Body Structure Aware Deep Crowd Counting

Siyu Huang, Xi Li, Zhongfei Zhang, Fei Wu, Shenghua Gao, Rongrong Ji, and Junwei Han

***Abstract*— Crowd counting is a challenging task, mainly due to the severe occlusions among dense crowds. This paper aims to take a broader view to address crowd counting from the perspective of semantic modeling. In essence, crowd counting is a task of pedestrian semantic analysis involving three key factors: pedestrians, heads, and their context structure. The information of different body parts is an important cue to help us judge whether there exists a person at a certain position. Existing methods usually perform crowd counting from the perspective of directly modeling the visual properties of either the whole body or the heads only, without explicitly capturing the composite body-part semantic structure information that is crucial for crowd counting. In our approach, we first formulate the key factors of crowd counting as semantic scene models. Then, we convert the crowd counting problem into a multi-task learning problem, such that the semantic scene models are turned into different sub-tasks. Finally, the deep convolutional neural networks are used to learn the sub-tasks in a unified scheme. Our approach encodes the semantic nature of crowd counting and provides a novel solution in terms of pedestrian semantic analysis. In experiments, our approach outperforms the state-of-the-art methods on four benchmark crowd counting data sets. The semantic structure information is demonstrated to be an effective cue in scene of crowd counting.**

***Index Terms*— Crowd counting, pedestrian semantic analysis, visual context structure, convolutional neural networks.**

Manuscript received December 14, 2016; revised May 15, 2017; accepted August 6, 2017. Date of publication August 15, 2017; date of current version December 5, 2017. This work was supported in part by NSFC under Grant 61672456, Grant U1509206, and Grant 61472353, in part by the Fundamental Research Funds for Central Universities in China, in part by the Zhejiang Provincial Engineering Research Center on media data cloud processing and analysis technologies, in part by the ZJU Converging Media Computing Laboratory, in part by the Key Program of Zhejiang Province under Grant 2015C01027, in part by the National Basic Research Program of China under Grant 2015CB352302, and in part by the Alibaba-Zhejiang University Joint Institute of Frontier Technologies. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shuicheng Yan. *(Corresponding author: Xi Li.)*

S. Huang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China (e-mail: siyuhuang@zju.edu.cn).

X. Li is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China, and also with the Alibaba–Zhejiang University Joint Institute of Frontier Technologies, Hangzhou 310027, China (e-mail: xilizju@zju.edu.cn).

Z. Zhang is with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and also with the Computer Science Department, Watson School, The State University of New York Binghamton University, Binghamton, NY 13902 USA (e-mail: zhongfei@zju.edu.cn).

F. Wu is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: wufei@cs.zju.edu.cn).

S. Gao is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (e-mail: gaoshh@shanghaitech.edu.cn).

R. Ji is with the School of Information Science and Engineering, Xiamen University, Xiamen 361005, China (e-mail: jirongrong@gmail.com).

J. Han is with the School of Automation, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: junweihan2010@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2740160

I. Introduction

CROWD counting is a challenging task of accurately counting the crowds in dense scenes. It has drawn much attention from researchers because of a series of practical demands including crowd control and public safety. As illustrated in the top left of Fig. 1, there is a common crowd scene. The occlusions among people are severe and the perspective distortions vary significantly in different areas. In addition, the crowd distributions are visually diverse. These difficulties have restricted the performance of existing crowd counting methods.

In principle, crowd counting seeks for pedestrian semantic analysis involving three key factors: pedestrians, heads and their context structure. Most existing methods focus on modelling the visual properties of either the whole pedestrians or the heads only, while ignoring the context structure of different body parts which is also significant for counting the crowds. For instance, when we humans count the pedestrians, we will naturally use the composite body-part semantic structure information as an auxiliary cue to judge whether a head seen by us is exactly a pedestrian at that position or something else. It indicates that the semantic structures of pedestrians could provide abundant information for recognizing the pedestrians. However, many existing detection-based crowd counting methods [1], [2] model the pedestrians by constructing pedestrian detectors or head-shoulder detectors that they are limited by the severe occlusions in dense crowds. In more recent literatures [3], [4], researchers focus on modelling the density distributions of pedestrians, while, ignoring the body-part semantic structure information that is essential for the cognition of human beings.

Motivated by the above observations, we address the crowd counting problem from the viewpoint of semantic modelling in this work. The three key factors of crowd counting, including pedestrians, heads, and their context structure, are formulated as two types of semantic scene models. The first semantic scene model is denoted as the **body part map** in this paper. It models the visual appearance and context structure of pedestrian body parts. In body part map, different pedestrian body parts are formulated as different semantic categories, in the meantime, the spatial context structure of different parts are also formulated into the map. Fig. 1 provides an intuitive illustration of our approach. The body part map is created based on the single pedestrian parsing model [5], which is a pre-trained neural network model calculating the semantic segmentation mask of an input pedestrian image. And then we merge the segmentation masks of all the pedestrians to create the body part map. The top right of Fig. 1 shows the body part map highlighted by the areas of head, body, and legs respectively with colors of light blue, yellow, and red.

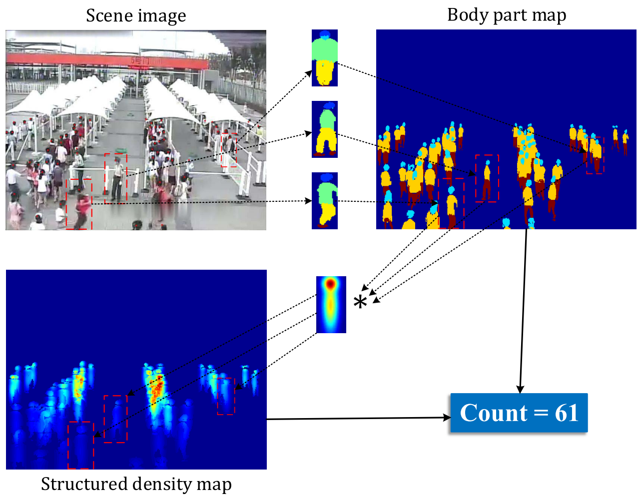


Fig. 1. Brief illustration of our approach. We build semantic scene models including the body part map and the structured density map to encode the semantic nature of a crowd scene. The crowd count is estimated based on the two semantic scene models.

The second semantic scene model is denoted as the **structured density map** in this paper. The conventional density maps in existing works [3], [6] are proposed to model the density distributions of crowds, while the shapes of individual pedestrians are ignored. Motivated by this, we create the structured density map according to specific shapes of individual pedestrians which are provided by the body part map. As an improvement of the conventional density map, the structured density map aims to model more fine-grained semantic structure information and so it can provide more accurate pixel-wise labels. As illustrated in bottom left of Fig. 1, the structured density map denotes the density information of crowds, meanwhile, preserving the shapes of specific pedestrians. In summary, the two semantic scene models, body part map and structured density map, are proposed to encode the semantic nature of crowd counting and they recover rich semantic structure information from crowd images.

For the purpose of accurately estimating the count of pedestrians, we reformulate crowd counting as a multi-task learning problem. There are three sub-tasks: inferring two types of semantic scene models and estimating the crowd count. We build deep convolutional neural networks (CNNs) to jointly learn these sub-tasks. The CNNs first model the mappings from scene image to semantic scene models including the body part map and structured density map, followed by calculating the crowd count based on them. The CNNs are able to extract powerful visual representations from images. The feature extraction and multi-task crowd counting problem are addressed in a unified scheme.

