# RSANET: Deep Recurrent Scale-aware Network for Crowd Counting

Yujun Xie    Yao Lu    Shunzhou Wang

Session: ARS-15 -- Image & Video Mid-Level Analysis

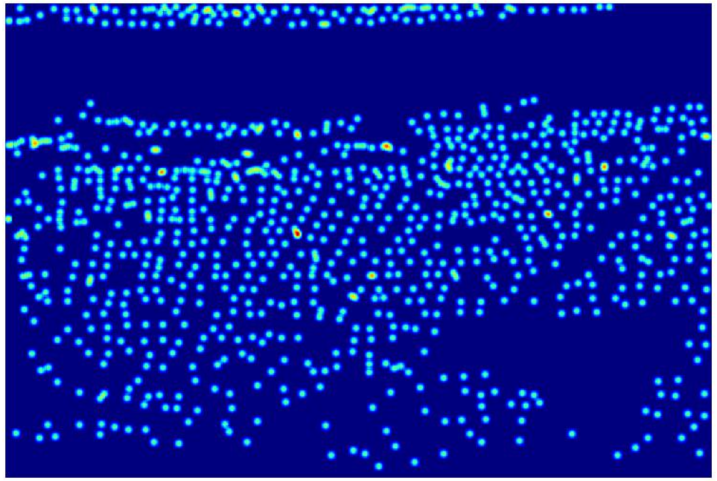
# Introduction

What is the task of crowd counting?


input image

Estimated Count:

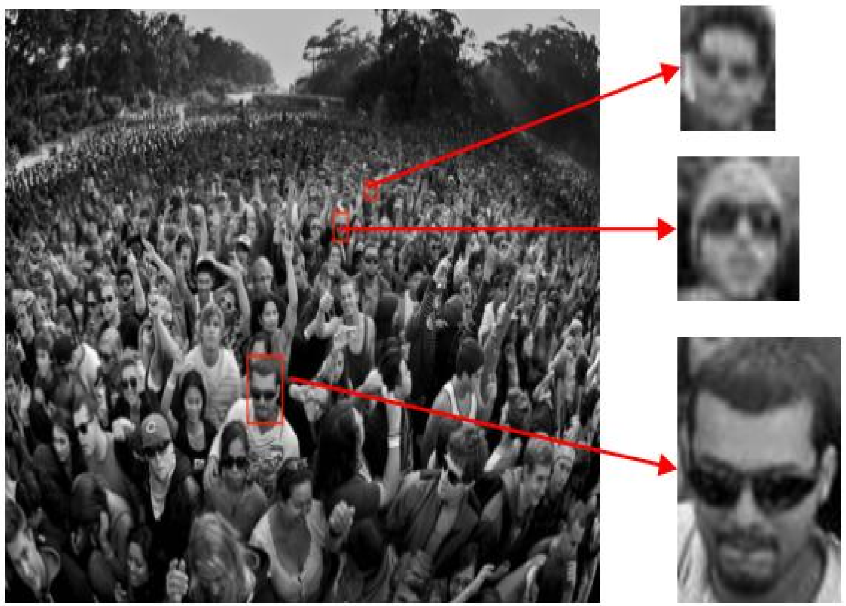
Density Map:



