Si-yue YU, Jian PU, 2020. Aggregated context network for crowd counting. *Frontiers of Information Technology & Electronic Engineering*, 21(11):1626-1638. https://doi.org/10.1631/FITEE.1900481

Aggregated context network for crowd counting

Key words: Crowd counting; Convolutional neural network; Density estimation; Semantic segmentation; Multi-task learning

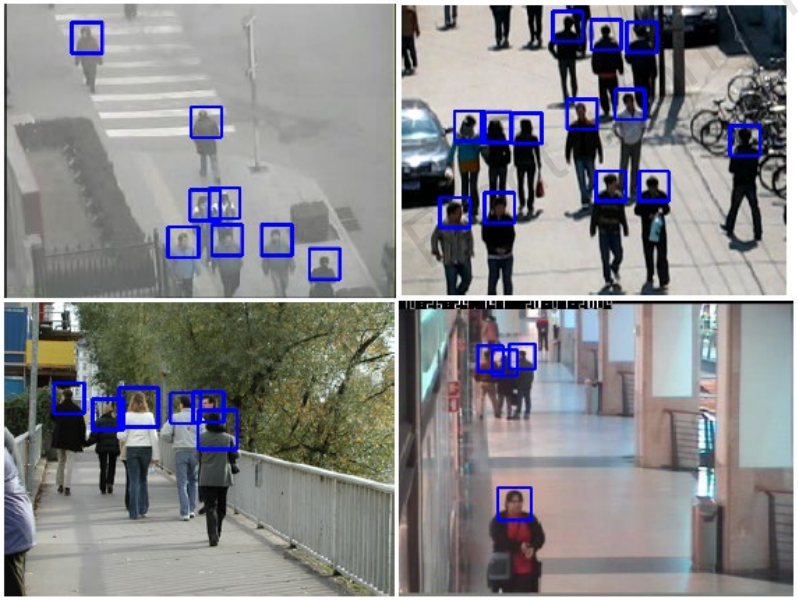
Corresponding author: Jian PU
E-mail: jianpu@fudan.edu.cn
ORCID: https://orcid.org/0000-0002-2949-4273

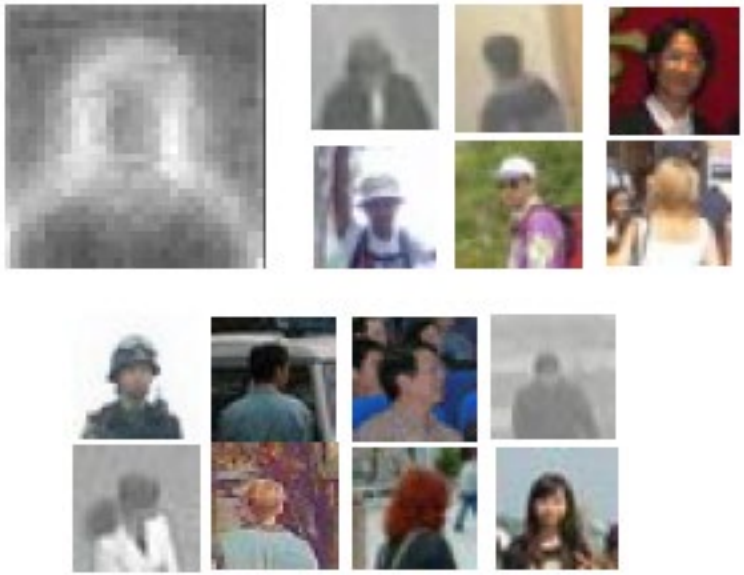
Related works on crowd counting

- **Traditional method**

(1) Detection-based

Training classifiers for the whole body is a straightforward method using low-level features, such as histograms of oriented gradient (HOG). However, most of these methods fail to deal with high-density crowds' images, since targeted pedestrians are severely obscured.



