Liu, Z. Qiu, G. Li, S. Liu, W. Ouyang, and L. Lin, "Crowd counting with deep structured scale integration network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1774–1783.
[12] V. Sindagi and V. Patel, "Multi-level bottom-top and top-bottom feature fusion for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1002–1012.
[13] H. Idrees, K. Soomro, and M. Shah, "Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 1986–1998, Oct. 2015.
[14] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 7, pp. 1198–1211, Jul. 2008.
[15] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection," in *Proc. 19th Int. Conf. Pattern Recognit.*, Dec. 2008, pp. 1–4.
[16] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *Proc. CVPR*, Jun. 2009, pp. 2913–2920.
[17] A. B. Chan, Z.-S. J. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–7.
[18] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. Brit. Mach. Vis. Conf.*, 2012, vol. 1, no. 2, p. 3.
[19] A. B. Chan and N. Vasconcelos, "Bayesian Poisson regression for crowd counting," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 545–551.
[20] J. Wan and A. Chan, "Adaptive density map generation for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1130–1139.
[21] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.
[22] E. Bondi, L. Seidenari, A. D. Bagdanov, and A. Del Bimbo, "Real-time people counting from depth imagery of crowded environments," in *Proc. 11th IEEE Int. Conf. Adv. Video Signal Surveill. (AVSS)*, Aug. 2014, pp. 337–342.

[23] X. Zhang, J. Yan, S. Feng, Z. Lei, D. Yi, and S. Z. Li, "Water filling: Unsupervised people counting via vertical Kinect sensor," in *Proc. IEEE 9th Int. Conf. Adv. Video Signal-Based Surveill.*, Sep. 2012, pp. 215–220.
[24] H. Fu, H. Ma, and H. Xiao, "Real-time accurate crowd counting based on RGB-D information," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 2685–2688.
[25] M. Xu, Z. Ge, X. Jiang, G. Cui, P. Lv, B. Zhou, and C. Xu, "Depth information guided crowd counting for complex crowd scenes," *Pattern Recognit. Lett.*, vol. 125, pp. 563–569, Jul. 2019.
[26] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 483–499.
[27] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4724–4732.
[28] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 833–841.
[29] H. Idrees, M. Tayyab, K. Athrey, D. Zhang, S. Al-Maadeed, N. Rajpoot, and M. Shah, "Composition loss for counting, density map estimation and localization in dense crowds," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 532–546.
[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.
[31] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
[32] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Crowd counting using multiple local features," in *Digital Image Computing: Techniques and Applications*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 81–88.
[33] D. Kong, D. Gray, and H. Tao, "Counting pedestrians in crowds using viewpoint invariant training," in *Proc. Brit. Mach. Vis. Conf.*, 2005, p. 2.
[34] B. Liu and N. Vasconcelos, "Bayesian model adaptation for crowd counts," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4175–4183.
[35] Z.-Q. Cheng, J.-X. Li, Q. Dai, X. Wu, and A. Hauptmann, "Learning spatial awareness to improve crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6152–6161.
[36] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6142–6151.
[37] S. D. Khan, H. Ullah, M. Uzair, M. Ullah, R. Ullah, and F. A. Cheikh, "Disam: Density independent and scale aware model for crowd counting and localization," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 4474–4478.
[38] S. Basalamah, S. D. Khan, and H. Ullah, "Scale driven convolutional neural network model for people counting and localization in crowd scenes," *IEEE Access*, vol. 7, pp. 71576–71584, 2019.
[39] V. Sindagi, R. Yasarla, and V. Patel, "Pushing the frontiers of unconstrained crowd counting: New dataset and benchmark method," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1221–1231.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556
[41] Z. Yan, Y. Yuan, W. Zuo, X. Tan, Y. Wang, S. Wen, and E. Ding, "Perspective-guided convolution networks for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 952–961.
[42] D. Song, Y. Qiao, and A. Corbetta, "Depth driven people counting using deep region proposal network," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Jul. 2017, pp. 416–421.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[44] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Perspective-aware CNN for crowd counting," Inria Rennes—Bretagne Atlantique, Rennes, France, Tech. Rep. ffhal-01831109v1f, 2018, pp. 1–10.
[45] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.
[46] A. Zhang, L. Yue, J. Shen, F. Zhu, X. Zhen, X. Cao, and L. Shao, "Attentional neural fields for crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5714–5723.
[47] L. Liu, H. Wang, G. Li, W. Ouyang, and L. Lin, "Crowd counting using deep recurrent spatial-aware network," 2018, *arXiv:1807.00601*. [Online]. Available: http://arxiv.org/abs/1807.00601
[48] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5245–5254.
[49] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2547–2554.
[50] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[51] V. A. Sindagi and V. M. Patel, "CNN-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Proc. 14th IEEE Int. Conf. Adv. Video Signal Surveill. (AVSS)*, Aug. 2017, pp. 1–6.
[52] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[53] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5099–5108.
[54] V. A. Sindagi and V. M. Patel, "HA-CCN: Hierarchical attention-based crowd counting network," *IEEE Trans. Image Process.*, vol. 29, pp. 323–335, 2020.



DESEN ZHOU received the B.E. degree in information engineering from Shanghai Jiao Tong University, China, in 2014. He is currently pursuing the Ph.D. degree with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, and ShanghaiTech University, Shanghai, China. His research interests include human pose estimation, instance segmentation, human–object interaction detection, and 2D/3D scene understanding.

QIAN HE received the B.E. degree in information and communication engineering from Zhejiang University, China, in 2016. She is currently pursuing the Ph.D. degree with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, and ShanghaiTech University, Shanghai, China. Her research interests include object shape estimation and 3D scene understanding.