We summarize our main contributions as follows:

1) We provide a novel solution for crowd counting in terms of pedestrian semantic analysis. We formulate three key factors of crowd counting and model them as two types of semantic scene models. The models recover rich semantic structure information from images and are effective in learning our crowd counting framework.
2) We reformulate the crowd counting problem as a multi-task learning problem such that the semantic scene models are converted into its sub-tasks. We present a unified framework to jointly learn these sub-tasks based on the CNNs. Experiments show that our method achieves better results compared to the state-of-the-art methods.

II. Related Work

We introduce the literatures related to our work in this section. We first discuss the crowd counting methods proposed in existing literatures. And then we discuss several related works on pedestrian semantic analysis, as we address the crowd counting problem from the viewpoint of pedestrian semantic modelling in this paper. In addition, we introduce the background of CNNs, as our crowd counting framework is built upon the deep neural networks.

A. Crowd Counting

In general, most of the methods for crowd counting can be grouped into three categories: (1) detection-based, (2) global regression and (3) density estimation. The earlier literatures of crowd counting propose the *detection-based* methods [7]–[9] to model the semantic structure of pedestrians. Various kinds of detectors are employed to match individual pedestrians in images. Li *et al.* [10] use a HOG based head-shoulder detector to detect heads within foreground areas. Wu and Nevatia [1] detect local human body parts by part detectors and combine their responses to form people detections. Lin and Davis [2] learn a generic human detector by matching a part-template tree to images hierarchically. The detection-based methods perform better in relatively low dense scenes, while, they are limited by the heavy occlusions in dense crowds. Different from them, we model the semantic structure of pedestrians as semantic scene models. They are more robust to learn under the crowded scene and are more suitable for deep learning based framework.

To overcome the difficulties of detection-based methods in high dense scenes, researchers take a different way that they propose the *global regression* based methods to learn the mapping between low-level features and pedestrian counts. These methods are more suitable for crowded environments than the detection-based approaches. Diverse kinds of low-level features are employed, including textures [11]–[13], edge information [13], [14], and segment shape [12], [15], [16]. In addition, regression algorithms including linear regression [17], Bayesian regression [13], ridge regression [12], and Gaussian process regression [15], [18] are commonly used. The global regression based methods only utilize the information of pedestrian counts, while, the spatial information and body structure information of pedestrians are ignored.

To model the spatial information of pedestrians, researchers [6], [19]–[21] formulate the latent density distributions of crowds as an intermediate ground truth, namely, the *density estimation* based crowd counting. Lempitsky and Zisserman [6] first generate the density map based on the annotated points with a 2D Gaussian kernel and learn a linear regression function between scene image and density map. Following their work, other density estimation methods including random forest [22], [23] and deep neural networks [3], [4], [24], [25] are proposed. These methods demonstrate good performance on crowd counting. But from the perspective of semantic modelling, the body-part structures of individual pedestrians are ignored in these approaches. In this work, we focus on analyzing the semantic nature of crowd counting. We build semantic scene models by recovering rich semantic structure information from images and take them as novel supervised labels for crowd counting.

B. Pedestrian Semantic Analysis

The semantic analysis of pedestrian is an important prerequisite to many practical applications for intelligent surveillance systems operating in real world environments, including several typical computer vision topics like pedestrian detection [26]–[28], pedestrian parsing [5], [29] and crowd segmentation [30], [31]. In recent years, some high-level tasks of pedestrian analysis have also drawn much attention from researchers in recent years, including action recognition [32], [33], crowd attribute analysis [34], and pedestrian path prediction [35], [36]. What these approaches have in common is that they learn and model different aspects of semantic structure prior of pedestrians.

In this work, we address the crowd counting problem by focusing on pedestrian semantic analysis, because the visual cues of pedestrian body-part appearance can provide abundant information for recognizing the crowds. The success of parts-based methods [1], [2] on pedestrian detection also demonstrates this idea. Rather than directly detecting the holistic pedestrian, the parts-based methods utilize the information of pedestrian body structure and is able to handle occlusions more robustly. Different from the conventional parts-based methods, we formulate the body-part semantic structure of pedestrians as the semantic scene models in our approach, which are more suitable for learning under deep neural network framework and are more effective and robust in dense crowded scenes.

C. Convolutional Neural Networks

Our crowd counting framework is built upon the CNNs. The CNNs are a popular and leading visual representation technique, for they are able to learn powerful and interpretable visual representations [37]–[42]. Specifically, we use the fully convolutional networks (FCNs) to learn the semantic scene models proposed in our approach. As a kind of CNN architecture, FCNs are end-to-end models for pixelwise problems. They have given the state-of-the-art performance on many scene analysis tasks, including scene parsing [43]–[45], crowd segmentation [31] and action estimation [46]. For crowd counting, Zhang *et al.* [4] propose a multi-column FCN to map the crowd image to the density map. Their models are adaptive to the variations in pedestrian size and achieve the state-of-the-art performance on the benchmark datasets.

III. Our Approach

A. Problem Formulation

In this work, we aim to address the problem of single image crowd counting. Given a crowd image $\mathcal{X}$, our goal is to estimate the pedestrian number C in the image. It can be formulated as a mapping $\mathcal{X} \xrightarrow{F} C$. From the perspective of semantic modelling, we reformulate the original problem as a multi-task learning problem that contains three sub-tasks: the inference of two semantic scene models and the estimation of pedestrian number. The first semantic scene model is the body part map $\mathcal{B}$, which is built to model the body-part semantic structures of pedestrians. The second one is the structured density map $\mathcal{D}$, which is built to model the density distributions and shapes of pedestrians. Both two models are data-dependent that they respectively encode different semantic attributes of a crowd image. To address the multi-task learning problem, we build the CNNs to jointly learn the three sub-tasks in a unified framework. The learning process can be written as $\mathcal{X} \xrightarrow{F_1} (\mathcal{B}, \mathcal{D}) \xrightarrow{F_2} C$, such that $\mathcal{B}$ and $\mathcal{D}$ are used as auxiliary ground truths to better estimate C. For reading convenience, we summarize a collection of important notations used in this paper as Table I. We discuss our approach in more details in the following subsections.

TABLE I

The Detailed Description of the Variables

Variables	Description
$\mathcal{X} \in \mathbb{R}^{m \cdot n \cdot 3}$	the scene image
$\mathcal{M} \in \mathbb{R}^{m \cdot n}$	the perspective map
$\mathcal{B} \in \mathbb{R}^{m \cdot n \cdot 4}$	the body part map
$\mathcal{D} \in \mathbb{R}^{m \cdot n}$	the structured density map
$C \in \mathbb{R}$	the pedestrian count
$x \in \mathbb{R}^{m \cdot n \cdot 3}$	the image patch
$b \in \mathbb{R}^{m \cdot n \cdot 4}$	the body part map of image patch
$d \in \mathbb{R}^{m \cdot n}$	the density map of image patch
$c \in \mathbb{R}$	the pedestrian count of image patch
$p \in \mathbb{R}^2$	the coordinate of arbitrary location
$P \in \mathbb{R}^2$	the coordinate of specific location
$\mathcal{N}$	the 2D Gaussian kernel
F	the mapping function
L	the loss function of neural network

B. Body Part Map

As one of the semantic scene models, the body part map $\mathcal{B}$ is proposed to model the body-part semantic structures of individual pedestrians, which can serve as an important cue to judge whether there exists a person at a certain location. We introduce it into our framework as a novel supervised label to address the difficulties in crowd counting problem.