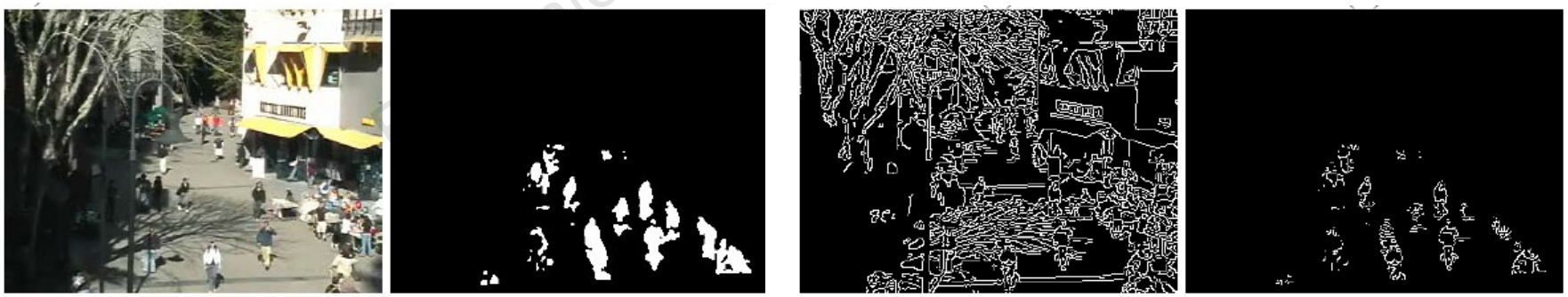


Related works on crowd counting

- **Traditional method**

(2) Regression-based

Features, such as edge features, foreground features, gradient features, and texture, are first independently extracted from images or image patches. Afterward, the relationship between these features and the crowd count are learned by regression. However, this method obtains only the information of the number of people without the spatial distribution of the crowds.



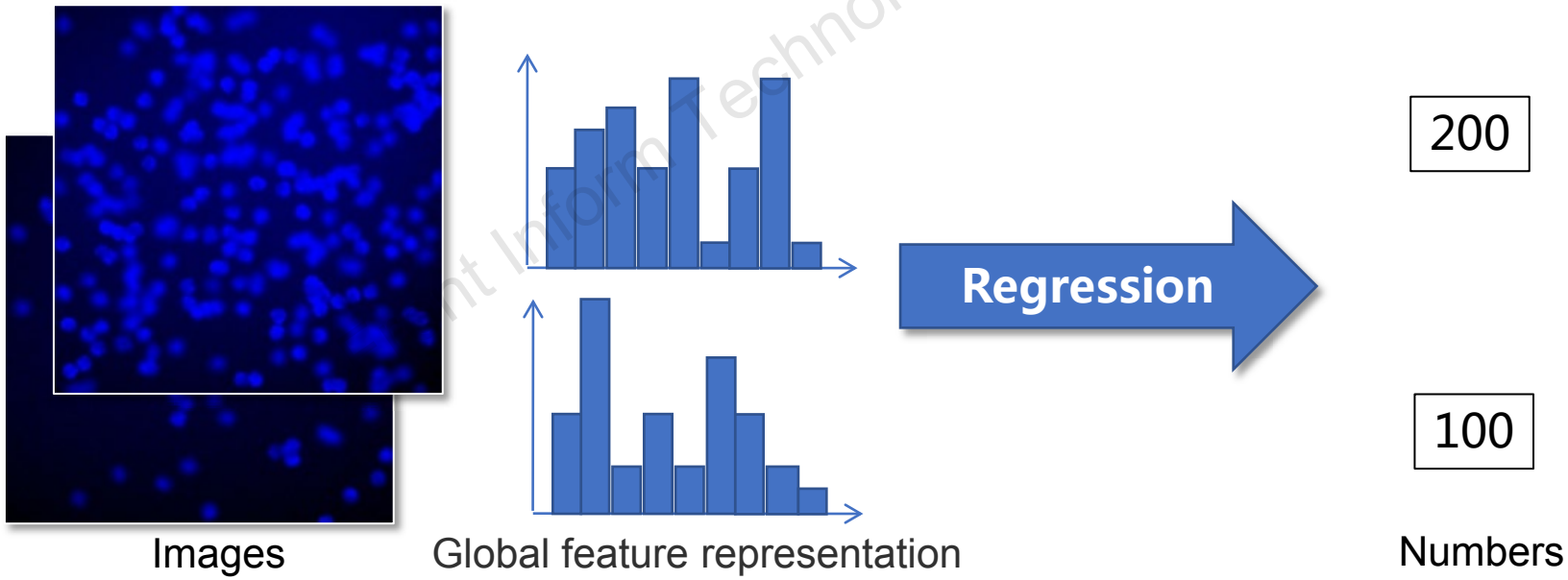
(a) (b) (c) (d)

