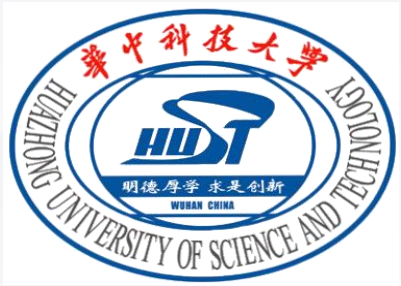


HANet: Hybrid Attention-aware Network for Crowd Counting

Xinxing Su*[1], Yuchen Yuan*[2], Xiangbo Su[2], Zhikang Zou[2], Shilei Wen[2], and Pan Zhou[1]

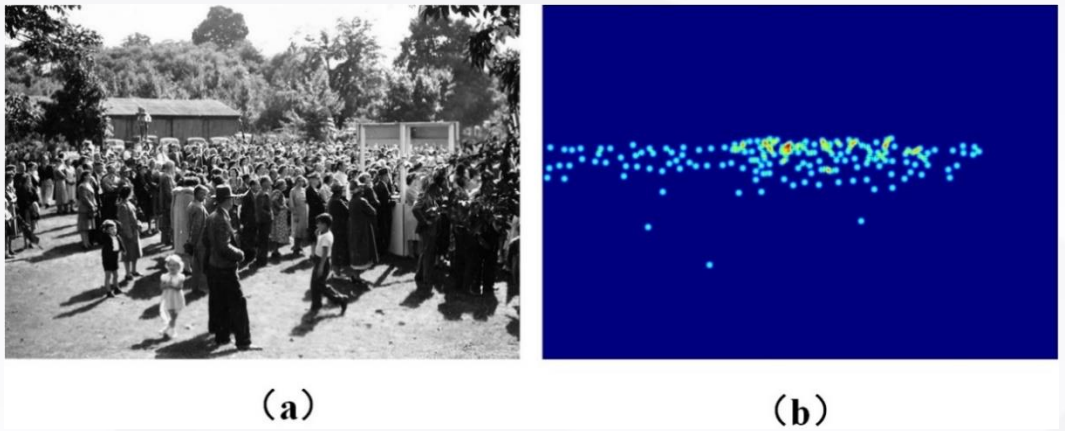
[1]Huazhong University of Science and Technology
[2]Baidu Inc.

CONTENTS

Motivation

◆ Task

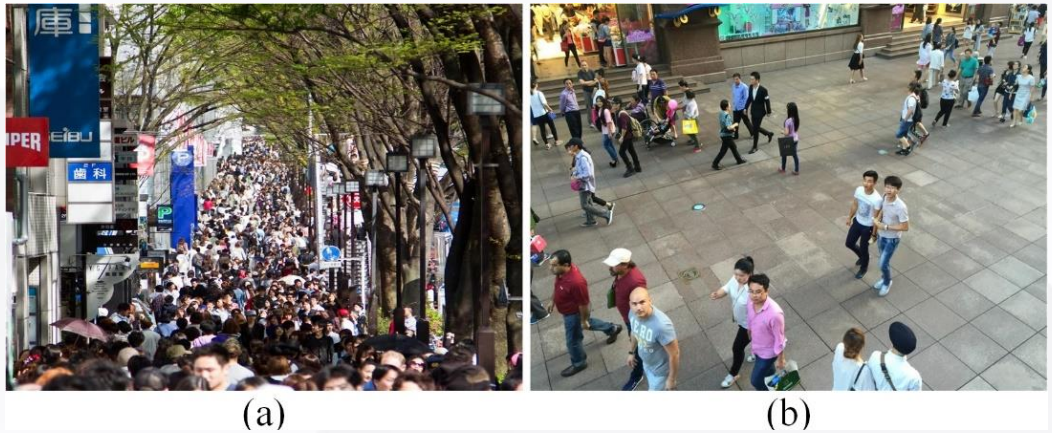
- Estimating the number of people
- Generating high-quality density maps