The body part map $\mathcal{B}$ is generated based on the given scene image $\mathcal{X}$, perspective map $\mathcal{M}$ and the locations of head points P_h^i. First, we have to obtain single pedestrian images. Due to the perspective distortions, we use the perspective map $\mathcal{M}$ (given in datasets) to normalize the scales of pedestrians. The pixel value $\mathcal{M}(p)$ denotes the number of pixels in the image representing one meter at location p in the actual scene. With the head location $P_h = (P_h^x, P_h^y)$ of a person, the top left corner P_{tl} and bottom right corner P_{br} of the person's bounding box are estimated as

$$\begin{aligned} P_{tl} &= \left(P_h^x - \alpha_1 \mathcal{M}(P_h), P_h^y - \beta_1 \mathcal{M}(P_h)\right), \\ P_{br} &= \left(P_h^x + \alpha_2 \mathcal{M}(P_h), P_h^y + \beta_2 \mathcal{M}(P_h)\right), \end{aligned} \quad (1)$$

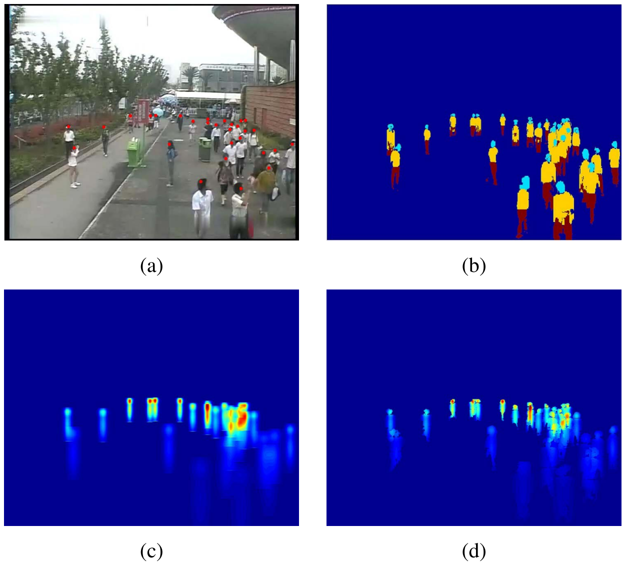


Fig. 2. Illustration of the semantic scene models. (a) is the scene image. The red points on (a) denote the annotations of pedestrian heads. (b) is the body part map. (c) is the conventional density map created by 2D Gaussian kernels only. (d) is the structured density map generated based on (b) and (c), modelling both the density information and shape information of individual pedestrians.

where the parameters are manually set as $\alpha_1 = 0.5$, $\alpha_2 = 0.5, \beta_1 = 0.25, \beta_2 = 1.75$ in the experiments to best approximate the actual situations, such that the width of bounding box is assumed as $\alpha_1+\alpha_2 = 1$ meter and the height of bounding box is assumed as $\beta_1 + \beta_2 = 2$ meters in the actual scene.

After obtaining the pedestrian images, we normalize them to the same size followed by inputting them into the single pedestrian parsing model [5] to calculate their semantic segmentation masks. The pedestrian parsing model uses a deep neural network to parse a single pedestrian image into several semantic regions, including hair, head, body, legs, and feet. The model is pre-trained that we only use it to generate the semantic segmentation of pedestrians. We merge the regions of hair into head, and also merge the regions of feet into legs. Finally, we resize the semantic masks of individual pedestrians to their original sizes to create the body part map $\mathcal{B}$. Fig. 2(a) shows a crowded scene image, and Fig. 2(b) shows the body part map corresponding to the scene image. Colors of light blue, yellow, red and dark blue respectively denote areas of head, body, legs and background. The body part maps containing labelled pixels of four categories models the semantic structure of every individual pedestrian in the scene images. And they are prepared for learning our crowd counting framework as discussed in subsection III-D.

C. Structured Density Map

The structured density map is proposed to capture both the density distributions and shapes of pedestrians. Different from the existing crowd counting methods [3], [4], [6], it is data-dependent in our approach that it is generated according to specific shapes of individual pedestrians.

We first discuss the conventional density map $\mathcal{D}_{\mathcal{N}}$ proposed in existing works [3] . It is usually created by a sum of 2D Gaussian kernels centered on the locations of pedestrians as:

$$\mathcal{D}_{\mathcal{N}}(p) = \sum_{i=1}^{C} \frac{1}{\|\mathrm{Z}\|} \Big(\mathcal{N}_h^i(p;\, P_h^i, \sigma_h^i) + \mathcal{N}_b^i(p;\, P_b^i, \sigma_b^i) \Big), \quad (2)$$

where $\mathcal{N}_h$ is a standard 2D Gaussian kernel for modelling the head part of a pedestrian and $\mathcal{N}_b$ is a bivariate 2D Gaussian kernel for modelling the body part of a pedestrian. p is the location of a pixel on $\mathcal{D}_{\mathcal{N}}$, P_h^i and P_b^i are respectively the i-th locations of person heads and bodies. In order to approximate the sizes of head and body in actual scene, we manually set the variance σ_h of kernel $\mathcal{N}_h$ as $\sigma_h = 0.25\mathcal{M}(P_h)$, and set the variance σ_b of kernel $\mathcal{N}_b$ as $\sigma_{bx} = 0.25\mathcal{M}(P_b)$ and $\sigma_{by} = \mathcal{M}(P_b)$. The body location P_b is set as $P_b = P_h + 0.8\mathcal{M}(P_h)$. $\|\mathrm{Z}\|$ is the normalization factor which normalizes the sum of density values for each person to 1, such that the sum of density values for all the persons is the count C. Fig. 2(c) shows the density map $\mathcal{D}_{\mathcal{N}}$ corresponding to the scene image in Fig. 2(a).

Since $\mathcal{D}_{\mathcal{N}}$ cannot well model the specific shapes of individual pedestrians, we further propose the structured density map $\mathcal{D}$:

$$\mathcal{D}_{\mathcal{N}}(p) = \sum_{i=1}^{C} \frac{1}{\|\mathrm{Z}\|} \Big(\mathcal{N}_h^i(p;\, P_h^i, \sigma_h^i) + \mathcal{N}_b^i(p;\, P_b^i, \sigma_b^i) \Big) \cdot \mathcal{B}_m(p) \quad (3)$$

The pedestrian mask $\mathcal{B}_m$ characterizes the shape of each pedestrian. It is obtained by binarizing the body part map $\mathcal{B}$, where the pixel values of foregrounds and backgrounds are respectively set to 1 and 0. The structured density map $\mathcal{D}$ is calculated by the element-by-element multiplication of $\mathcal{D}_{\mathcal{N}}$ and $\mathcal{B}_m$, followed by normalization. Fig. 2(d) shows the structured density map generated by our approach. Compared to the conventional density map in Fig. 2(c), we can see that the structured density map not only denotes the latent density distributions of crowds but also maintains specific shapes of every individual pedestrian.

