Density Map Regression Guided Detection Network for RGB-D Crowd Counting and Localization

Dongze Lian[1*], Jing Li[1*], Jia Zheng[1], Weixin Luo[1,2], Shenghua Gao[1]

[1]ShanghaiTech University [2]Yoke Intelligence

上海科技大学 ShanghaiTech University

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction:

➢ Crowd counting: Estimate the number of persons in images or videos.

- Recent methods: density map regression (cannot give the position of heads)
- Ours: revisiting detection-based method.

(a) (b) (c)

➢ Challenges:

- Underestimation: the number of detected heads is much smaller than the total number of heads [1] (especially for tiny heads);
- ground-truth annotation: heavy workload than point annotation.

➢ Contributions:

- A regression guided detection network (RDNet) for RGB-D crowd counting and localization;
- Depth-adaptive kernel for density map generation, depth-anchor for anchor initialization, Depth-based bounding boxes ground-truth generation for training [figure (b)];
- A large-scale RGB-D crowd counting dataset (ShanghaiTechRGBD);
- Our method can be easily extended to RGB based counting and localization.

Our approach:

➢ Density map regression module:

- Fixed-kernel density map:

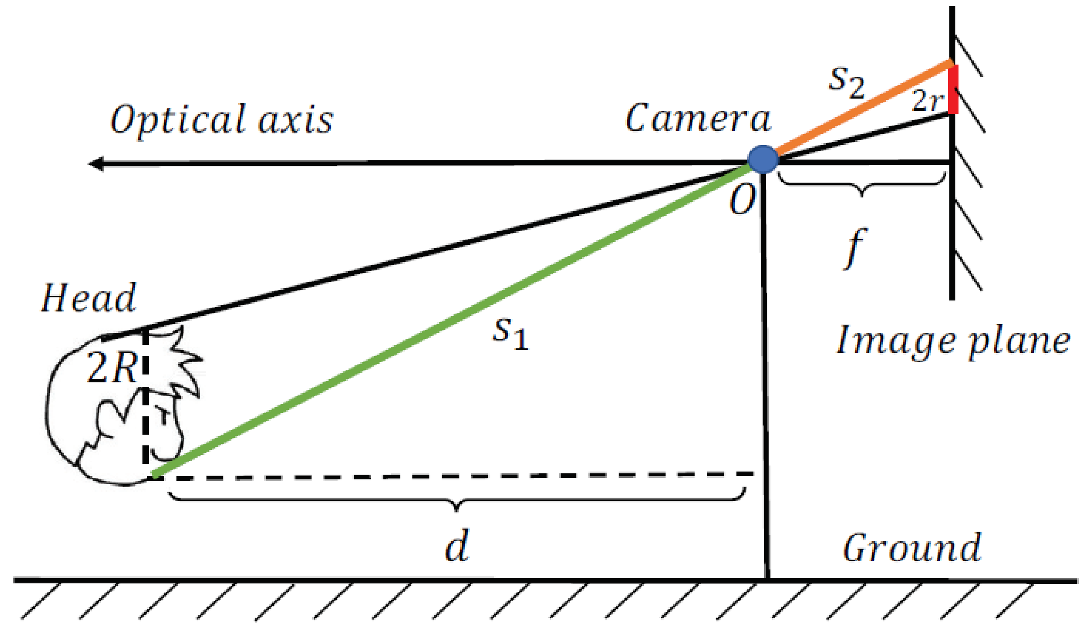
$$\mathcal{D}(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_\sigma(x).$$