Related works on crowd counting

- **Traditional method**

(3) Density estimation based

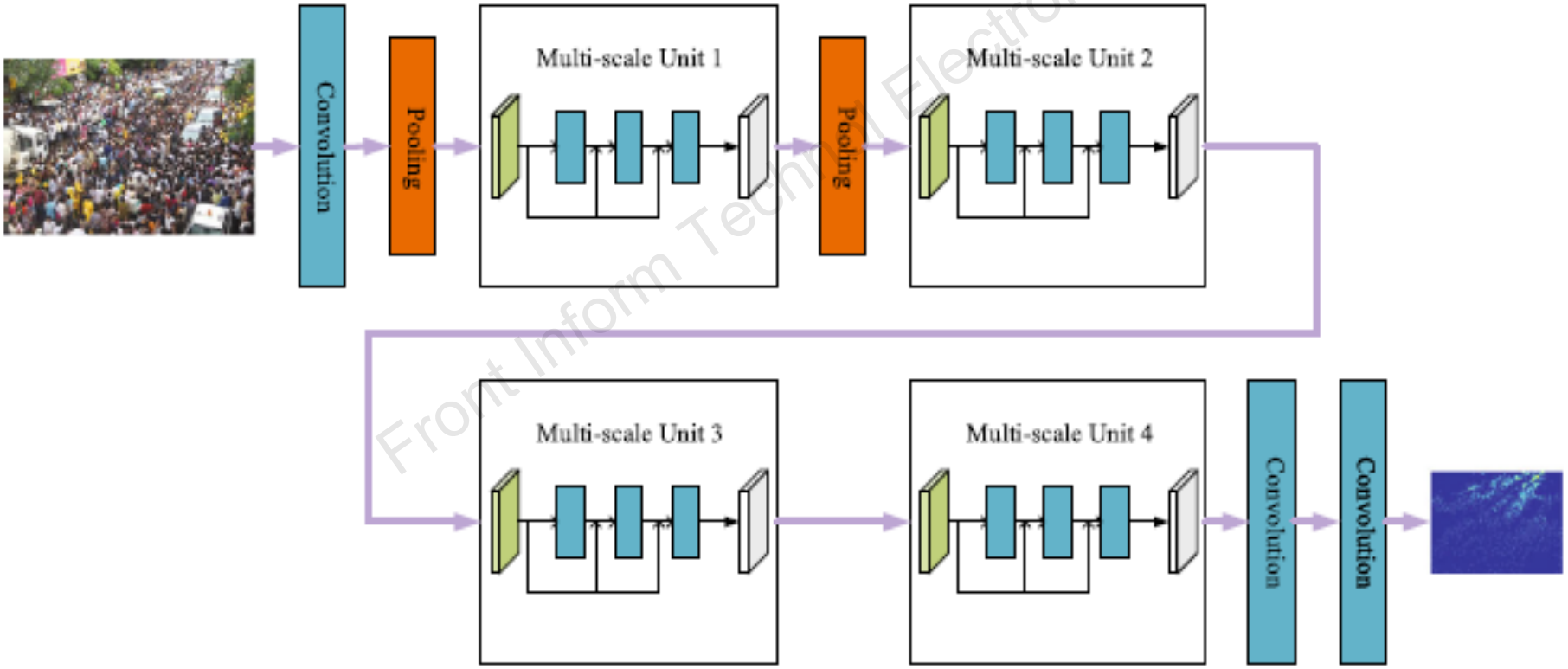
The density estimation based method can preserve more spatial information of crowd scenes.