D. Multi-Task Crowd Counting Framework

To estimate the accurate pedestrian number C, we reformulate the original crowd counting problem as a multi-task learning problem including three sub-tasks: the inference of semantic scene models $\mathcal{B}$ and $\mathcal{D}$, and the estimation of pedestrian number C. To jointly learn these three sub-tasks, we propose a unified multi-task learning framework based on the CNNs. See Fig. 3 for an illustration of our framework. We take a patch-wise strategy in which the input of networks is an image patch x cropped from scene image $\mathcal{X}$, where x is constrained to cover a 3-meter by 3-meter square in the actual scene according to the perspective map $\mathcal{M}$. In conventional neural network based density estimation method [3], there is only one stream of the mapping from scene patch x to density map d: $x \xrightarrow{F_d} d$, as illustrated in the blue blocks of Fig. 3. In our multi-task learning framework, we add another

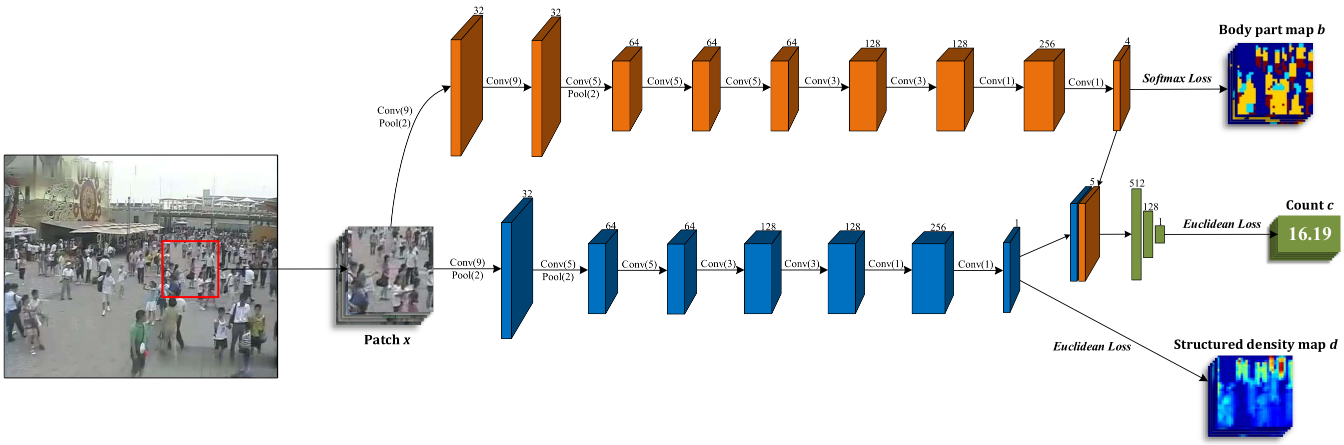


Fig. 3. Illustration of the proposed networks. The image patches are cropped from the scene image and are inputted into the CNNs. The convolutional layers denoted in blue blocks are built to infer the density distributions, and they are trained under the structured density map d with Euclidean loss. The layers denoted in orange blocks are built to infer the pedestrian body-part structures, and they are trained under the body part map b with Softmax loss. Finally, the crowd count c is regressed based on the two types of maps with fully connected layers as denoted in green blocks. **Best viewed in color.**

stream to learn the body part map b of patch x, as illustrated in the orange blocks of Fig. 3. The outputs of two streams are concatenated together and are mapped into the pedestrian count c by fully connected neural networks, as illustrated in the green blocks. As a whole, our crowd counting framework can be written as: $x \xrightarrow{F_1} (b, d) \xrightarrow{F_2} c$. Three supervised labels b, d, and c jointly train our model.

We discuss more details about the architectures of our networks. The networks between patch x and density map d are fully convolutional networks (FCNs) which contain 7 convolutional layers. The architecture is $conv_d1(9, 32) - pool_d1(2) - LRN - conv_d2(5, 64) - pool_d2(2) - LRN - conv_d3(5, 64) - conv_d4(3, 128) - conv_d5$(3,128)-$conv_d6$(1,256)-$conv_d7$(1,1), where '*conv*' represents a convolution layer, '*pool*' represents a max-pooling layer, and '*LRN*' represents a local response normalization layer. Numbers in the parentheses are respectively kernel size and number of channels. Each convolutional layer is followed by a ReLU activation function and is padded to keep the same as its last layer. The Euclidean loss is used to measure the difference between the estimated density map d and ground truth $\hat{d}$, as

$$L_d = \|d - \hat{d}\|_2^2. \tag{4}$$

The networks between patch x and body part map b are also FCNs which contain 9 convolutional layers. The architecture is $conv_b1(9, 32) - pool_b1(2) - LRN - conv_b2(9, 32) - conv_b3(5, 64) - pool_b3(2) - LRN - conv_b4(5, 64) - conv_b5(5, 64) - conv_b6(3, 128) - conv_b7(3, 128) - conv_b8(1, 256) - conv_b9$(1,1). We use a sum of Softmax loss at all 32×32 positions to measure the difference between the estimated body part map b and ground truth $\hat{b}$:

$$L_b = -\sum_{h=1}^{32}\sum_{w=1}^{32} \log \frac{\exp\left(\hat{b}(h, w)\right)}{\sum_{i=1}^{4} \exp(b(h, w, i))}, \tag{5}$$

where $\hat{b}(h, w)$ stands for the output of $conv_b9$ layer at spatial position (h, w) and channel of ground truth category. $b(h, w, i)$ is the output of $conv_b9$ layer at position (h, w) and i-th channel.

The two outputs d and b are concatenated to the fully connected layers $fc1(512)$-$fc2(128)$-$fc3(1)$ for estimating the counts c, where *fc* represents a fully connected layer. We use Euclidean loss to measure the difference between the estimated count c and ground truth $\hat{c}$, as

$$L_c = (c - \hat{c})^2. \tag{6}$$

The three loss functions L_d, L_b and L_c are combined as a joint multi-task loss L:

$$L = L_c + \lambda_d L_d + \lambda_b L_b. \tag{7}$$

λ_d and λ_b are loss weights respectively set to 10 and 1 in experiments. The whole network is jointly trained under L by back propagation and stochastic gradient descent.

Finally, the count C of an entire image is a sum of all the patch counts c of the image. Because C is composed of local count information within different areas, we further use the ground truth count $\hat{C}$ of image to fine-tune the count C predicted by our neural networks, as

$$\hat{C} = C\omega + \epsilon. \tag{8}$$

We employ the linear regression for fine-tuning. The regression coefficients ω and ϵ are estimated by the training data. In the testing phase, we use ω and ϵ to fine-tune the pedestrian count which is estimated by the neural networks.

IV. Experiments

A. Experimental Setup

We give the details on the networks, datasets, and the evaluation metric in the following.

1) Networks: We use the popular Caffe toolbox [47] to implement the proposed deep convolutional neural networks. Due to the effect of gradient vanishing for deep neural networks, it is not easy to learn all the parameters simultaneously. We use a trick in training phase that we first separately pretrain the two CNNs of mapping between the patch and two maps, and then use the pre-trained weights to initialize the entire CNNs and fine-tune all the parameters simultaneously.

TABLE II
SUMMARIZATION OF FOUR DATASETS

Dataset	Scenes	Frames	Resolution	FPS	Counts	Average	Total
UCSD	1	2000	158 × 238	10	11-46	24.9	49885
UCF_CC_50	50	50	different	image	94-4543	1279.5	63974
WorldExpo'10	108	4.44 million	576×720	50	1-253	50.2	199923
Shanghaitech-B	716	716	768×1024	image	9-578	123.6	88488

(a) (b) (c) (d)

Fig. 4. Example frames of (a) UCSD dataset, (b) UCF_CC_50 dataset, (c) WorldExpo'10 dataset, and (d) Shanghaitech-B dataset.