(a) Raw image; (b) Ground truth density map

◆ Challenges

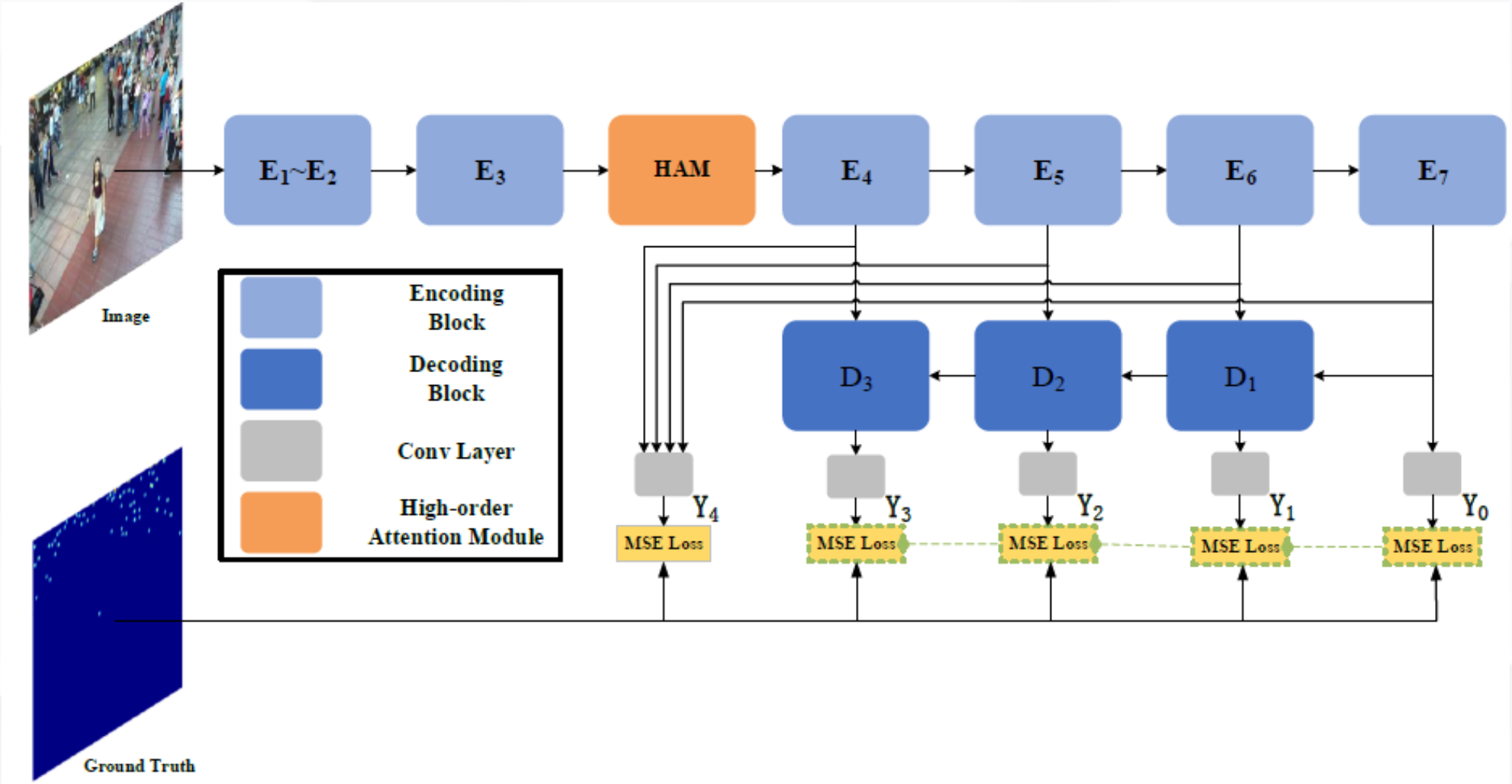
- non-uniform crowd distribution, noisy background, occlusions
- restricted to the 2D position-wise (or first-order) attention
- independence between different supervisions



Non-uniform crowd distribution, noisy background, occlusions

Overview

Pipeline



Our proposed HANet.
The green dotted line is the adaptive compensation loss (ACLoss).