Related works on crowd counting

- **CNN-based method**

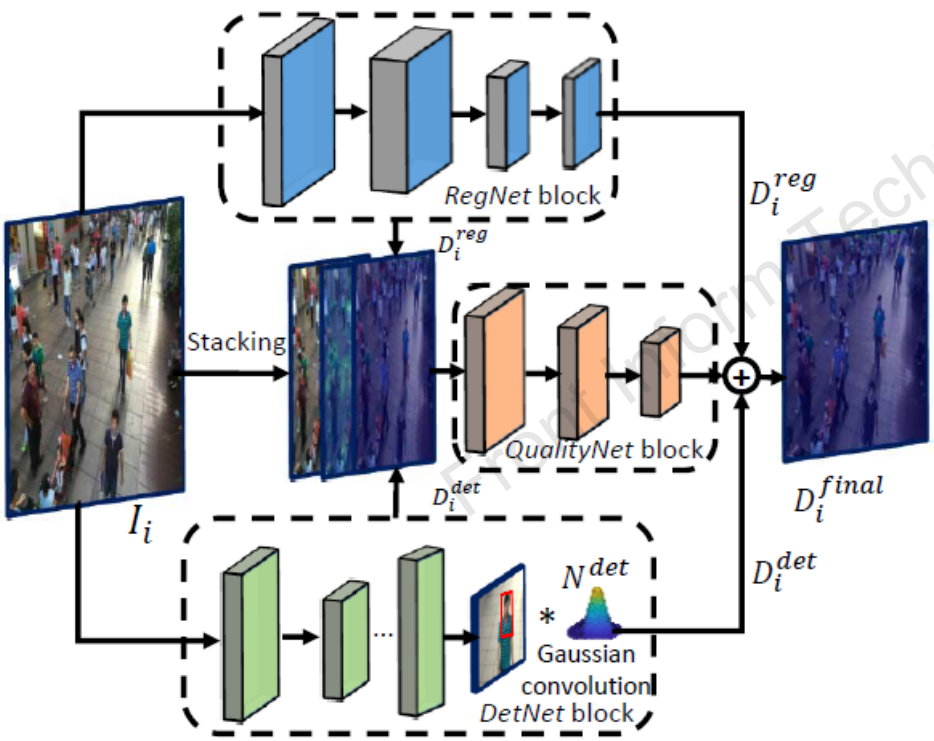
(1) To overcome perspective distortions

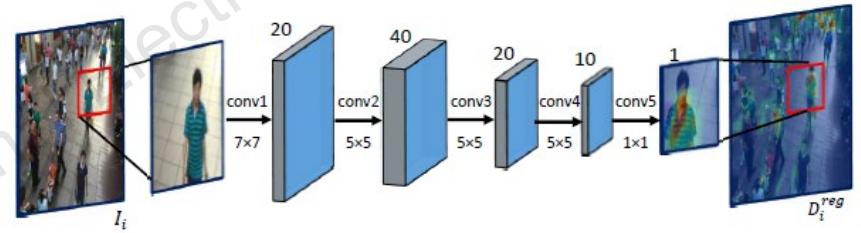


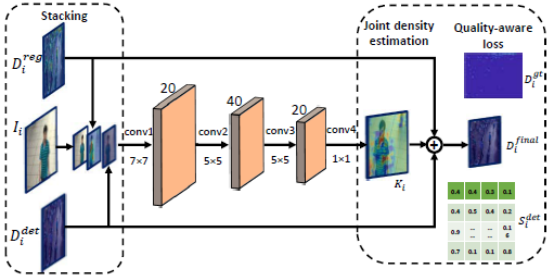
Related works on crowd counting

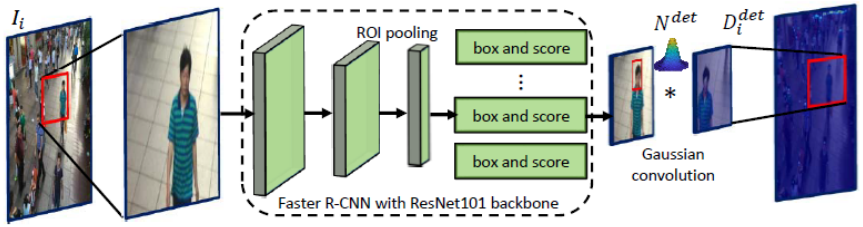
- **CNN-based method**

(2) To overcome background interference





