Cross-scene Crowd Counting via Deep Convolutional Neural Networks

Cong Zhang[1;2] Hongsheng Li[2] Xiaogang Wang[2] Xiaokang Yang[1]

[1]Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University

[2]Department of Electronic Engineering, The Chinese University of Hong Kong

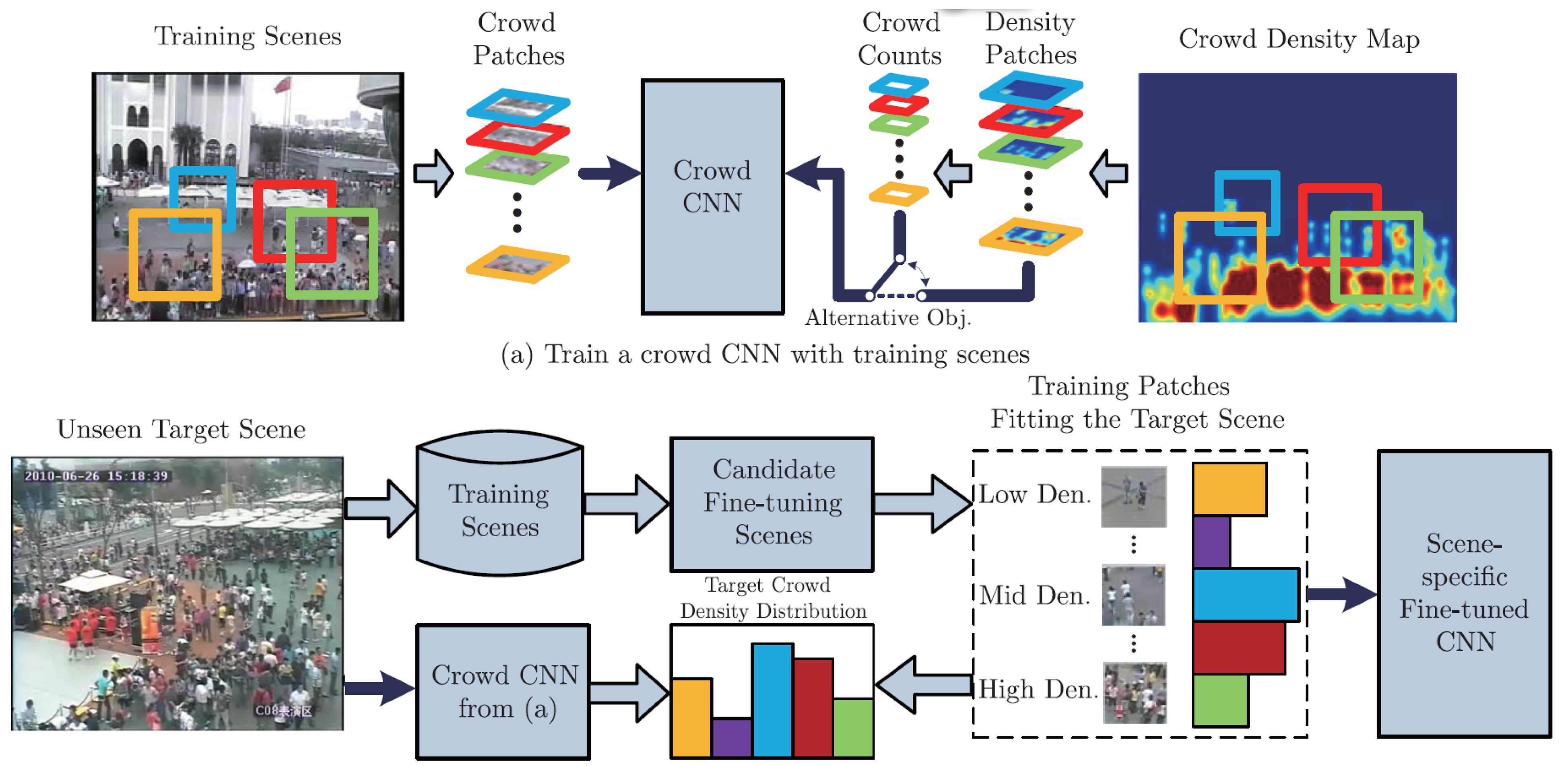
IEEE 2015 Conference on Computer Vision and Pattern Recognition

CVPR 2015 BOSTON JUNE 7-12

• Challenges for Cross-scene Crowd Counting

- Develop effective features to describe crowd. Previous works used general hand-crafted features, which have low representation capability for crowd.
- Without additional training data, the model trained in one specific scene has difficulty being used for other scenes.
- Foreground segmentation is indispensable for crowd counting.
- Existing datasets are not sufficient to evaluate cross-scene counting research.