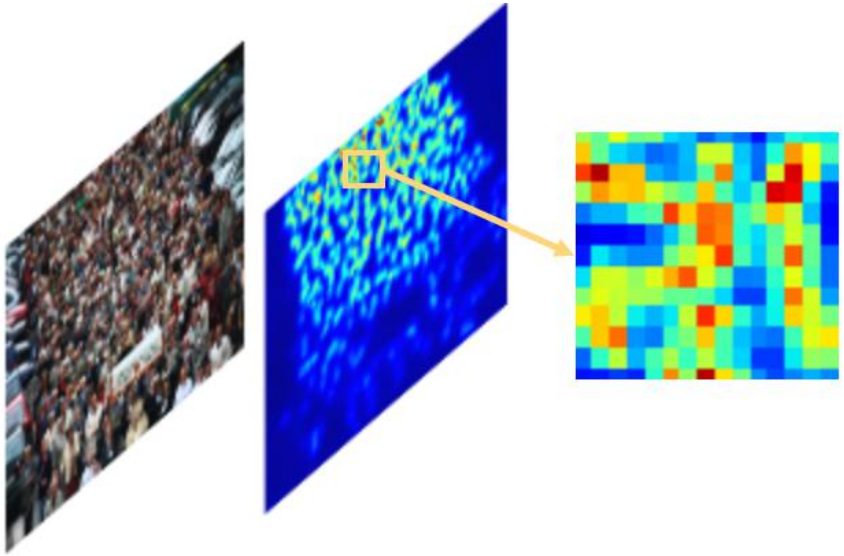
# Introduction

# Motivation

## Scale Variation



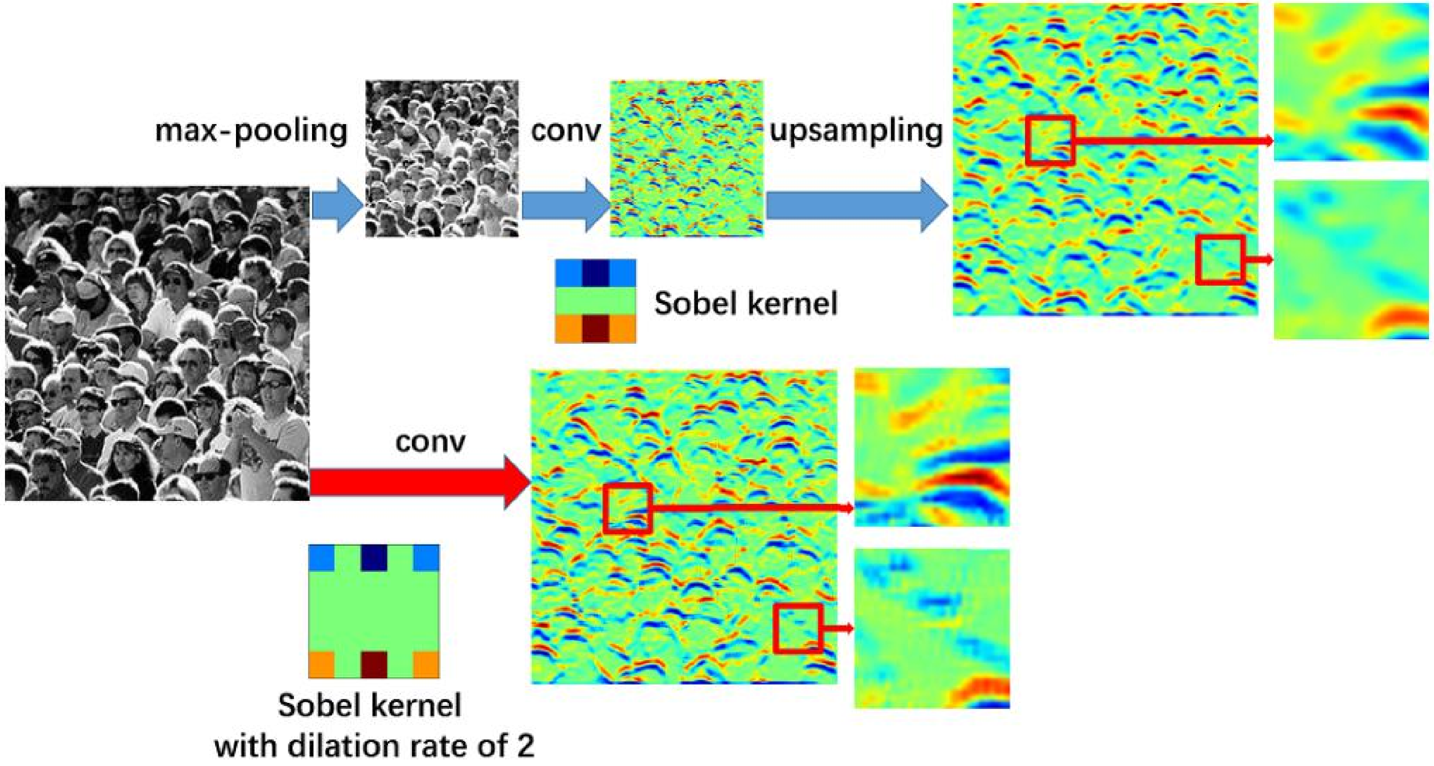
## High-Resolution



How to learn effective multi-scale features and resore high-resolution density maps?

