Encoder-Decoder Based Convolutional Neural Networks with Multi-Scale-Aware Modules for Crowd Counting

Pongpisit Thanasutives[1], K. Fukui[1], M. Numao[1], B Kijsirikul[2]
[1]Osaka University; [2]Chulalongkorn University

ICPR 2020

Codes: https://github.com/Pongpisit-Thanasutives/Variations-of-SFANet-for-Crowd-Counting

Introduction (1)

(Image from Zhang *et al.*, 2016 [1])

Crowd Counting: To count a number of people in a given image for public safety, surveillance monitoring, etc.

Introduction (2)

Problem (Gao *et al.*, 2019 [2]) :

- Heavy occlusion (noisy image, blurred objects)
- Perspective distortion (different camera angles)
- Scale variation (different sizes of head and surrounding context), etc.

Goal: Solve these problems using a combination of multi-scale-aware modules and dual-path decoder..

(a) Occlusion (b) Complex background (c) Scale variation (d) Non-uniform distribution

(e) Perspective distortion (f) Rotation (g) Illumination variation (h) Weather changes

Introduction (3)

Data preprocessing (based on Zhang *et al.*, 2016 [1]): Convolve the head annotation with Gaussian kernel (G) which has fixed standard deviation (σ). Assuming that there is a head annotation at pixel x_i represented as $\delta(x-x_i)$. The density map D(x) can be defined as

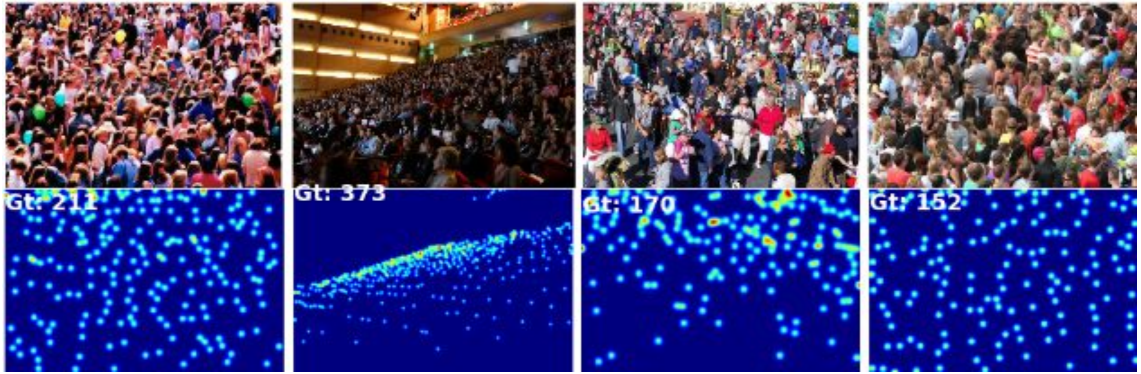
C: Headcounts

$$\boxed{D(x) = \sum_{i=1}^{C} \delta\left(x - x_i\right) * G_\sigma(x)}$$

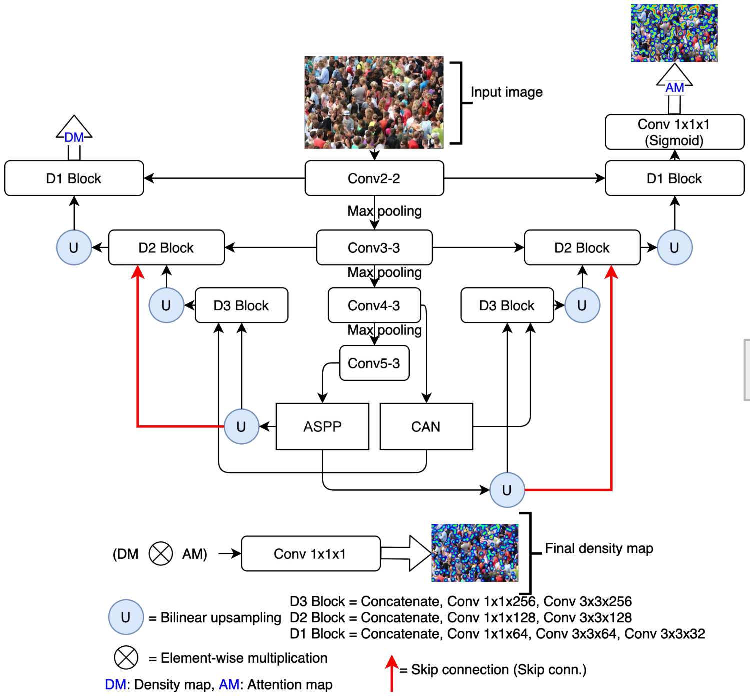
$$\boxed{A(\mathbf{i}) = \begin{cases} 0 & 0.001 > D(\mathbf{i}) \\ 1 & 0.001 \leq D(\mathbf{i}) \end{cases}}$$