• Illustration of Our Proposed Method



(a) Train a crowd CNN with training scenes

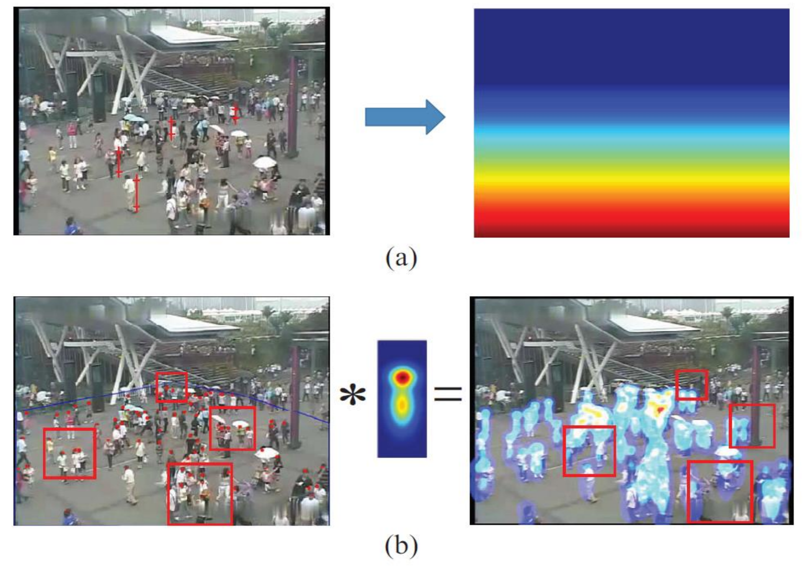
(b) Fine-tuning (adapting) the pre-trained CNN in (a) to a target scene

- Crowd CNN model is trained for crowd scenes by a switchable learning process.
- The target scenes require no extra labels in our framework for counting.
- The framework does not rely on foreground segmentation results.
- A new dataset is introduced for evaluating cross-scene crowd counting methods.

• Normalized Crowd Density Map for Training

- The crowd density map is created by the combination of several distributions with perspective normalization.
- The total crowd number in a selected training patch is calculated through integration over the density map

$$D_i(p) = \sum_{P \in \mathbf{P}_i} \frac{1}{\|Z\|}\left(N_h(p; P_h, \sigma_h) + N_b(p; P_b, \Sigma)\right)$$