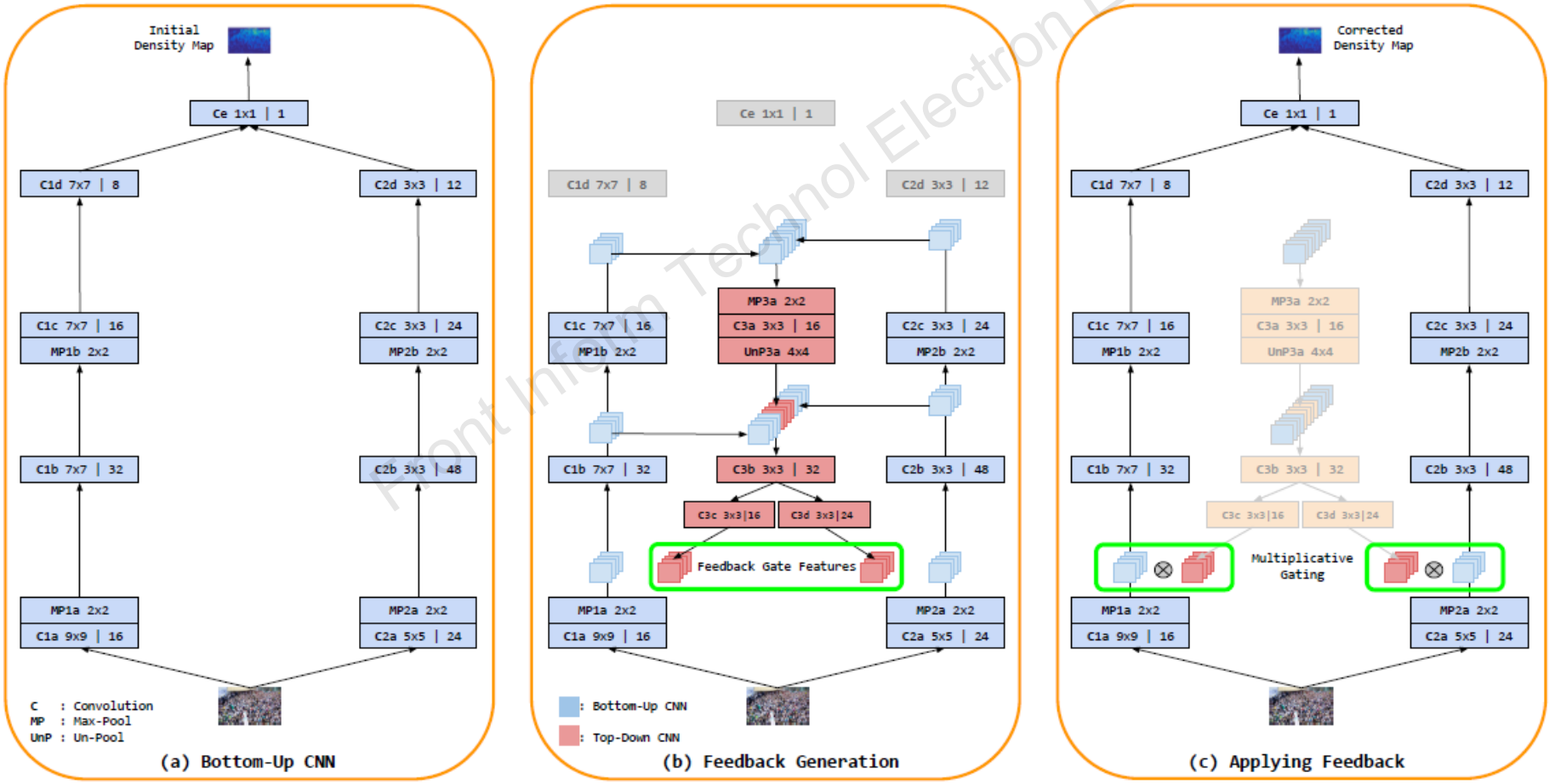




Related works on crowd counting

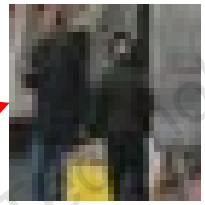
- **CNN-based method**

(3) Crowd counting based on multi-task learning



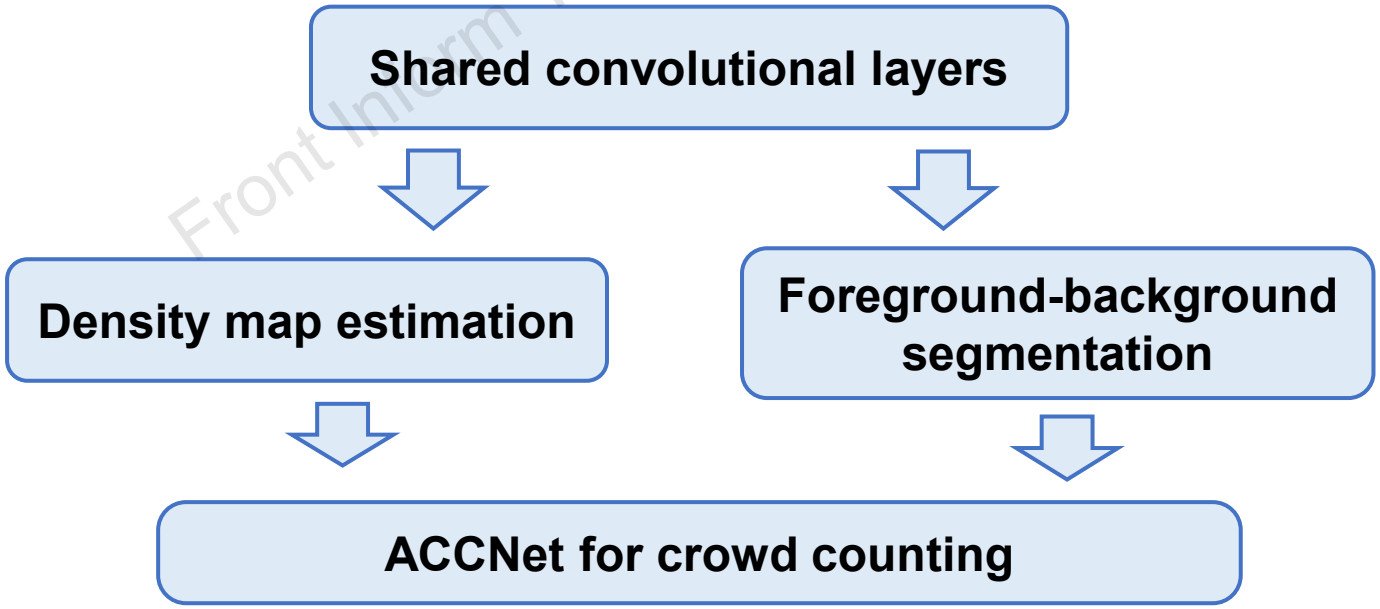
Challenges

- Context information, such as perspective distortion and background interference, is a crucial factor in achieving high performance for crowd counting.