A: Attention map

Gaussian convolution



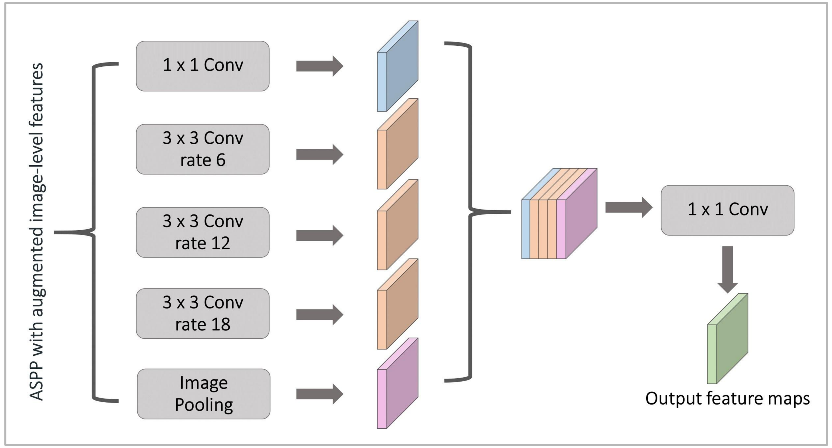
1st Proposed model - M-SFANet (1)



The architecture of M-SFANet

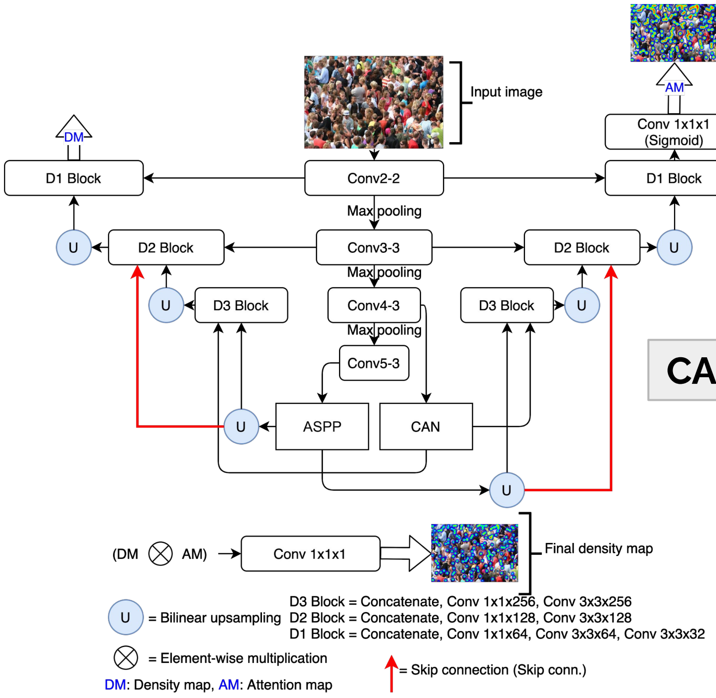
Inspired by SFANet by Zhu *et al.*, 2019 [3].