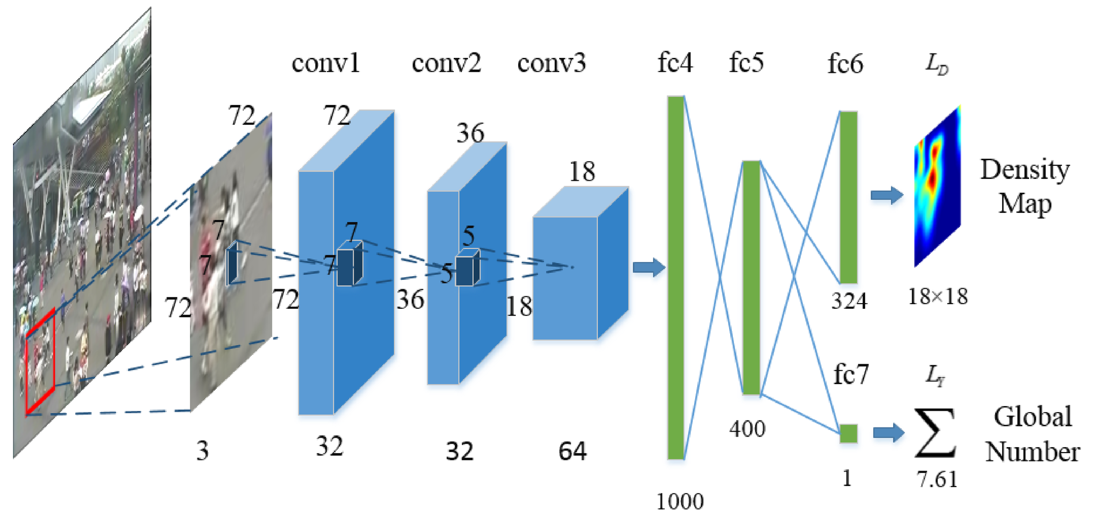
• Crowd Convolutional Neural Networks

- Our CNN model is trained for crowd scenes by a switchable learning process with two learning objectives, crowd density maps and crowd counts.
- The two different but related objectives can alternatively assist each other to obtain better local optima.
- Euclidean distance is adopted in these two objective losses.

- Switching training scheme vs. Joint training scheme

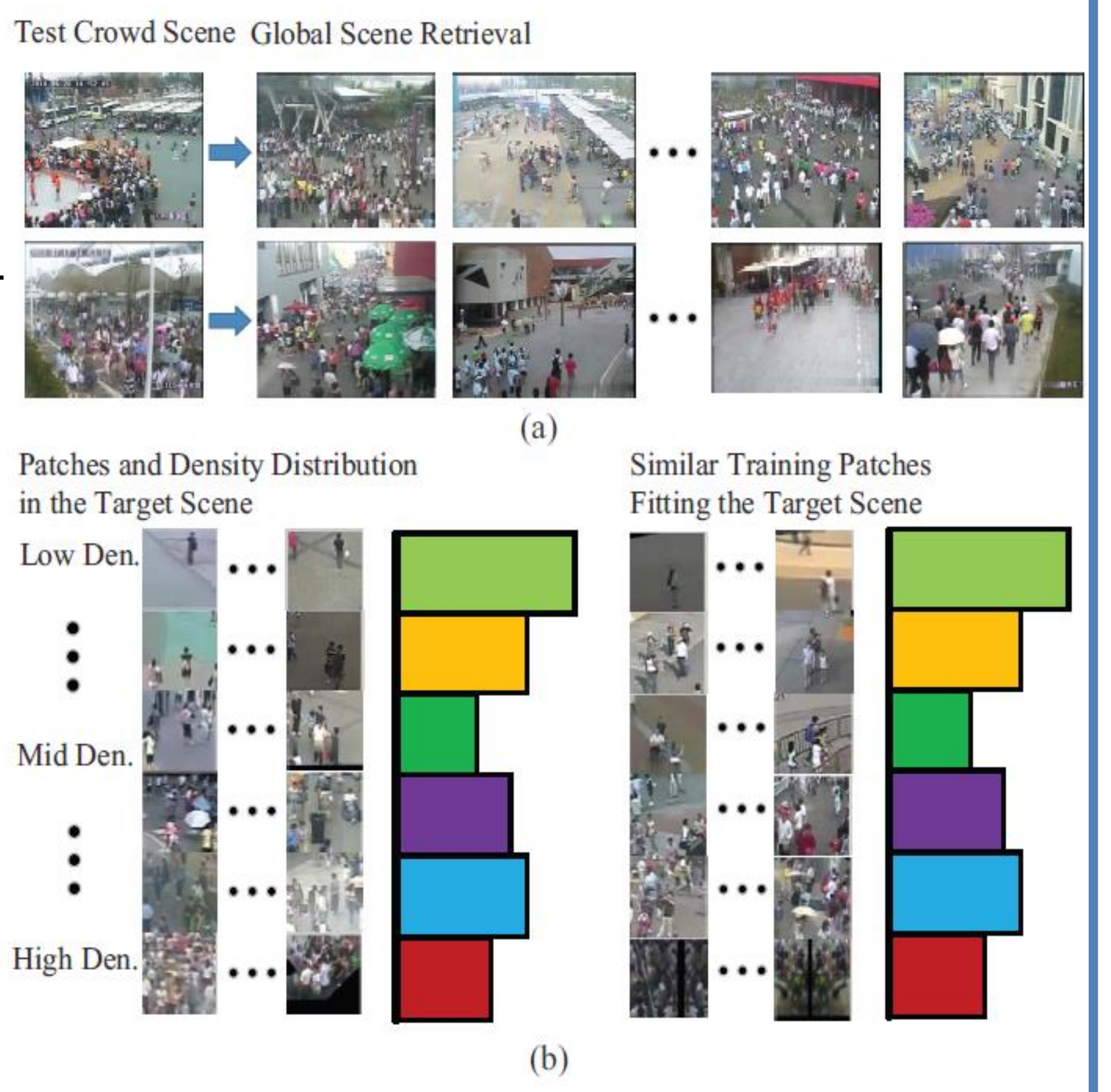
t	1	2	3	4	5	6
AMSE	17.4	15.5	14.9	14.3	14.1	14.3