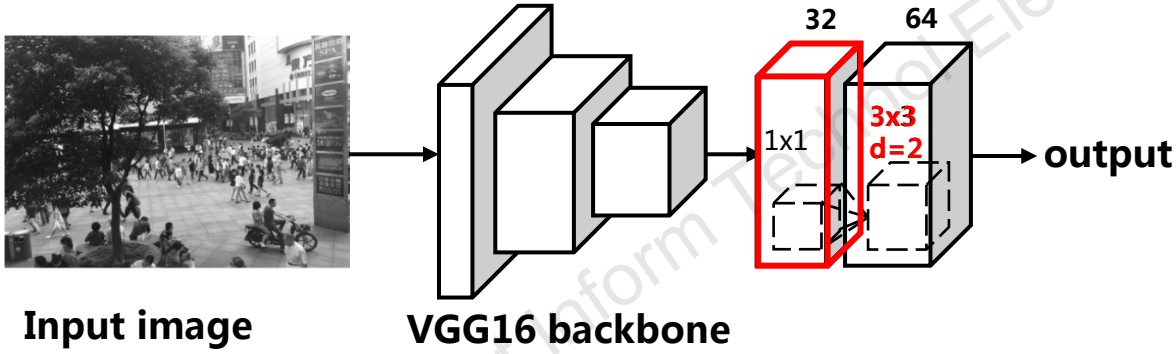
Proposed ACCNet

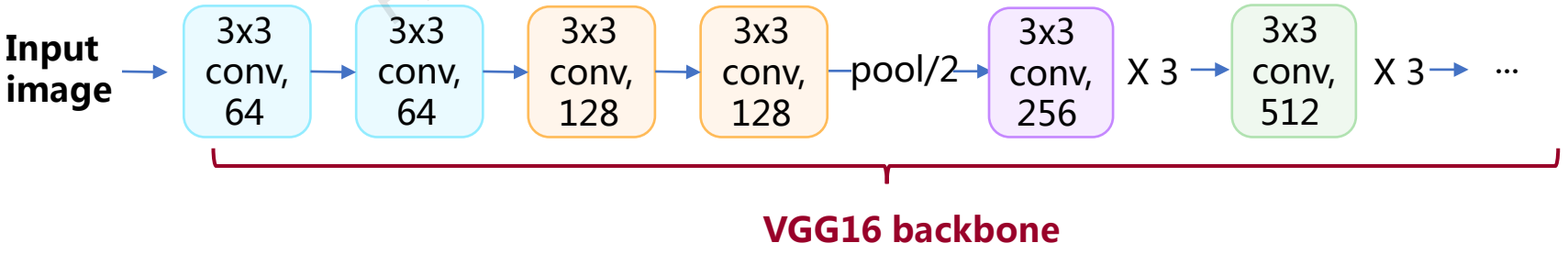
- This study proposed a multi-task network combining density map estimation with semantic segmentation to provide a mutual promotion of crowd counting and semantic segmentation.
- The main task is to extract the multi-scale and spatial context information to learn the density map. The auxiliary semantic segmentation task gives a comprehensive view of the background and foreground information, and the extracted information is finally incorporated into the main task by late fusion.



(1) Shared convolutional layers

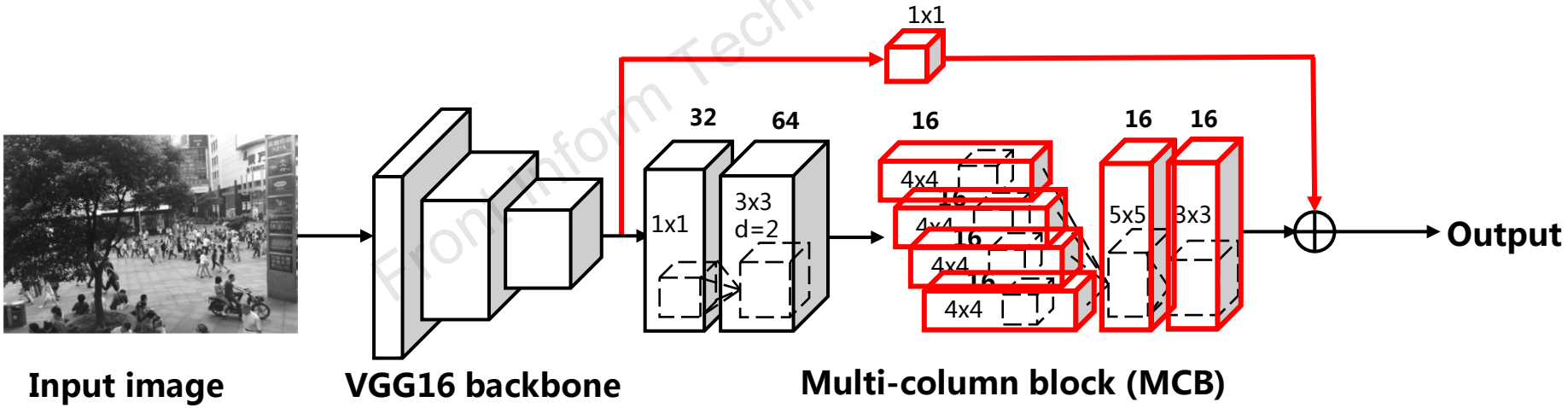
- Use part of VGG-16.
- For three max-pooling layers of the original VGG-16 network, to maintain sufficient invariance and generate high-quality density maps, ACCNet removes the first and third pooling layers and keep only the second max-pooling layer.