ASPP

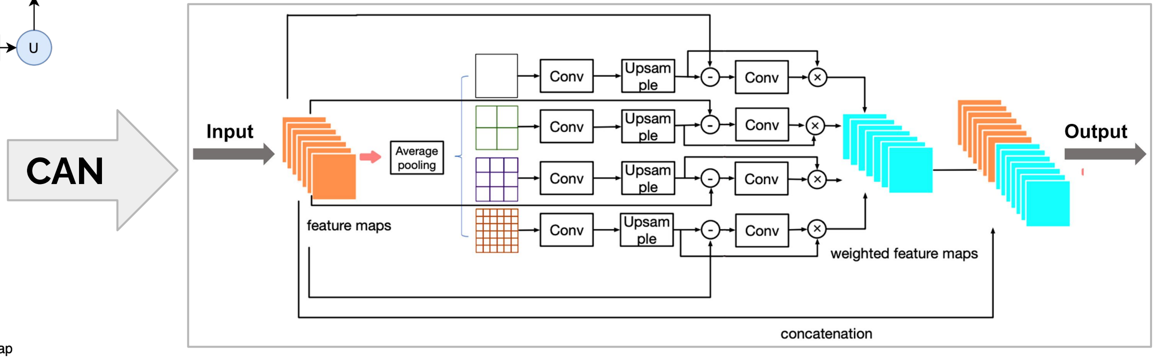


Atrous spatial pyramid pooling (**ASPP**) with augmented image-level features by Chen *et al.*, 2018 [4].

1st Proposed model - M-SFANet (2)

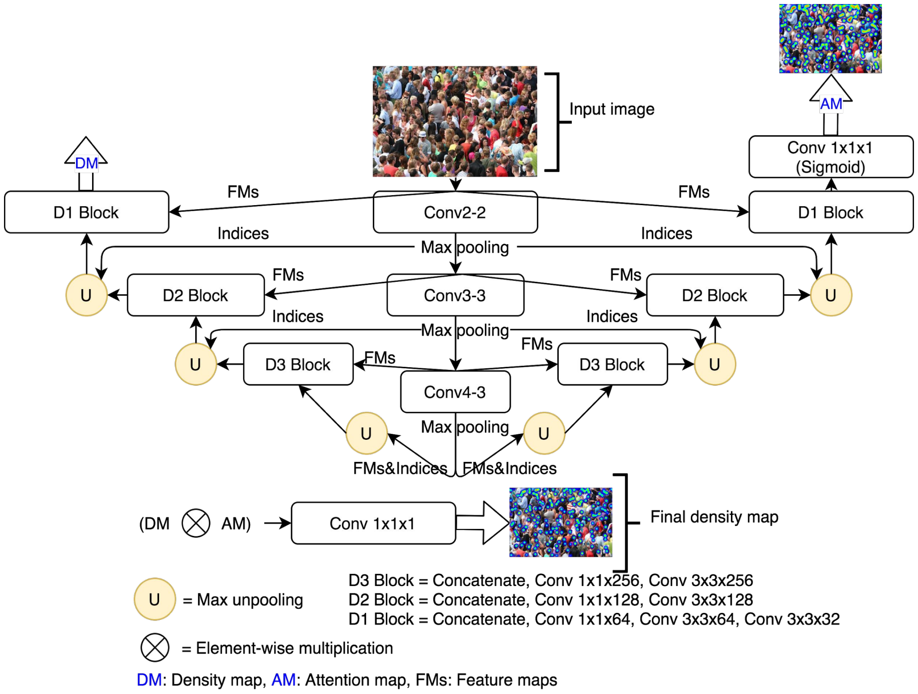


The architecture of M-SFANet



Context-aware module (**CAN**), Liu *et al.*, 2019 [5].