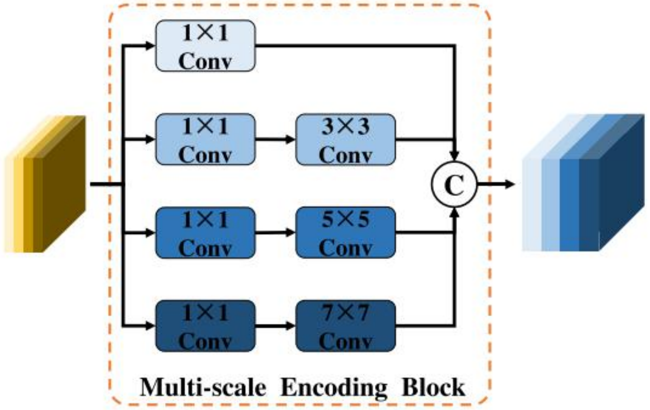
# Related Works - Network Architecture

| Configurations of CSRNet | | | |
|---|---|---|---|
| A | B | C | D |
| input(unfixed-resolution color image) | | | |
| front-end (fine-tuned from VGG-16) | | | |
| conv3-64-1 | | | |
| conv3-64-1 | | | |
| max-pooling | | | |
| conv3-128-1 | | | |
| conv3-128-1 | | | |
| max-pooling | | | |
| conv3-256-1 | | | |
| conv3-256-1 | | | |
| conv3-256-1 | | | |
| max-pooling | | | |
| conv3-512-1 | | | |
| conv3-512-1 | | | |
| conv3-512-1 | | | |
| back-end (four different configurations) | | | |
| conv3-512-1 | conv3-512-2 | conv3-512-2 | conv3-512-4 |
| conv3-512-1 | conv3-512-2 | conv3-512-2 | conv3-512-4 |
| conv3-512-1 | conv3-512-2 | conv3-512-2 | conv3-512-4 |
| conv3-256-1 | conv3-256-2 | conv3-256-4 | conv3-256-4 |
| conv3-128-1 | conv3-128-2 | conv3-128-4 | conv3-128-4 |
| conv3-64-1 | conv3-64-2 | conv3-64-4 | conv3-64-4 |
| conv1-1-1 | | | |