TABLE III
MEAN ABSOLUTE ERRORS (MAE) OF THE WORLDEXPO'10 DATASET

Method	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Average
LBP+RR	13.6	59.8	37.1	21.8	23.4	31.0
Crowd CNN [3]	10.0	15.4	15.3	25.6	4.1	14.1
Fine-tuned Crowd CNN [3]	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
Ours (conventional density map)	4.0	20.9	15.8	12.6	5.6	11.8
Ours	4.1	21.7	**11.9**	**11.0**	**3.5**	**10.5**

The network from patch to body part map is trained for 100K iterations with a batch size of 100 and learning rate of 10^{-5}. The network from patch to structured density map is trained for 100K iterations with a batch size of 100 and learning rate of 10^{-4}. Finally, the entire network is initialized with these pre-trained weights and is trained for 300K iterations with a batch size of 40 and learning rate of 10^{-5}. The input image patches are uniformly resized to 128×128. For a fair comparison with other crowd counting methods, we do not use pre-trained weights of other deep learning models.

2) Datasets: We evaluate our method in four benchmark datasets including the WorldExpo'10 dataset [3], the Shanghaitech-B dataset [4], the UCSD dataset [15] and the UCF_CC_50 dataset [48]. The details of the four datasets are summarized in Table II, where **Scenes** is the number of scenes; **Frames** is the number of frames; **Resolution** is the resolution of images; **FPS** is the number of frames per second; **Counts** is the minimum and maximum numbers of people in the ROI of a frame; **Average** is the average pedestrian count; **Total** is the total number of labeled pedestrians. Fig. 4 shows example frames of the four datasets. The scenes, crowd densities, crowd distributions, and perspective distortions vary significantly among these datasets such that they can be used to comprehensively evaluate the crowd counting methods.

3) Evaluation Metric: By following the convention of existing works [3], [4] for crowd counting, we use the mean absolute error (MAE) and mean squared error (MSE) to evaluate the performance of different crowd counting methods:

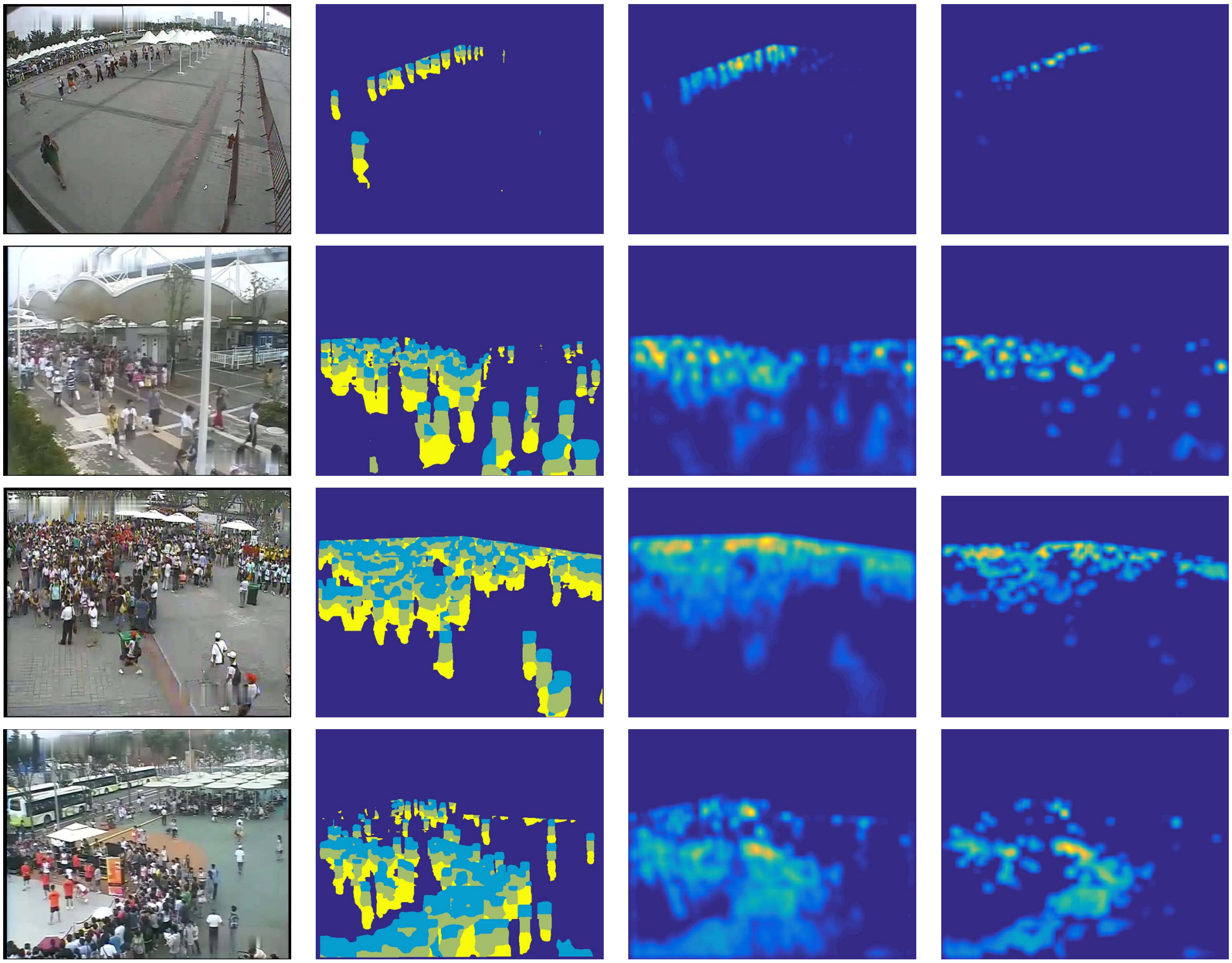
$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|C_i - \hat{C}_i|, \tag{9}$$

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_i - \hat{C}_i)^2}, \tag{10}$$

where N is the number of test images, C_i and $\hat{C}_i$ are respectively the estimated people count and ground truth people count in the i-th image.

B. WorldExpo'10 Dataset

The WorldExpo'10 dataset [3] contains 1132 annotated video sequences captured by 108 surveillance cameras, all from Shanghai 2010 WorldExpo. This dataset provides a total of 199,923 annotated pedestrians at the centers of their heads in 3980 frames. The testing dataset includes five video sequences of different scenes, and each video sequence contains 120 labeled frames. The regions of interest (ROI) and



(a) (b) (c) (d)

Fig. 5. Qualitative results of our method on test scenes of WorldExpo'10 dataset, including (a) scene images, (b) body part maps, (c) density maps, and (d) head density maps. These maps model the body-part structures and the density distributions of crowds.

perspective maps of scenes are provided for the train and test scenes.