Overview

Inspired by Kernel Pooling[26], we introduce the High-order Attention Module (HAM), which captures higher-ordered feature of the target via a kernel function $t(\mathrm{x})$:

$$t(\mathrm{x}) = \langle W, \phi(\mathrm{x}) \rangle = \sum_{r=1}^{R} \langle W^r, \otimes_r \mathrm{x} \rangle = \langle W^1, \mathrm{x} \rangle + \sum_{r=2}^{R} \langle W^r, \otimes_r \mathrm{x} \rangle ,$$

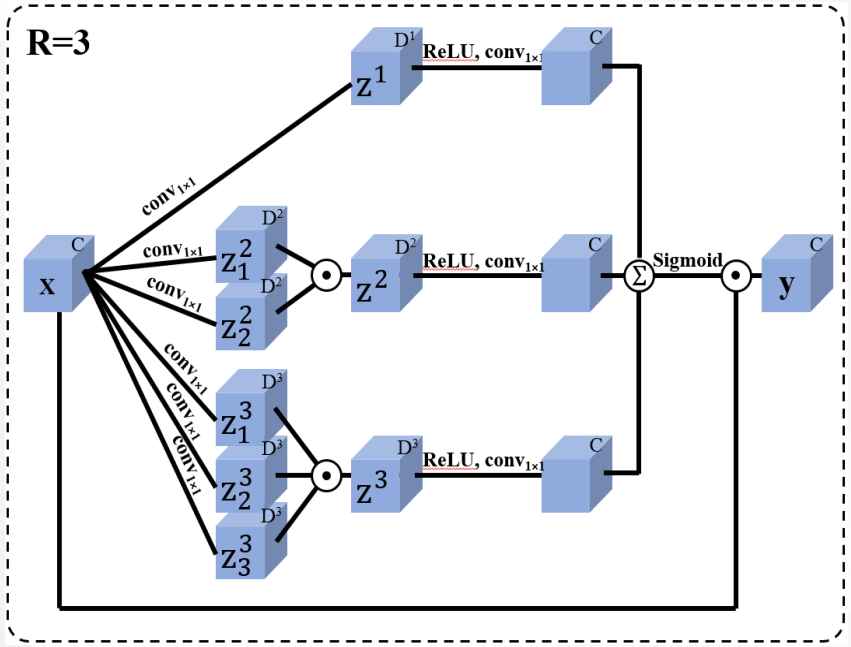
where $\langle \cdot, \cdot \rangle$ calculates the inner product, R is the maximum rank of the polynomial, W^r is a r-ranked learnable tensor.

Due to the sparsity of the CNN, in order to minimum the storage cost and risk of overfitting, we decompose W^r by [55]:

$$W^r = \sum_{d=1}^{D^r} \beta^{r,d} v_1^{r,d} \otimes v_2^{r,d} \otimes \cdots \otimes v_r^{r,d} ,$$

where W^r is decomposed into the summation of D^r tensors; v is the obtained tensor, and $\beta^{r,d}$ is the corresponding weight.

High-order Attention Module (HAM)



HAM with R=3

$t(\mathrm{x})$ is thus rewritten as:

$$t(\mathrm{x}) = \langle W^1, \mathrm{x}\rangle + \sum_{r=2}^{R} \left\langle \sum_{d=1}^{D^r} \beta^{r,d} v_1^{r,d} \otimes v_2^{r,d} \otimes \cdots \otimes v_r^{r,d}, \otimes_r \mathrm{x} \right\rangle$$

$$= \langle W^1, \mathrm{x}\rangle + \sum_{r=2}^{R} \sum_{d=1}^{D^r} \beta^{r,d} \prod_{s=1}^{r} \langle v_s^{r,d}, \mathrm{x}\rangle$$

$$= \langle W^1, \mathrm{x}\rangle + \sum_{r=2}^{R} \langle \beta^r, z^r\rangle$$

where $z^r = [z^{r,1}, z^{r,2}, \ldots, z^{r,D^r}]^T$和$z^{r,d} = \prod_{s=1}^{r} \langle v_s^{r,d}, \mathrm{x}\rangle$, $\beta^r = [\beta^{r,1}, \ldots, \beta^{r,D^r}]^T$. Let $W^1 = v^1\beta^1$, where $v^1 \in \mathbb{R}^{C\mathrm{x}D^1}$, $\beta^1 \in \mathbb{R}^{D^1\mathrm{x}C}$, $t(\mathrm{x})$ is then transformed as:

$$t(\mathrm{x}) = \langle \beta^1, (v^1)^T\mathrm{x}\rangle + \sum_{r=2}^{R} \langle \beta^r, z^r\rangle = \sum_{r=1}^{R} \langle \hat{\beta}^r, \hat{z}^r\rangle.$$