- Depth-adaptive density map:

$$\frac{r}{R} = \frac{s_2}{s_1} = \frac{f}{d}$$

$$\sigma = \beta r = \beta \frac{Rf}{d} = \beta \frac{\gamma}{d}.$$

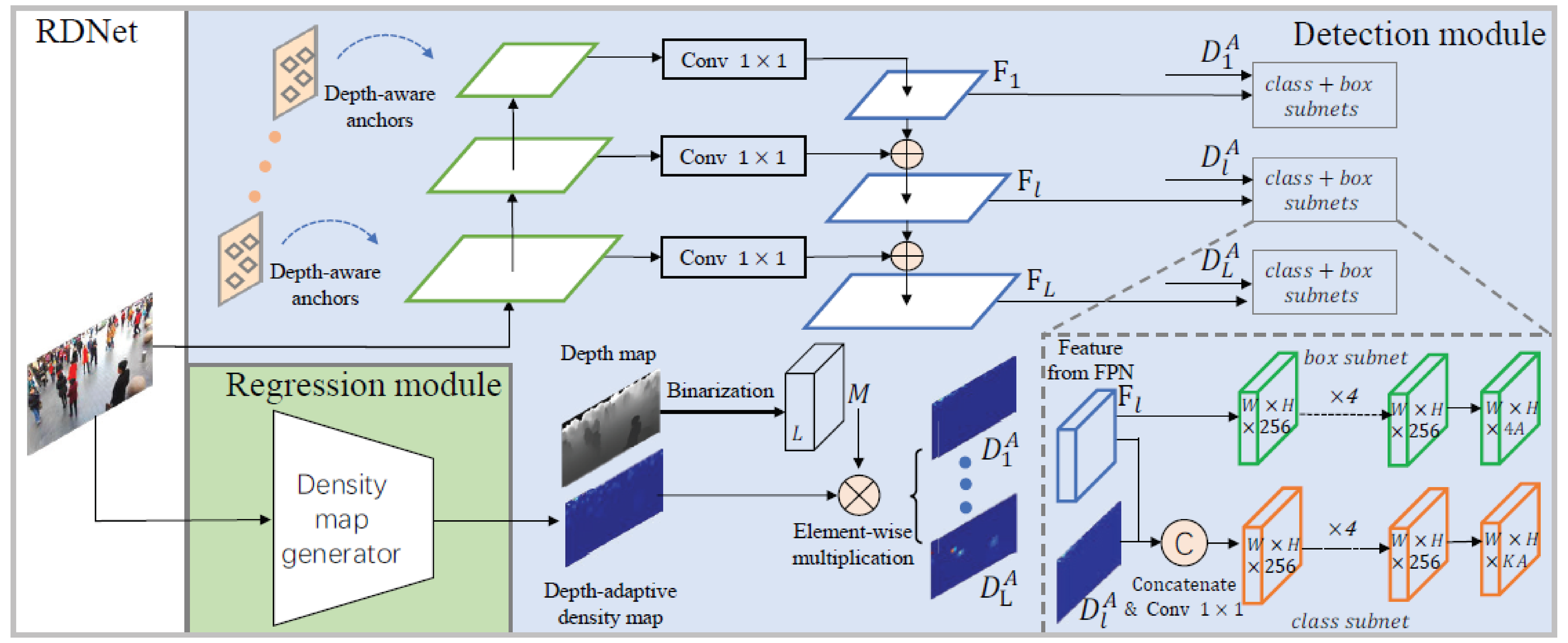
$$\mathcal{D}^A(x) = \sum_{i=1}^{N} \delta(x - x_i) * G_{\sigma(d_i)}(x).$$



➢ Detection module:

- Density map guided classification:
  - Decoding layer l (l=1,…,L), head size $[r_1, r_2]$, corresponding depth $d \in [\frac{\gamma}{r_2}, \frac{\gamma}{r_1}]$
  - Masked density map: $D_l^A = D^A \odot M_l$
- Depth-aware anchor: $H(m,n) = \frac{\gamma}{d(m,n)}$ for anchor initialization.
- Generation of bounding box for training:
  - Bounding boxes: $\mathcal{B} = \{b_1, ..., b_N\}$ for N heads. Head width: $w_i = \frac{\gamma}{d_i}$.

➢ Network architecture:



Dataset:

➢ Our ShanghaiTechRGBD dataset:

- 2,193 images with 144,512 annotated head counts;
- 1,193 images for training.