λ	10	1	0.1	0.05	0.01	0.005
AMSE	50.8	50.8	18.5	15.5	15.3	15.5



• Nonparametric Fine-tuning for Target Scene

- In order to bridge the distribution gap between the training and test scenes, we design a nonparametric fine-tuning scheme to adapt our pre-trained CNN model to unseen target scenes.
- Given a target video from the unseen scenes, samples with similar properties from the training scenes are retrieved and added to training data to fine-tune the crowd CNN model.
- The retrieval task consists of two steps, candidate scenes retrieval and local patch retrieval.



Method	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Average
LBP+RR	13.6	58.9	37.1	21.8	23.4	31.0
Crowd CNN	10.0	15.4	15.3	25.6	4.1	14.1
Fine-tuned Crowd CNN	9.8	14.1	14.3	22.2	3.7	12.9

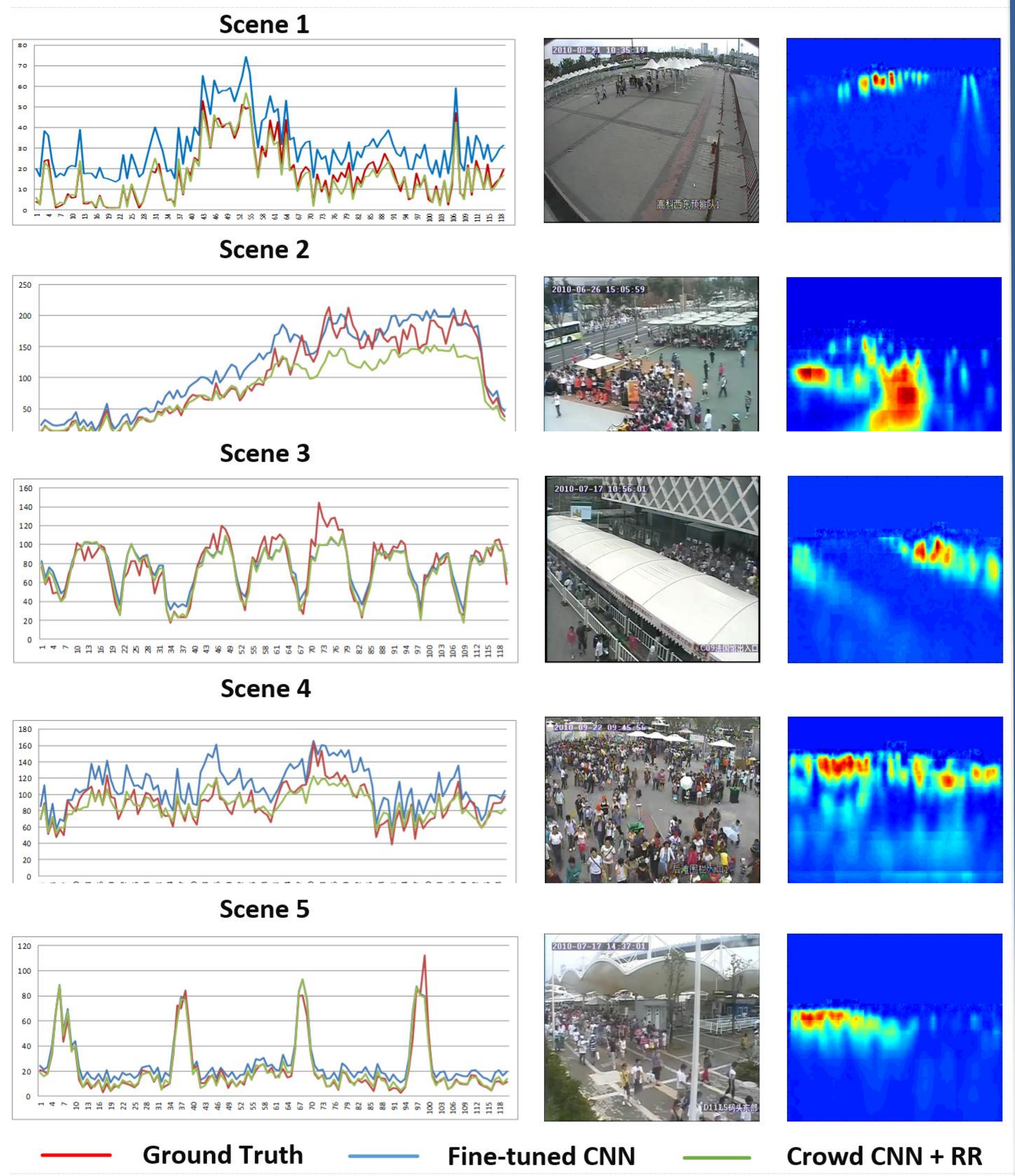
• WorldExpo'10 Crowd Counting Dataset

- To the best of our knowledge, the largest dataset for evaluating crowd counting algorithms. 103 scenes are treated as training and validation sets. The test set has 5 one-hour long video sequences from 5 different unseen scenes.

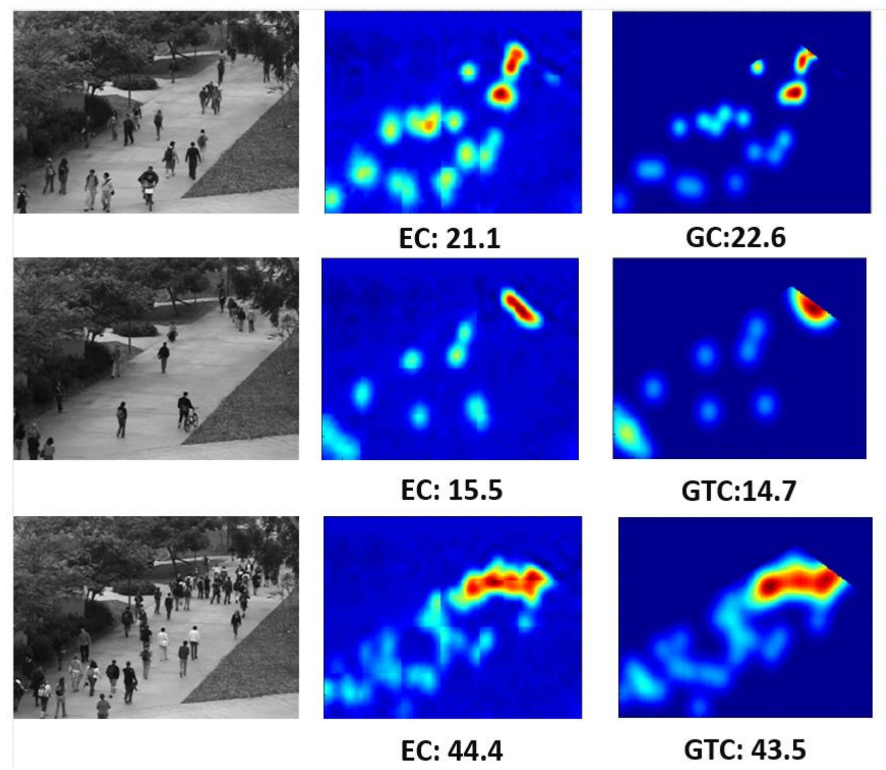
Dataset	N_f	N_c	R	FPS	D	T_p
UCSD	2000	1	158*238	10	11-46	49885
UCF_CC_50	50	50	–	image	94-4543	63974
WorldExpo	4.44 million	108	576*720	50	1-253	199923

• Experiment Results

- WorldExpo'10 dataset



- USCD dataset



Method	MAE	MSE
Kernel Ridge Regression [1]	2.16	7.45
Ridge Regression [6]	2.25	7.82
Gaussian Process Regression [4]	2.24	7.97
Cumulative Attribute Regression [5]	2.07	6.86
Our Crowd CNN Model	**1.60**	**3.31**