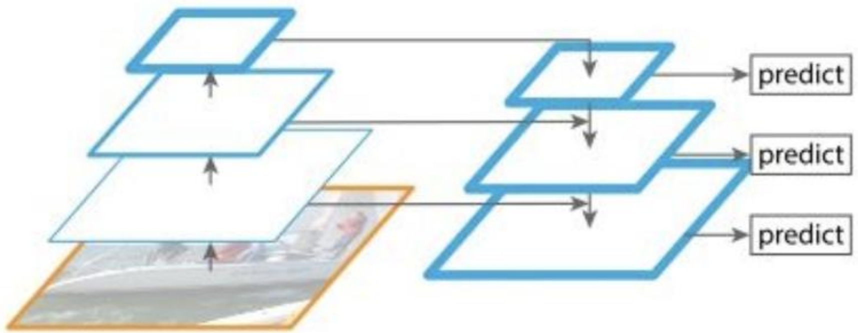



Y. Li, X. Zhang, and D. Chen, “CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes”. in CVPR, 2018.

# Related works - Scale Variation




mixed kernels




feature pyramid

perspective information

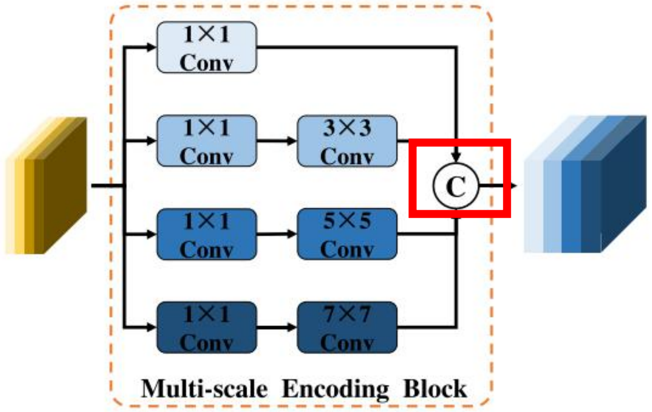
X. Jiang, Z. Xiao, B. Zhang, “Crowd counting and density estimation by trellis encoder-decoder networks,” in CVPR, 2019.

T. Lin , D. Piotr, R. Girshick, “Feature Pyramid Networks for Object Detection,” in CVPR, 2016.

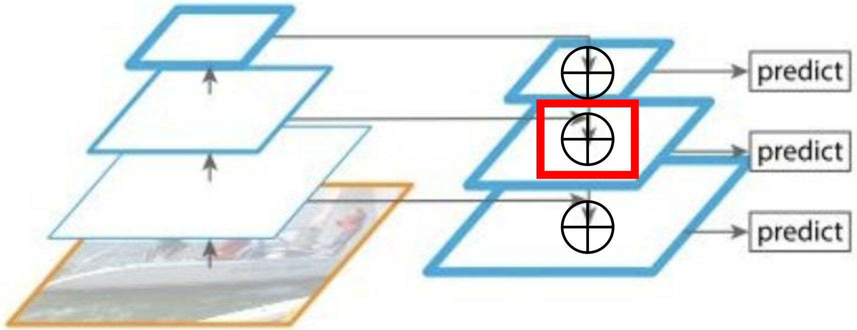
M. Shi, Z. Yang, C. Xu, and Q Chen, “Revisiting perspective information for efficient crowd counting,” in CVPR, 2019.