Table III reports the MAE errors of different methods on WorldExpo'10 dataset. The results of LBP features based ridge (RR) regression method are listed at the top row. Zhang *et al.* [3] propose the Crowd CNN model to estimate the density maps and crowd counts of image patches based on deep neural networks. Results of their model are listed at the second row. The third row lists the results of the Crowd CNN model with the scene-specific fine-tuning technique which utilizes the information of test scenes. Zhang *et al.* [4] propose a Multi-column CNN (MCNN) model which uses filters of different sizes to estimate the geometry-adaptive density map. Results of their model are listed at the fourth row. The last row lists the results of our method. Our method achieves the best performance in terms of average MAE. In scene 3, 4, and 5, our method achieves the best performance compared with the other methods. It indicates that the semantic structure information modelled by our method is effective in different scenes. In scene 2, the performance of our method is relatively worse, mainly because the crowds of this scene are extremely dense within small areas such that there are very few body-part semantic cues. As a comparison, the fifth row lists the results of our method using conventional density map instead of structured density map. The introduction of structured density map makes 11% improvement over conventional density map for our model.

Fig. 5 shows some qualitative results of our method on the test scenes of WorldExpo'10 dataset. The figures are respectively the (a) scene images, (b) body part maps, (c) density maps and (d) head density maps from left to right. In the body part maps, colors of light blue, green, yellow and dark blue respectively denote the regions of head, body, legs and background. The body part maps show that our method can detect precise pedestrian body parts even under the heavy occlusions among crowds. The density maps are constructed to model the density distributions of pedestrians. The head density maps are inferred based on the body part maps and density maps, indicating that our method is also able to predict precise locations and densities of pedestrian heads.

TABLE IV
COMPARISON OF DIFFERENT METHODS ON THE SHANGHAITECH-B DATASET

Method	MAE	MSE
LBP+RR	59.1	81.7
Crowd CNN [3]	32.0	49.8
MCNN-CCR [4]	70.9	95.9
MCNN [4]	26.4	41.3
Ours	**20.2**	**35.6**

Thus the accurate pedestrian counts are estimated based on these effective body part maps and density maps.

C. Shanghaitech-B Dataset

The Shanghaitech-B dataset is a part of Shanghaitech dataset which was first introduced by Zhang *et al.* [4]. It contains 716 annotated images which are taken from different cameras at the busy streets of metropolitan areas in Shanghai. This dataset has a total of 88,488 annotated pedestrians at the centers of their heads. In this dataset, 400 images are used for training and 316 images are used for testing. Because the perspective maps are not provided and the perspective distortions of scenes are similar among scenes, we manually create a single perspective map which is used for all the images.

Table IV reports the MAE and MSE errors of different methods on the Shanghaitech-B dataset. Following the convention of Zhang *et al.* [4], we compare our method with the LBP+RR method, the Crowd CNN method [3], and the MCNN method [4]. The MCNN-CCR is the MCNN model trained without the ground truth of the density map. Our method outperforms the state-of-the-art methods by large margins in terms of both the MAE and MSE. It indicates that our method has a good generalization capability over many different scenes. Compared to MCNN-CCR which is based on pedestrian count regression, the MCNN method performs much better because it preserves more density information of the image. Likewise, our method proves better performance than MCNN because we further capture the pedestrian body-part structure information to help improve the count accuracy.

D. UCSD Dataset

The UCSD dataset [15] contains 2000 frames of a single scene. The video in this dataset is recorded at 10 fps with the frame size of 158×238. The crowd density of this dataset is relatively low that there are only about 25 persons on average in each frame. The annotations of pedestrian head locations and ROI are provided. Following the convention of the existing works [4], [15], we use frames 601-1400 as the training data, and the remaining 1200 frames as the test data. Since the perspective map is not provided in this dataset and the perspective distortions of the scene are not severe, we fix the perspective values of all the pixels to the same.

Table V reports the MAE and MSE errors of our method and other methods [3], [4], [12], [15], [49], [50]. Four hand-crafted features based regression methods are compared in Table V, including the kernel ridge regression [49], the multi-output ridge regression [12], the Gaussian process regression [15], and the cumulative attribute regression [50]. Results of the CNN based density estimation methods [3], [4] are also listed in Table V. Our method outperforms both the regression based methods and the CNN based methods in terms of MAE. The MSE of our method is a little larger than the MCNN method, mainly because the multi-size filters proposed in their methods are more robust for datasets without annotated perspective values. Except the MCNN method, other methods including ours do not specifically optimize the perspective distortions. Our method outperforms these methods by large margins in terms of both two metrics, because there are abundant pedestrian body-part information in relatively low-density scenarios. It also demonstrates the effectiveness of semantic scene models proposed in this paper.

TABLE V
COMPARISON OF DIFFERENT METHODS ON THE UCSD DATASET

Method	MAE	MSE
Kernel Ridge Regression [49]	2.16	7.45
Multi-output Ridge Regression [12]	2.25	7.82
Gaussian Process Regression [15]	2.24	7.97
Cumulative Attribute Regression [50]	2.07	6.86
Crowd CNN [3]	1.60	3.31
MCNN [4]	1.07	**1.35**
Ours	**1.00**	1.40

TABLE VI
COMPARISON OF DIFFERENT METHODS ON THE UCF_CC_50 DATASET

Method	MAE	MSE
Density-aware detection [19]	655.7	697.8
Density estimation [6]	493.4	**487.1**
Multi-source fusion [48]	419.5	541.6
Crowd CNN [3]	467.0	498.5
Ours+no global fine-tune	424.4	584.1
Ours	**409.5**	563.7

E. UCF_CC_50 Dataset

The UCF_CC_50 dataset [48] contains 50 images of different scenes. It is very challenging, because of not only the limited number of images, but also the extremely dense crowds in images. Following the conventional setting [48], we split the dataset randomly and perform 5-fold cross validation. Because of the limitation of the training samples, we randomly crop 1000 patches from each image for training. The perspective values are fixed as the perspective maps are not provided.

Table VI reports the MAE and MSE errors of our method and the other methods [3], [4], [6], [19], [48]. Rodriguez *et al.* [19] propose the density map estimation to improve the head detection performance in crowd scenes. Lempitsky and Zisserman [6] learn the density regression model based on dense SIFT features and the MESA distance. Idrees *et al.* [48] estimate the crowd counts based on multi-source features. The deep learning based approach [3] are also evaluated on this dataset. Our method performs the

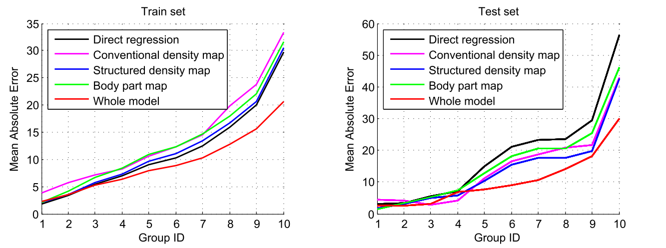


Fig. 6. Comparison of different ground truths on WorldExpo'10 dataset.

best in terms of MAE, indicating that the semantic structure information is still effective in extremely dense scenes. In addition, we also evaluate the performance of our model without the global fine-tuning operation described as Eq. 8 in subsection III-D. The results show that the global fine-tuning operation is able to help improve the performance of the patch-wise based crowd counting models.