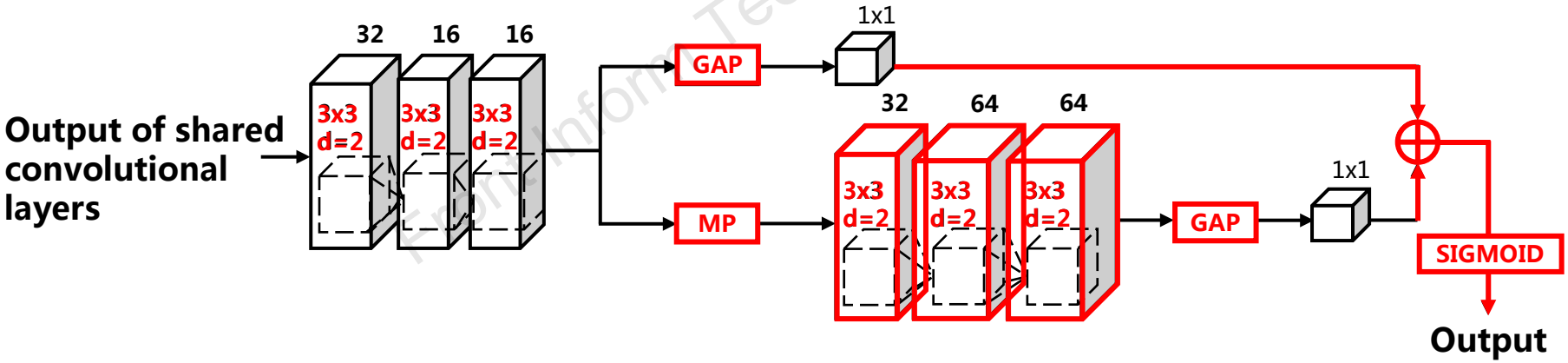
(2) Density map estimation

- To tackle perspective distortion, ACCNet contains the specially designed multi-column block and the skip connection.
- The density map estimation task with the multi-column block and skip connection is to extract spatial information and multi-scale information.