# Related works - Feature Fusion

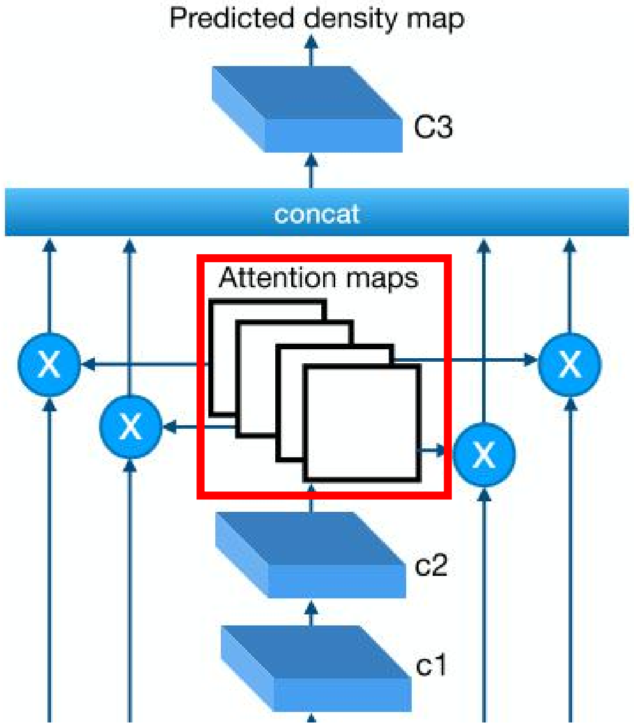
concatenation




element-wise addition




attention-guided fusion

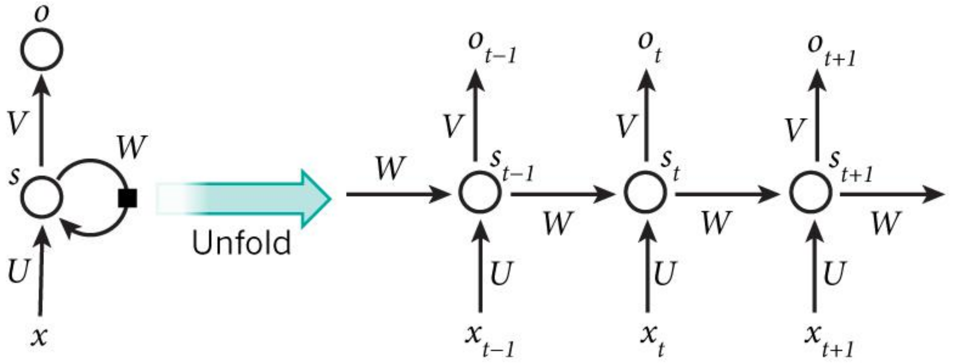



X. Jiang, Z. Xiao, B. Zhang, “Crowd counting and density estimation by trellis encoder-decoder networks,” in CVPR, 2019.

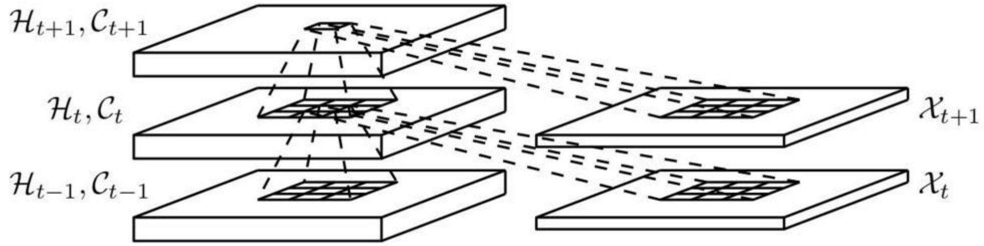
T. Lin , D. Piotr, R. Girshick, “Feature Pyramid Networks for Object Detection,” in CVPR, 2016.

V. A. Sindagi and V. M. Pateld, “Multi-level bottom-top and top-bottom feature fusion for crowd counting,” in ICCV, 2019.