2nd Proposed model - M-SegNet



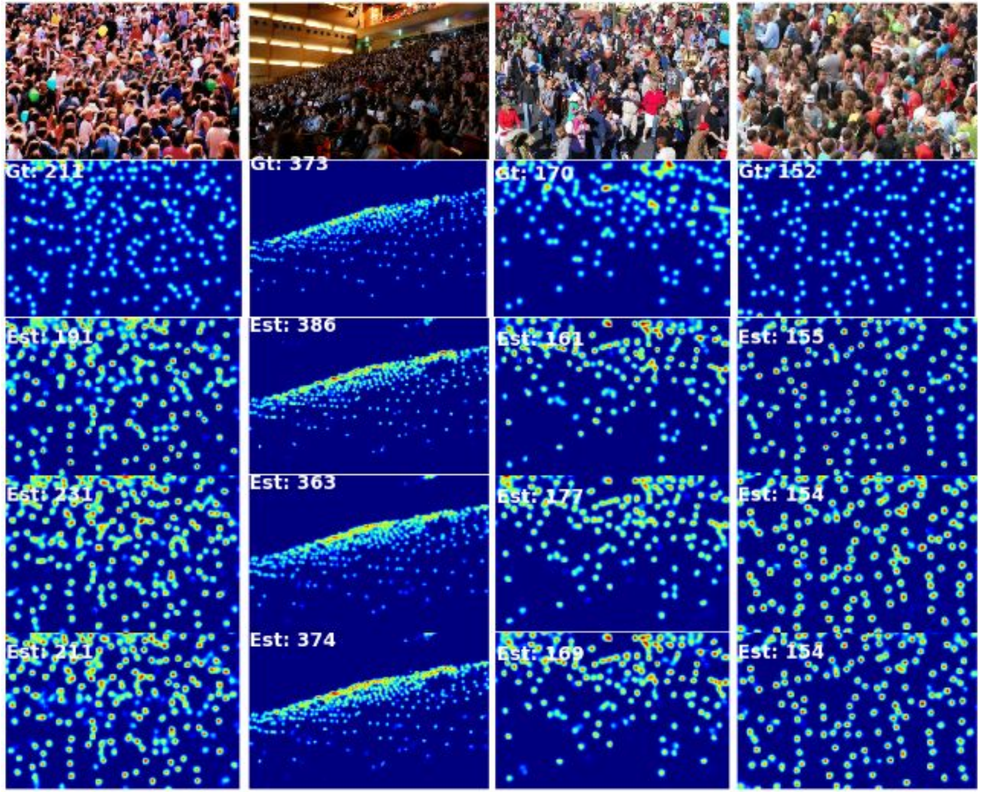
The architecture of M-SegNet

- There are no **CAN** and **ASPP** to additionally emphasize multi-scale information.

- The bilinear upsampling is replaced with max unpooling operation using the memorized max-pooling indices (Badrinarayanan et al., 2017 [6]).

- Less computational resources than M-SFANet with competitive performance. More suitable for speed-constrained applications.

Results (1)