The Adaptation Compensation Loss (ACLoss)

Our supervision is:

$$\mathcal{L}_{total} = \sum_{l=1}^{B} \sum_{i=0}^{4} \| \hat{Y}_{l,i} - Y_{l,i} \|_2^2 ,$$

where B is the batchsize, i is the index of the decoder, $\hat{Y}_{l,i}$ is the density map of image l decoded from module i, and $Y_{l,i}$ is the corresponding ground truth. Furthermore, the ACLoss can be updated as:

$$\mathcal{L}_{total} = \sum_{l=1}^{B} \left(\sum_{i=1}^{3} \left(1 + w_{l,i}\right) \| \hat{Y}_{l,i} - Y_{l,i} \|_2^2 + \| \hat{Y}_{l,0} - Y_{l,0} \|_2^2 + \| \hat{Y}_{l,4} - Y_{l,4} \|_2^2 \right),$$

Where $w_{l,i} = \left| \varphi\left(\hat{Y}_{l,i-1}\right) - \varphi\left(Y_{l,i-1}\right) \right|$, $\varphi(\cdot)$ is the Sigmoid function.

Experimental Results

Quantitative Results

EXPERIMENTAL RESULTS ON THE SHANGHAI PART A, SHANGHAI PART B AND UCF_CC_50 DATASETS

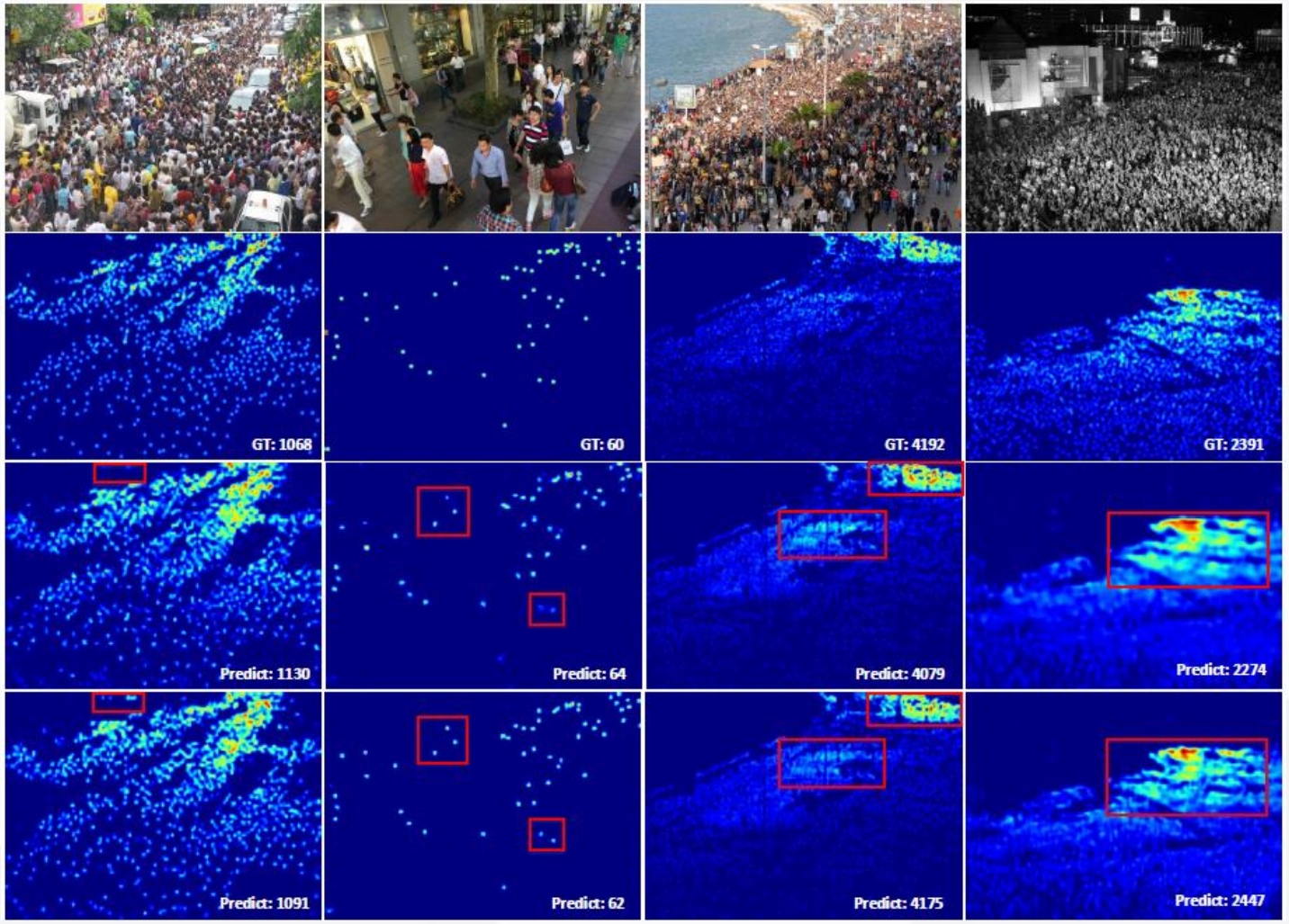
Method	Shanghai Part A		Shanghai Part B		UCF_CC_50	
	MAE↓	RMSE↓	MAE↓	RMSE↓	MAE↓	RMSE↓
MCNN [17]	110.2	173.2	26.4	41.3	377.6	509.1
CSRNet [6]	68.2	115.0	10.6	16.0	266.1	397.5
ADCrowdNet [10]	66.1	102.1	7.6	13.9	257.9	357.7
TEDnet [15]	64.2	109.1	8.2	12.8	249.4	354.4
DADNet [12]	64.2	99.9	8.8	13.5	285.5	389.7
CAN [11]	62.3	100.0	7.8	12.2	**212.2**	**243.7**
ADMG [37]	64.7	97.1	8.1	13.6	239.8	319.4
ANF [31]	63.9	99.4	8.3	13.2	250.2	340.0
MBTTBF-SCFB [28]	60.2	94.1	8.0	15.5	233.1	300.9
RANet [30]	59.4	102.0	7.9	12.9	239.8	319.4
HANet (ours)	**58.5**	**92.4**	**7.0**	**10.3**	225.9	269.1