# Related works - Feature Fusion

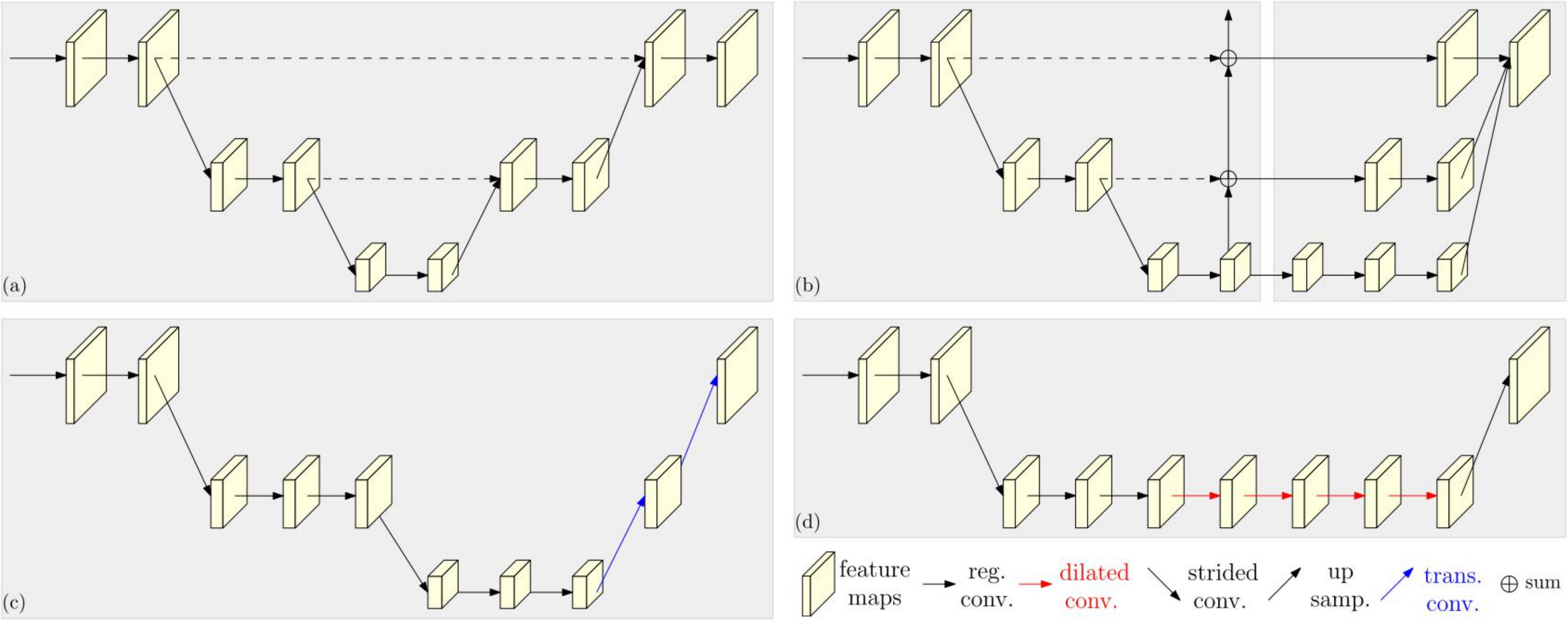



RNN




ConvLSTM

# Related works - High Resolution



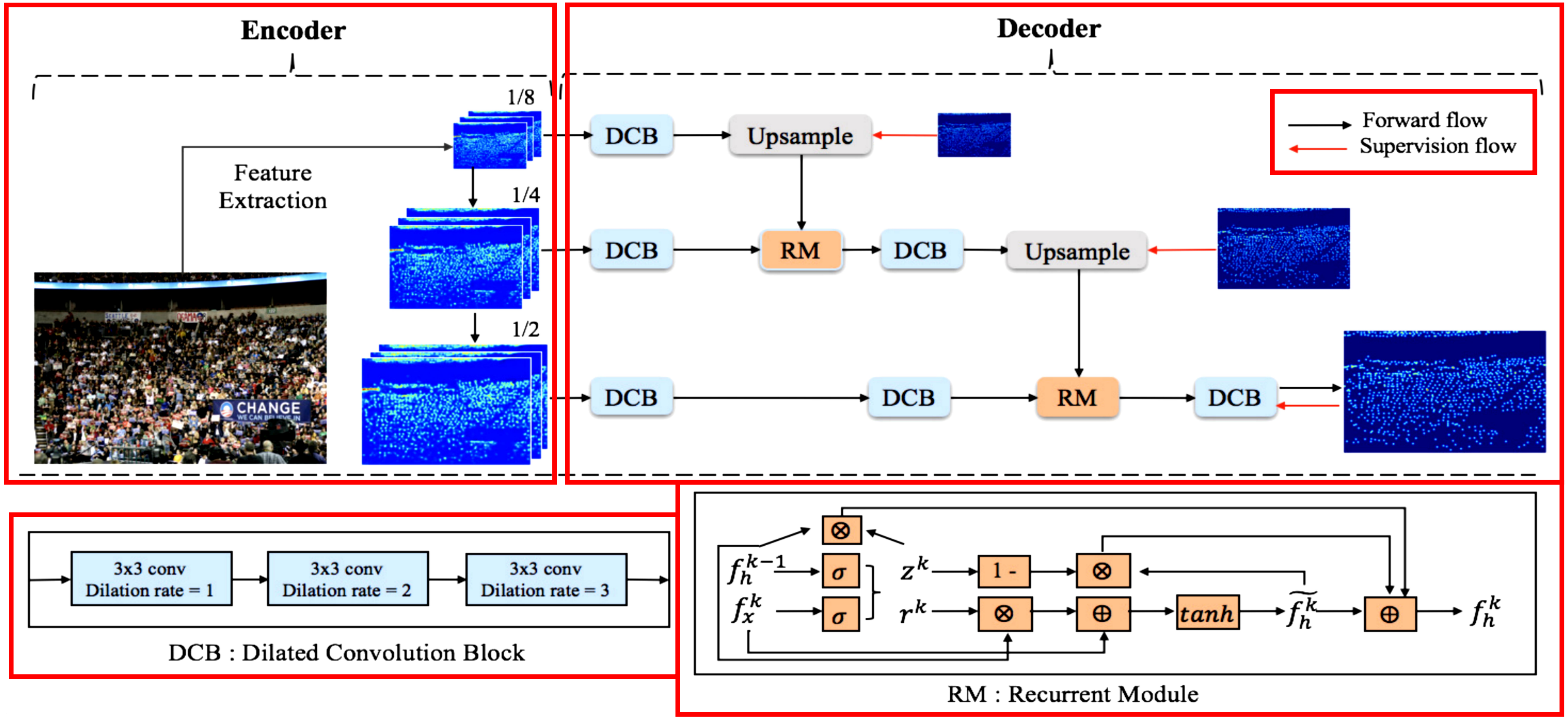

K, Sun, B. Xiao, D. Liu. “Deep High-Resolution Representation Learning for Human Pose Estimation,” in CVPR, 2019.

# Contributions

**RSANet -** a recurrent scale-aware network

- **a coarse-to-fine scheme** is introduced to gradually restore the high-resolution feature map;
- **a recurrent module** is deployed in feature fusion process to enrich feature representations;
- **a multi-resolution supervision strategy** is used for our network.

# Our Method - Recurrent scale-aware network



# Our Method - Multi-resolution supervision loss

$$L = \sum_{i=1}^{K} \sum_{j=1}^{N} ||P_{near}(D(I_j;\theta), k_i) - P_{near}(D_j^{GT}, k_i)||_2^2 \quad (5)$$

