ICPR2020

Spatial-related and Scale-aware Network for Crowd Counting

Lei Li[1] , Yuan Dong[1], Hongliang Bai[2]

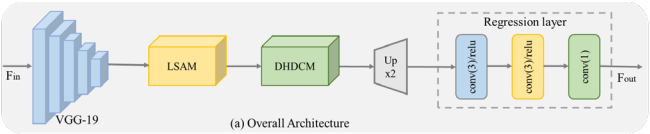
[1]*Beijing University of Posts and Telecommunications*

[2]*Beijing Faceall Co*

I. Introduction

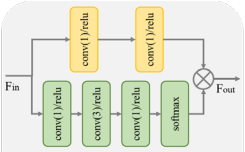
- Crowd counting aims to estimate the number of people in images. The challenge of crowd counting task lies in the variations of scale, different perspective, cluttering, background noise and occlusions. In this paper, we propose a learnable spatial attention module which can get the spatial relations to diminish the negative impact of backgrounds. Besides, a dense hybrid dilated convolution module is also brought up to preserve information derived from varied scales. With these two modules, our network can deal with the problems caused by scale variance and background interference. Our method can achieve state-of-the-art results on three representative crowd counting benchmarks.

II. Methodology

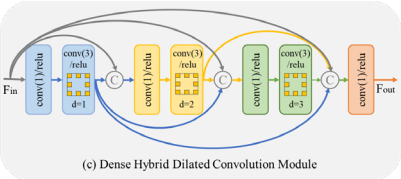


(a) Overall Architecture

- ***Overview of the proposed framework:*** We use a standard image classification network VGG-19 as our backbone, with the last pooling and the subsequent fully connected layers removed. Our LSAM and DHDCM are added after the backbone. Next, we upsample the output after the above modules by two times with bilinear interpolation, and then feed it to a regression head consisting of three convolutional layers to get the density map.



(b) Learnable Spatial Attention Module



(c) Dense Hybrid Dilated Convolution Module

- ***LSAM:*** To get the spatial relations in the feature maps, we use a learnable convolution operation to capture the spatial attentions. The lower branch consists of three convolutional layers and a softmax layer. The convolutional layers are designed to get the spatial attentions.
- ***DHDCM:*** To solve the problem of scale variation, we use dilated convolutional layers to get the most representative features of people across different scales. A module to densely connect a set of dilated convolutional layers, so that the generated multi-scale features can cover a denser scale range without significantly increasing the parameters.

III. Expriments

- ***Results:*** The largely varied heads of different sizes can be captured, background and crowd region can be easily distinguished. Apart from that, the estimated numbers of people is close to the ground truth, proving the robustness and accuracy of our model, not only in congested crowd scenes but also sparse scenes.

- ***Ablation Study of LSAM and DHDCM:*** Simultaneously take the background and scale influence into consideration can respectively bring improvements compared with the baseline method.
- ***Comparison with SOTA methods:*** All the results in the three datasets prove that the proposed LSAM and DHDCM are effective to optimize the performance, not only in dense scenes but also in sparse scenes.

Ablation study of LSAM and DHDCM on UCF-QNRF Benchmark.

LSAM	DHDCM	MAE	MSE
-	-	98.6	172.8
✓	-	96.2	168.5
-	✓	95.4	164.3
✓	✓	**93.2**	**158.2**

Comparison with the state of the Art methods on UCF-QNRF Benchmark.

Method	MAE	MSE
MCNN[1]	277	426
SwitchCNN[6]	228	445
CL[2]	132	191
S-DCNet[28]	104.4	176.1
SFCN[29]	102.0	171.4
Ours	**93.2**	**158.2**

Comparison with the state of the Art methods on ShanghaiTech Benchmark.

Method	PartA		PartB	
	MAE	MSE	MAE	MSE
MCNN[1]	110.2	173.2	26.4	41.3
SwitchCNN[6]	90.4	135	21.6	33.4
IC-CNN[27]	68.5	116.2	10.7	16.0
CSRNet[9]	68.2	115.0	10.6	16.0
BL[26]	64.5	104.0	7.9	13.3
BL+[26]	62.8	101.8	7.7	12.7
ADCrowdNet[12]	63.2	98.9	7.7	12.9
Ours	**60.95**	**97.55**	**7.5**	**12.6**

Comparison with the state of the Art methods on UCF_CC_50 Benchmark.

Method	MAE	MSE
MCNN[1]	377.6	509.1
SwitchCNN[6]	318.1	439.2
CSRNet[9]	266.1	397.5
SANet[7]	258.4	334.9
ADCrowdNet[12]	257.1	363.5
BL+[26]	229.3	308.2
Ours	**220.5**	**302.9**

IV. Conclusions

a. Learnable Spatial Attention Module can get spatial attentions.

b. Dense Hybrid Dilated Convolutional Module can solve the scale variation problem.

c. Both modules can be transferred to other related task.

d. The state-of-the-art results on all three datasets.