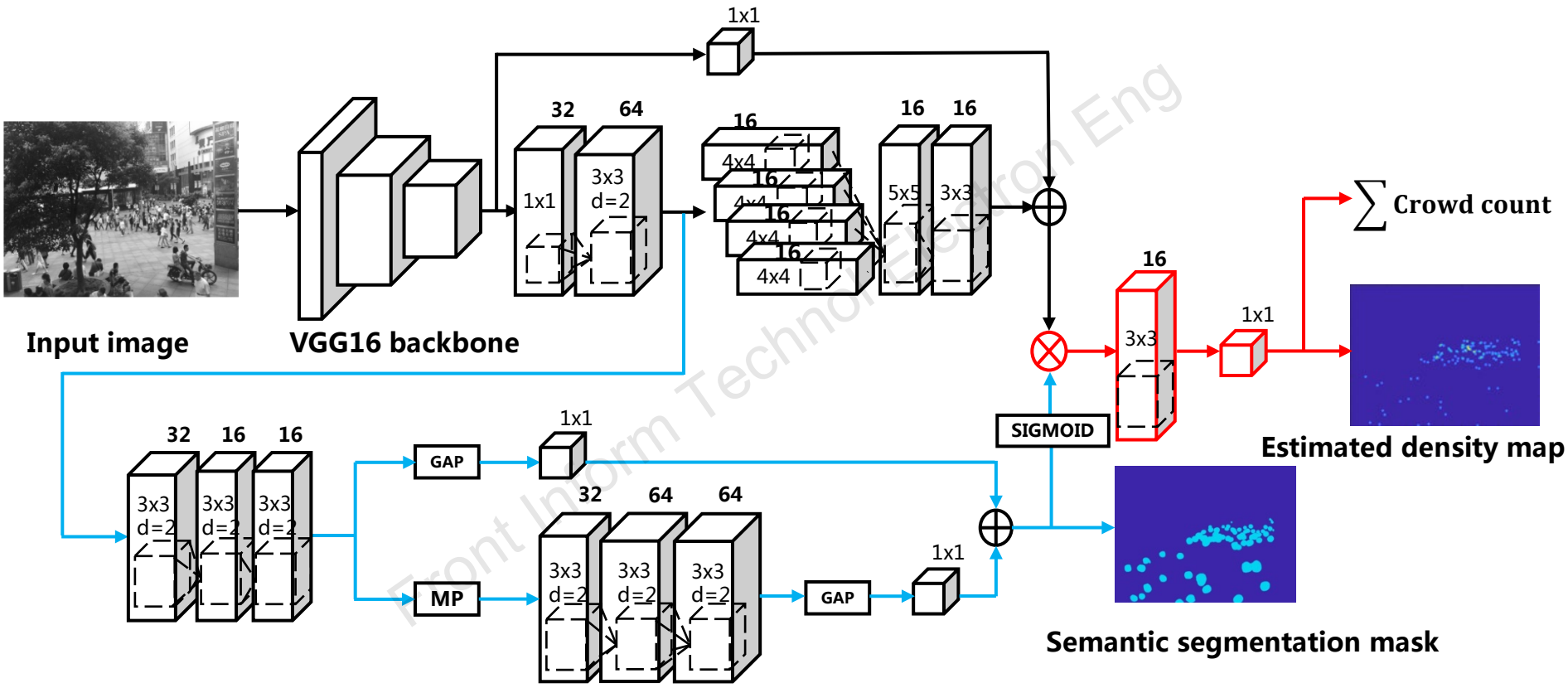


(3) Foreground-background segmentation

- The primary goal of the auxiliary task is to provide global context information to boost the main task of crowd counting.
- The global average pooling and dilated convolutional layers are adopted to provide a comprehensive view of the foreground and background information.



(4) Fuse two tasks for ACCNet



Future outlook

- Enhance the generalization performance of the model, so that the model can handle the richness of the crowd scene well.
- Study real-time statistics of crowd counting in the video to improve the low latency.
- The ground truth generation method for the density map can be improved to better fit the crowd characteristics under the actual scene.
- Considering the network model without annotation or with a small amount of annotations, the time and labor cost of data set annotation can be saved.

References

[1] Li M, Zhang ZX, Huang KQ, et al., 2008. Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection. Int Conf on Pattern Recognition, p.1-4.

[2] Oñoro-Rubio D, López-Sastre RJ, 2016. Towards perspective-free object counting with deep learning. European Conf on Computer Vision, p.615-629.

[3] Li YH, Zhang XFF, Chen DM, 2018. CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. IEEE Conf on Computer Vision and Pattern Recognition, p.1091-1100.

[4] Zhang YY, Zhou DS, Chen SQ, et al., 2016. Single-image crowd counting via multi-column convolutional neural network. IEEE Conf on Computer Vision and Pattern Recognition, p.589-597.

[5] Wang LY, Yin BQ, Guo AX, et al., 2018. Skip-connection convolutional neural network for still image crowd counting. *Appl Intell*, 48:3360-3371.

[6] Sam DB, Babu RV, 2018. Top-down feedback for crowd counting convolutional neural network. AAAI Conf on Artificial Intelligence, p.7323-7330.

[7] Deb D, Ventura J, 2018. An aggregated multicolumn dilated convolution network for perspective-free counting. IEEE Conf on Computer Vision and Pattern Recognition Workshops, p.195-204.

[8] Huang SY, Li X, Zhang ZF, et al., 2018. Body structure aware deep crowd counting. *IEEE Trans Image Process*, 27:1049-1059.

[9] Peng YX, He XT, Zhao JJ, 2018. Object-part attention model for fine-grained image classification. *IEEE Trans Image Process*, 27(3):1487-1500.

[10] Sam DB, Sajjan NN, Babu RV, 2018. Divide and grow: capturing huge diversity in crowd images with incrementally growing CNN. IEEE Conf on Computer Vision and Pattern Recognition, p.3618-3626.

Jian PU received his Ph.D. degree from Fudan University, China, in 2014. He was a postdoctoral researcher at the Institute of Neuroscience, Chinese Academy of Sciences from 2014 to 2016. He was an associate professor at the School of Computer Science and Technology, East China Normal University in China from 2016 to 2019. Currently, he is a professor at the Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, China. He is a corresponding expert of *Frontiers of Information Technology & Electronic Engineering*. His current research interests include machine learning, computer vision, autopilot, and medical image analysis.