Table 1. Comparisons of ShanghaiTechRGBD with some existing datasets: **Num** is the number of images; **Max** is the maximal crowd count within one image; **Min** is the minimal crowd count; **Ave** is the average crowd count; **Total** is total number of labeled heads.

Dataset		Resolution	Num	Max	Min	Ave	Total	Modality
CBSR [36]	Dataset 1	240 × 320	2834	7	0	1.6	4,541	Depth
	Dataset 2	240 × 320	1500	7	0	1	1,553	RGB + depth
MICC [1]		480 × 640	3358	11	0	5.32	17,630	RGB + depth
ShanghaiTechRGBD		1080 × 1920	2193	234	6	65.9	144,512	RGB + depth

RGB

Depth

Experiments:

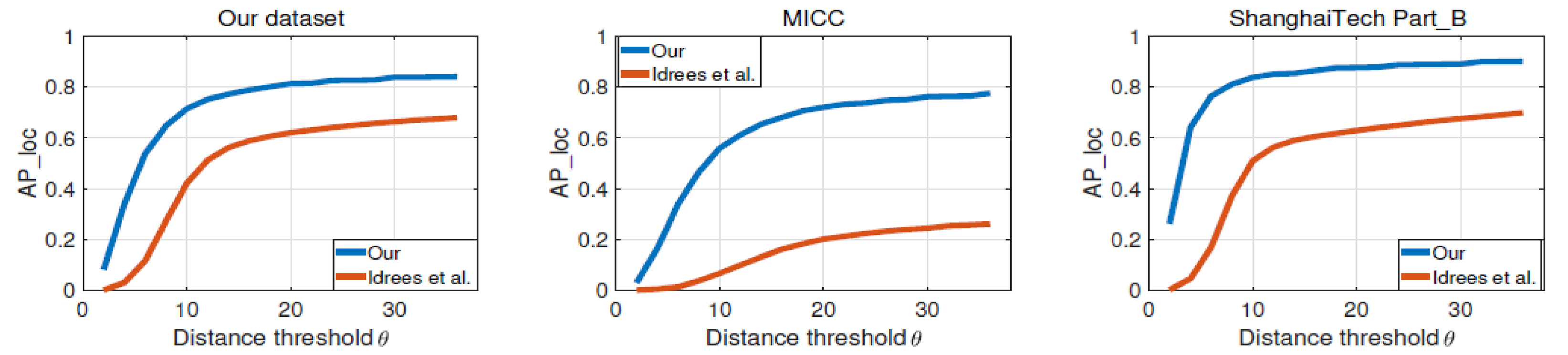
➢ Counting and localization:

- MAE: $\text{MAE} = \frac{1}{\mathcal{M}} \sum_{j=1}^{\mathcal{M}} |N_j - \hat{N}_j|,$
- MSE: $\text{MSE} = \sqrt{\frac{1}{\mathcal{M}} \sum_{j=1}^{\mathcal{M}} (N_j - \hat{N}_j)^2}$
- AP_det for head detection evaluation [2];
- AP_loc for localization evaluation [3].

$$\text{Prediction} = \begin{cases} \text{Positive}, & \text{if dist} \le \theta \\ \text{Negative}, & \text{otherwise} \end{cases}$$

Table 2. Performance evaluations on ShanghaiTechRGBD.

Methods	MAE	MSE	AP_det
MCNN [37]	7.56	10.92	-
MCNN-adaptive	7.14	9.99	-
CSRNet [12]	5.11	7.34	-
CSRNet-adaptive	**4.91**	**7.11**	-
RetinaNet [14]	10.25	14.56	0.356
RetinaNet* [14]	21.84	36.19	0.136
DecideNet (DetNet) [15]	9.74	13.14	0.383
Idrees *et al.* [8]	7.32	10.48	-
RDNet	**4.96**	**7.22**	**0.610**



➢ Visualization:

➢ Reference:

[1] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation, in CVPR 2018.

[2] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. In ICCV, 2017.

[3] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In ECCV 2018.

Code & dataset