where $K$ is the number of scales, $N$ is the number of images in a batch, $D(I_j;\theta)$ is the estimated density map for training image $I_j$ with parameters $\theta$, $D_j^{GT}$ is the ground truth density map of $I_j$, $P_{ner}$ is the nearest neighbor upsampling operation to ensure the same scale of two density maps, $k_i$ is the specified output resolution of density map.

# Results - Datasets

## Congested Crowd

ShanghaiTech Part A: Congestion

482 images

241,677 heads

UCF-QNRF: High-resolution Data

1535 images

1,251,642 heads

## Sparse Crowd

ShanghaiTech Part B: Free-view Scenes

716 images

88,488 heads

# Results - ShanghaiTech dataset

| | part A | | part B | |
|---|---|---|---|---|
| Method | MAE | MSE | MAE | MSE |
| MCNN [1] | 110.2 | 173.2 | 26.4 | 41.3 |
| CSRNet [3] | 68.2 | 115.0 | 10.6 | 16.0 |
| SANet [2] | 67.0 | 104.5 | 8.4 | 13.6 |
| PACNN [20] | 66.3 | 106.4 | 8.9 | 13.5 |
| SFCN [21] | 64.8 | 107.5 | **7.6** | 13.0 |
| TEDNet [19] | 64.2 | 109.1 | 8.2 | 12.8 |
| Ours(RSANet) | **63.5** | **97.4** | 8.5 | **12.6** |