F. The Effectiveness of Different Supervised Labels

There are three different ground truth supervised labels used in this work, including the structured density map $\mathcal{D}$, the body part map $\mathcal{B}$ and the pedestrian number C. We compares the effectiveness of themselves in crowd counting, as shown in Fig. 6. We evenly group the training images and testing images of WorldExpo'10 dataset into 10 groups according to increasing pedestrian number. The vertical axis denotes the MAE error in each group. The black curve represents the direct regression network trained by C only. The purple curve represents the network trained by conventional density map $\mathcal{D}_{\mathcal{N}}$ and C. The blue curve represents the network trained by structured density map $\mathcal{D}$ and C. The green curve represents the network trained by $\mathcal{B}$ and C. The red curve represents our whole model which is trained by all the three ground truths $\mathcal{B}$, $\mathcal{D}$, and C.

The whole model performs the best on both the training set and test set. It indicates that the joint modelling of different semantic attributes of crowd images provides an effective and robust solution for crowd counting. The direct regression network performs well on the train set but the worst on the test set. It is in line with our intuition that the crowd counting model trained by only the count information is short of generalization capability. On the test set, the body part map network performs better than the direct regression network, indicating that the body part map is an effective supervised label in crowd counting. While, it performs a little worse than the density map networks, indicating that the density map may make bigger contribution to our framework than body part map. In addition, the structured density map performs better than conventional density map in most cases, indicating that the structured density map can help improve the crowd counting performance.

In addition, Fig. 7 shows more qualitative results of our crowd counting models which are trained by different supervised labels. From left to right, the figures are respectively the scene images, the ground truth body part maps, the body part maps inferred by our model, and the pedestrian numbers estimated by different models. The scene images are of the test set of WorldExpo'10 dataset. The body part maps inferred by our model can well model the body-part semantic structures of pedestrians in scenes of different crowd densities and diverse crowd distributions, thus helps estimate the accurate pedestrian numbers.

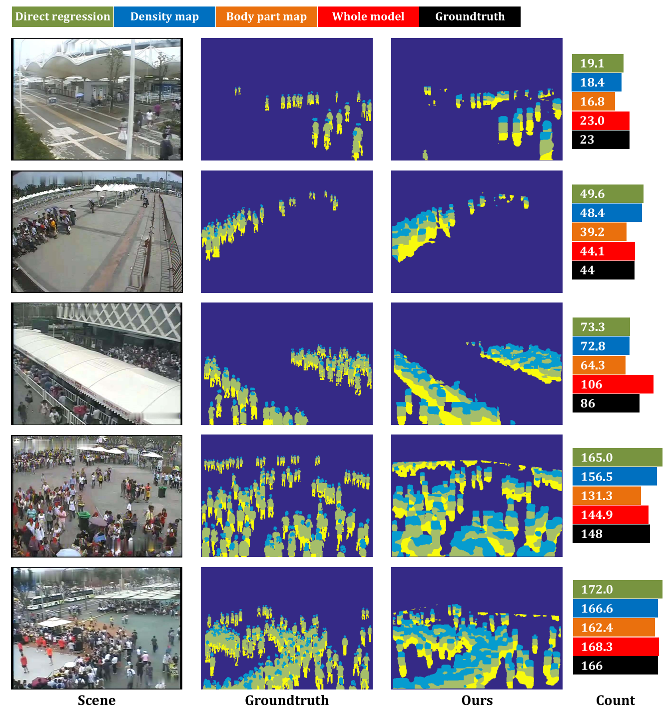


Fig. 7. From left to right: scene images, ground truth body part maps, body part maps generated by our method, and estimated pedestrian numbers. **Best viewed in color.**

V. Conclusion

In this paper, we have presented a novel approach to accurately estimate the count of crowds in images. Our approach has focused on discovering the semantic nature of crowd counting. We have built two semantic scene models to recover rich semantic structure information from images. In addition, we have reformulated the crowd counting problem as a multi-task learning problem such that the semantic scene models have been turned into different sub-tasks. We have built the CNNs to jointly learn these sub-tasks in a unified scheme. In experiments, our approach has achieved better performance compared to the state-of-the-art methods on four benchmark datasets.

References

[1] B. Wu and R. Nevatia, "Detection of multiple, partially occluded humans in a single image by Bayesian combination of edgelet part detectors," in *Proc. IEEE ICCV*, vol. 1. Oct. 2005, pp. 90–97.

[2] Z. Lin and L. S. Davis, "Shape-based human detection and segmentation via hierarchical part-template matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 604–618, Apr. 2010.

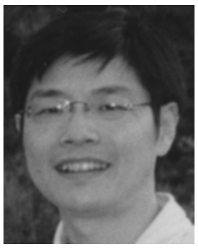
[3] C. Zhang, H. Li, X. Wang, and X. Yang, "Cross-scene crowd counting via deep convolutional neural networks," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 833–841.

[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. CVPR*, Jun. 2016, pp. 589–597.