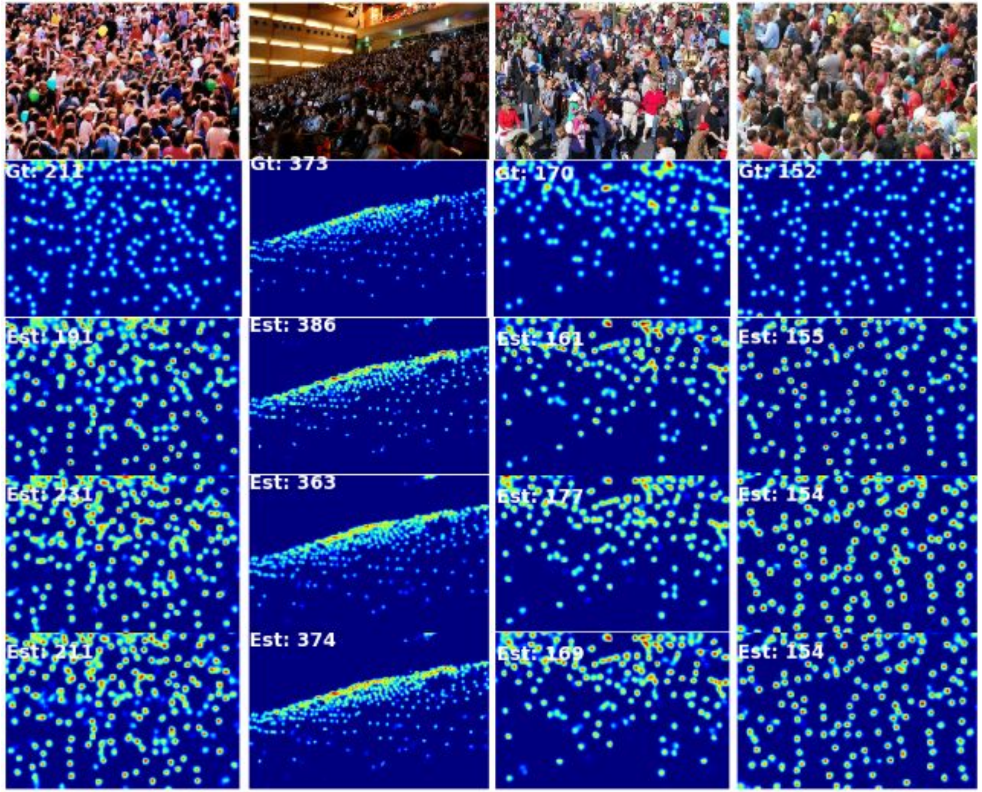


Performance comparison

Method	Part A		Part B		UCF_CC_50	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
CAN	62.3	100.0	7.8	12.2	212.2	**243.7**
SFANet	59.8	99.3	6.9	10.9	219.6	316.2
S-DCNet	58.3	95.0	6.7	10.7	204.2	301.3
SANet + SPANet	59.4	**92.5**	6.5	**9.9**	232.6	311.7
M-SegNet	60.55	100.80	6.80	10.41	188.40	262.21
M-SFANet	59.69	95.66	6.76	11.89	**162.33**	276.76
M-SFANet + M-SegNet	**57.55**	94.48	**6.32**	10.06	167.51	256.26

More results on the paper.

Results (2)



Ablation study

Method	Part A		Part B	
	MAE	RMSE	MAE	RMSE
M-SFANet w/o **CAN**	62..41	101.13	7.40	12.14
M-SFANet w/o **ASPP**	61.25	102.37	7.67	13.28
M-SFANet w/o skip conn.	**60.07**	**99.47**	**7.34**	**12.10**

ASPP: Suitable for sparse scenes.
CAN: Suitable for dense scenes.

Summary

- For M-SFANet, we add the multi-scale-aware modules to SFANet architecture for better tackling drastic scale changes of target objects.

- Furthermore, the decoder structure of M-SFANet is adjusted to have more residual connections in order to ensure that the learned multi-scale features of high-level semantic information will impact how the model regress for the final density map.

- For M-SegNet, we change the up-sampling algorithm from bilinear to max unpooling using the memorized indices employed in SegNet. This yields the cheaper computation model while providing competitive counting performance applicable to real-world applications.

Selected references

[1] Zhang, Y., Zhou, D., Chen, S., Gao, S., & Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 589-597).

[2] Gao, G., Gao, J., Liu, Q., Wang, Q., & Wang, Y. (2020). CNN-based Density Estimation and Crowd Counting: A Survey. *arXiv preprint arXiv:2003.12783*.

[3] Zhu, L., Zhao, Z., Lu, C., Lin, Y., Peng, Y., & Yao, T. (2019). Dual path multi-scale fusion networks with attention for crowd counting. *arXiv preprint arXiv:1902.01115*.

[4] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 801-818).

[5] Liu, W., Salzmann, M., & Fua, P. (2019). Context-aware crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5099-5108).

[6] Badrinarayanan, V., Kendall, A., & Cipolla, R. (2017). Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, *39*(12), 2481-2495.

Thank you for your attention!