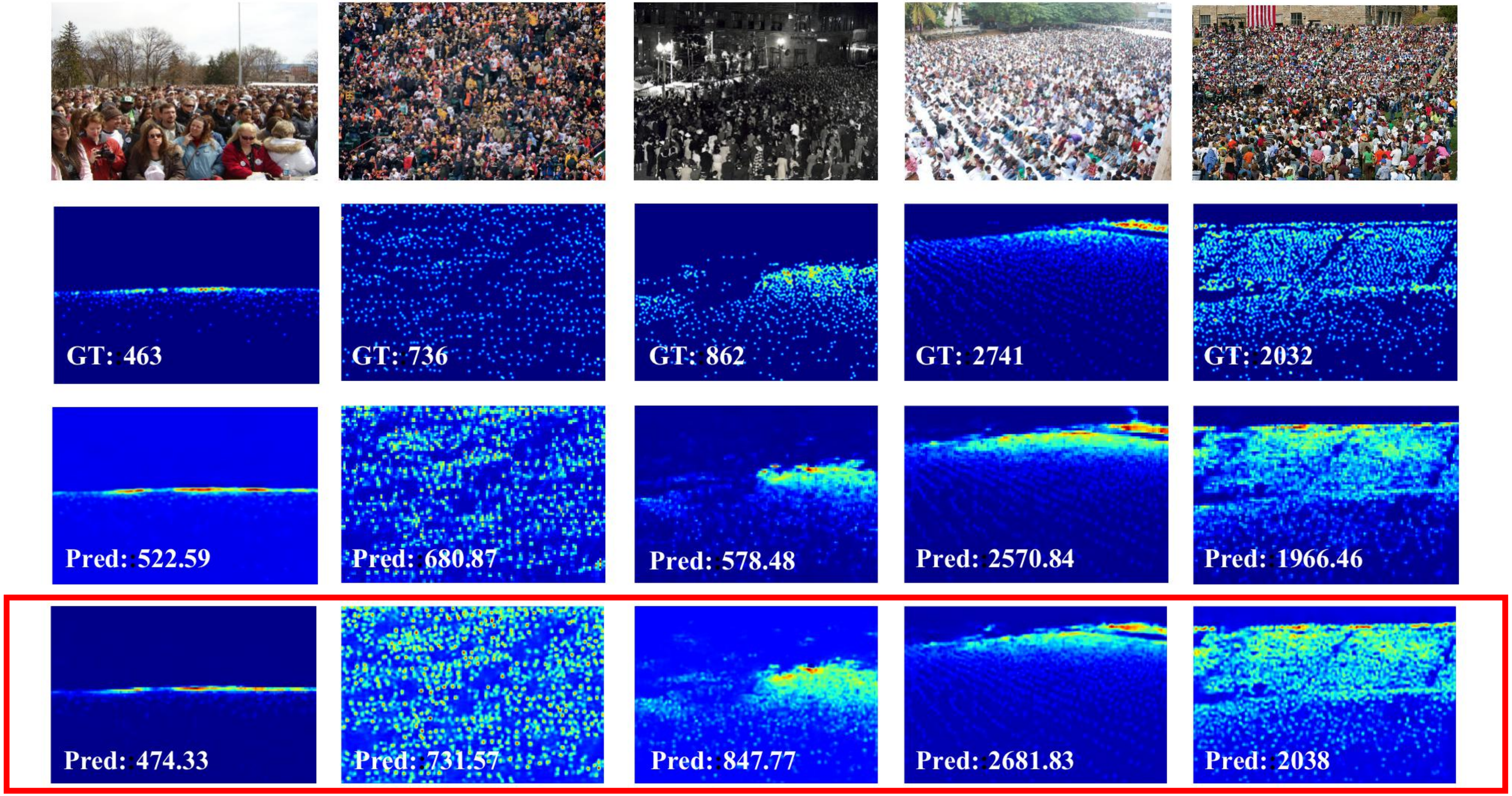
# Results - UCF-QNRF dataset

| Method | MAE | MSE |
|---|---|---|
| MCNN [1] | 277 | - |
| HA-CCN [22] | 118.1 | 180.4 |
| TEDNet [19] | 113 | 188 |
| CAN [18] | 107 | 183 |
| S-DCNet [23] | 104.4 | **176.1** |
| Ours(RSANet) | **102.9** | 181.9 |

# Results - Ablation study

| | Configuration | MAE | MSE |
|---|---|---|---|
| (a) | BaseNet | 68.2 | 112.6 |
| (b) | BaseNet+MRS | 65.3 | 104.2 |
| (c) | BaseNet+MRS+ConvLSTM | 65.2 | 104.5 |
| (d) | BaseNet+MRS+ConvGRU | **63.5** | **97.4** |

# Results - Visualization



# Conclusion

- Detailed information from low-level feature helps to restore high-resolution density map.
- Recurrent network is an effective way to guide learning process of scales adaptively.
- Task of crowd counting can draw on the experience of other density prediction tasks such as semantic segmentation, pose estimation, etc.

# Thank you!

RSANET: Deep Recurrent Scale-aware Network for Crowd Counting

Yujun Xie Yao Lu Shunzhou Wang