[5] P. Luo, X. Wang, and X. Tang, "Pedestrian parsing via deep decompositional network," in *Proc. IEEE ICCV*, Dec. 2013, pp. 2648–2655.
[6] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Proc. Adv. NIPS*, 2010, pp. 1324–1332.
[7] W. Ge and R. T. Collins, "Marked point processes for crowd counting," in *Proc. IEEE Conf. CVPR*, Jun. 2009, pp. 2913–2920.
[8] M. Wang and X. Wang, "Automatic adaptation of a generic pedestrian detector to a specific traffic scene," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 3401–3408.
[9] V. B. Subburaman, A. Descamps, and C. Carincotte, "Counting people in the crowd using a generic head detector," in *Proc. IEEE Conf. AVSS*, Sep. 2012, pp. 470–475.
[10] 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. IEEE ICPR*, Dec. 2008, pp. 1–4.
[11] A. N. Marana, L. F. Costa, R. A. Lotufo, and S. A. Velastin, "On the efficacy of texture analysis for crowd monitoring," in *Proc. IEEE SIBGRAPI*, Oct. 1998, pp. 354–361.
[12] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting," in *Proc. BMVC*, 2012, vol. 1. no. 2, p. 3.
[13] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and Bayesian regression," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2160–2177, Apr. 2012.
[14] D. Kong, D. Gray, and H. Tao, "A viewpoint invariant approach for crowd counting," in *Proc. IEEE ICPR*, vol. 3. Aug. 2006, pp. 1187–1190.
[15] 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. CVPR*, Jun. 2008, pp. 1–7.
[16] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Crowd counting using multiple local features," in *Proc. IEEE DICTA*, Dec. 2009, pp. 81–88.
[17] N. Paragios and V. Ramesh, "A MRF-based approach for real-time subway monitoring," in *Proc. IEEE Conf. CVPR*, vol. 1. Jun. 2001, pp. I-1034–I-1040.
[18] M. von Borstel, M. Kandemir, P. Schmidt, M. K. Rao, K. Rajamani, and F. A. Hamprecht, "Gaussian process density counting from weak supervision," in *Proc. ECCV*, 2016, pp. 365–380.
[19] M. Rodriguez, I. Laptev, J. Sivic, and J.-Y. Audibert, "Density-aware person detection and tracking in crowds," in *Proc. IEEE ICCV*, Nov. 2011, pp. 2423–2430.
[20] Y. Wang, Y. X. Zou, J. Chen, X. Huang, and C. Cai, "Example-based visual object counting with a sparsity constraint," in *Proc. IEEE ICME*, Jul. 2016, pp. 1–6.
[21] Y. Wang and Y. Zou, "Fast visual object counting via example-based density estimation," in *Proc. IEEE ICIP*, Sep. 2016, pp. 3653–3657.
[22] L. Fiaschi, U. Köthe, R. Nair, and F. A. Hamprecht, "Learning to count with regression forest and structured labels," in *Proc. IEEE ICPR*, Nov. 2012, pp. 2685–2688.
[23] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "COUNT forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in *Proc. IEEE ICCV*, Dec. 2015, pp. 3253–3261.
[24] E. Walach and L. Wolf, "Learning to count with CNN boosting," in *Proc. ECCV*, 2016, pp. 660–676.
[25] C. Arteta, V. Lempitsky, and A. Zisserman, "Counting in the wild," in *Proc. ECCV*, 2016, pp. 483–498.
[26] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743–761, Apr. 2012.
[27] Q. Ye, Z. Han, J. Jiao, and J. Liu, "Human detection in images via piecewise linear support vector machines," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 778–789, Feb. 2013.
[28] A. Satpathy, X. Jiang, and H.-L. Eng, "Human detection by quadratic classification on subspace of extended histogram of gradients," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 287–297, Jan. 2014.
[29] Y. Bo and C. C. Fowlkes, "Shape-based pedestrian parsing," in *Proc. IEEE Conf. CVPR*, Jun. 2011, pp. 2265–2272.
[30] 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.
[31] K. Kang and X. Wang. (2014). "Fully convolutional neural networks for crowd segmentation." [Online]. Available: https://arxiv.org/abs/1411.4464
[32] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. NIPS*, 2014, pp. 568–576.
[33] Y. Yan, E. Ricci, S. Subramanian, G. Liu, and N. Sebe, "Multitask linear discriminant analysis for view invariant action recognition," *IEEE Trans. Image Process.*, vol. 23, no. 12, pp. 5599–5611, Dec. 2014.
[34] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 4657–4666.
[35] J. Walker, A. Gupta, and M. Hebert, "Patch to the future: Unsupervised visual prediction," in *Proc. IEEE Conf. CVPR*, Jun. 2014, pp. 3302–3309.
[36] S. Huang *et al.*, "Deep learning driven visual path prediction from a single image," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5892–5904, Dec. 2016.
[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. NIPS*, 2012, pp. 1097–1105.
[38] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.
[39] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 1–9.
[40] Y. Cheng, F. X. Yu, R. S. Feris, S. Kumar, A. Choudhary, and S.-F. Chang, "An exploration of parameter redundancy in deep networks with circulant projections," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 2857–2865.
[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. CVPR*, Jun. 2016, pp. 770–778.
[42] Y. Wang, C. Xu, S. You, D. Tao, and C. Xu, "CNNpack: Packing convolutional neural networks in the frequency domain," in *Proc. Adv. NIPS*, 2016, pp. 253–261.
[43] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. CVPR*, Jun. 2015, pp. 3431–3440.
[44] S. Zheng *et al.*, "Conditional random fields as recurrent neural networks," in *Proc. IEEE ICCV*, Dec. 2015, pp. 1529–1537.
[45] J. Yang, B. Price, S. Cohen, H. Lee, and M.-H. Yang, "Object contour detection with a fully convolutional encoder-decoder network," in *Proc. IEEE Conf. CVPR*, Jun. 2016, pp. 193–202.
[46] L. Wang, Y. Qiao, X. Tang, and L. Van Gool, "Actionness estimation using hybrid fully convolutional networks," in *Proc. IEEE Conf. CVPR*, Jun. 2016, pp. 2708–2717.
[47] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Multimedia*, 2014, pp. 675–678.
[48] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 2547–2554.
[49] S. An, W. Liu, and S. Venkatesh, "Face recognition using kernel ridge regression," in *Proc. IEEE Conf. CVPR*, Jun. 2007, pp. 1–7.
[50] K. Chen, S. Gong, T. Xiang, and C. C. Loy, "Cumulative attribute space for age and crowd density estimation," in *Proc. IEEE Conf. CVPR*, Jun. 2013, pp. 2467–2474.

Siyu Huang received the bachelor's degree in information and communication engineering from Zhejiang University, China, in 2014. He is currently pursuing the Ph.D. degree with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China. His advisors are Prof. Z. Zhang and Prof. X. li. His current research interests are primarily in computer vision, machine learning, and deep learning.

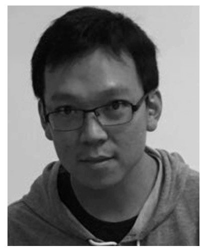
Xi Li received the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, in 2009. He is currently a Full Professor with the Zhejiang University, China. He was a Senior Researcher with the University of Adelaide, Australia. From 2009 to 2010, he was a Post-Doctoral Researcher with CNRS Telecomd ParisTech, France. His research interests include visual tracking, motion analysis, face recognition, Web data mining, and image and video retrieval.



Zhongfei Zhang received a B.S. degree (Hons.) in electronics engineering, an M.S. degree in information sciences from Zhejiang University, China, and the Ph.D. degree in computer science from the University of Massachusetts Amherst, USA. He is currently a QiuShi Chaired Professor with Zhejiang University, China, where he directs with Data Science and Engineering Research Center. He is on leave from The State University of New York Binghamton University, USA, where he is also a Professor with the Computer Science Department and also directs with the Multimedia Research Laboratory. His research interests include knowledge discovery from multimedia data and relational data, multimedia information indexing and retrieval, and computer vision and pattern recognition.

Fei Wu received the B.S. degree from the Lanzhou University, Lanzhou, Gansu, China, the M.S. degree from Macao University, Taipa, Macau, and the Ph.D. degree from Zhejiang University, Hangzhou, China. He is currently a Full Professor with the College of Computer Science and Technology, Zhejiang University. He was a Visiting Scholar with Prof. B. Yu's Group, University of California, Berkeley, from 2009 to 2010. His current research interests include multimedia retrieval, sparse representation, and machine learning.

Shenghua Gao received the B.E. degree from the University of Science and Technology of China, in 2008, and the Ph.D. degree from the Nanyang Technological University in 2013. He is currently an Assistant Professor with ShanghaiTech University, Shanghai, China. He was awarded the Microsoft Research Fellowship in 2010. His research interests include computer vision and machine learning.



Rongrong Ji is currently a Professor and the Director of the Intelligent Multimedia Technology Laboratory, and the Dean Assistant with the School of Information Science and Engineering, Xiamen University, Xiamen, China. His work mainly focuses on innovative technologies for multimedia signal processing, computer vision, and pattern recognition, with over 100 papers published in international journals and conferences. He is a Member of ACM. He also serves as Program Committee Member for several tier-1 international conferences. He was a recipient of the ACM Multimedia Best Paper Award and Best Thesis Award with the Harbin Institute of Technology. He serves as an Associate/Guest Editor for international journals and magazines, such as neurocomputing, signal processing, multimedia tools, and applications, the IEEE MULTIMEDIA MAGAZINE, AND MULTIMEDIA SYSTEMS.

Junwei Han received the Ph.D. degree in pattern recognition and intelligent systems from the School of Automation, Northwestern Polytechnical University, Xian, China, in 2003. He is currently a Professor with Northwestern Polytechnical University. His current research interests include multimedia processing and brain imaging analysis. He is an Associate Editor of the IEEE TRANSACTIONS ON HUMAN–MACHINE SYSTEMS, NEUROCOMPUTING, AND MULTIDIMENSIONAL SYSTEMS AND SIGNAL PROCESSING.