EXPERIMENTAL RESULTS ON THE UCF-QNRF DATASET

Method	MAE↓	RMSE↓
MCNN [17]	277.0	426.0
Idress *et al.* [18]	132.0	191.0
DADNet [12]	113.2	189.4
TEDnet [15]	113.0	188.0
RANet [30]	111.0	190.0
CAN [11]	107.0	183.0
ANF [31]	110.0	174.0
ADMG [37]	101.0	176.0
MBTTBF-SCFB [28]	97.5	165.2
HANet (ours)	**95.3**	**160.3**

Experimental Results

Qualitative Results



Top to bottom: input, ground truth density map, inference result of CSRNet[6] and that of HANet.

Experimental Results

Ablation Study

PERFORMANCE COMPARISONS ON SHANGHAITECH PART A WITH RESPECT TO THE ORDER AND THE PLACEMENT OF HAM.

Metric	encoder	E_2				E_3				E_4			
		R=1	R=2	R=4	R=6	R=1	R=2	R=4	R=6	R=1	R=2	R=4	R=6
MAE	64.8	**63.5**	63.7	64.1	63.7	63.2	62.9	**61.9**	62.7	63.7	63.2	**62.3**	63.5

HAM And ACLoss:

base net	HAM	ACLoss	Shanghai Part A		UCF-QNRF	
			MAE↓	RMSE↓	MAE↓	RMSE↓
√	-	-	62.6	104.9	100.1	180.2
√	√	-	59.6	100.1	97.3	175.0
√	-	√	61.6	100.9	96.4	171.1
√	√	√	58.5	92.4	95.3	160.3

Summary

Contributions

- Establishing a 3D attention mechanism that captures high-order statistics, both position-wisely and channel-wisely
- Exploiting the difference between the prediction and the ground truth from a higher-level supervision path as an additional attention guidance

T H A N K Y O U F O R W A T C H I N G