

Supplementary Material for

Robust Crowd Counting via Image Enhancement and Dynamic Feature Selection

Nayeong Kim
kimnay@postech.ac.kr

POSTECH
South Korea

Suha Kwak
suha.kwak@postech.ac.kr

This supplementary material provides a more detailed description of E-PFSNet, further analyses on it, and its additional experimental results, all of which are left out from the main paper due to the space limit. Section A details the architecture of Image enhancement network and PFSNet. In particular, we indicate the dimension of the feature map for each layer. Section B serves a detailed discussion about model complexity mentioned in Section 4.2. Section C provides more qualitative results. Visualization of level selection is in Section C.1 and Visualization of enhanced image is in Section C.2. Section D presents failure cases.

A Details of architecture

A.1 Image enhancement network

Detailed architecture of image enhancement network in Fig. A. We indicate the dimension of the feature map for each layer.

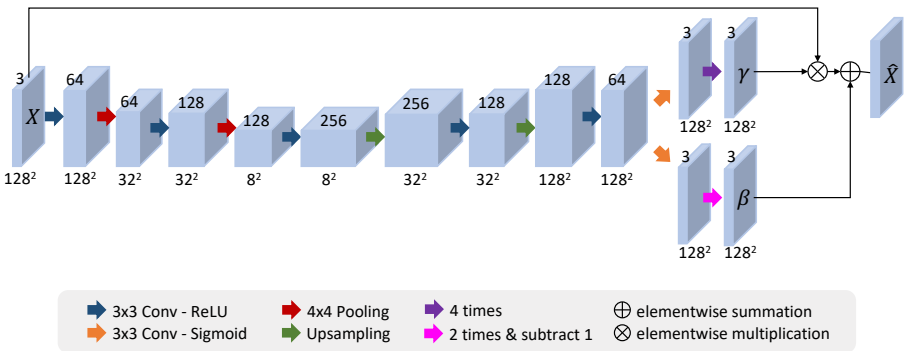


Figure A: Detailed architecture of image enhancement network. Boxes represent feature maps. The spatial dimension of each map is indicated on its below, and its number of channels are indicated above it.

A.2 PFSNet

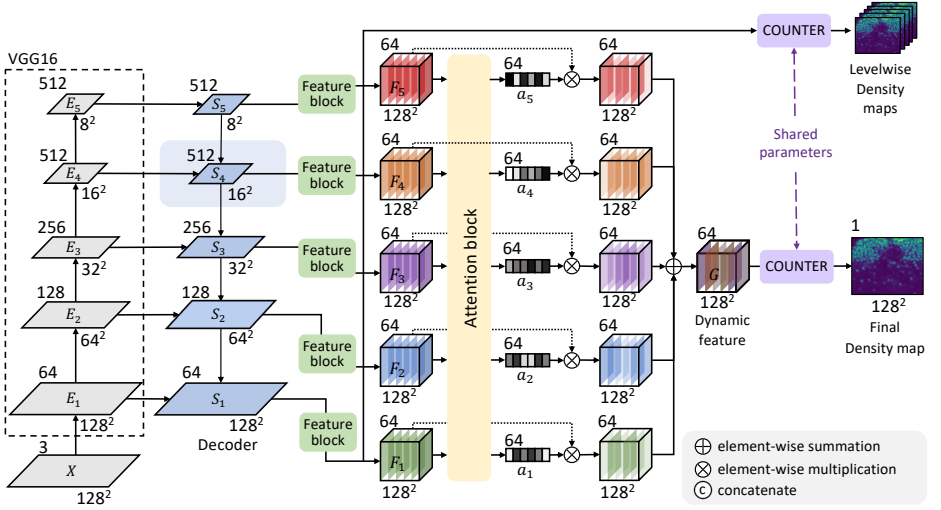


Figure B: Detailed architecture of PFSNet network. Boxes represent feature maps. The spatial dimension of each map is indicated on its below, and its number of channels are indicated above it.

Detailed architecture of PFSNet in Fig. B. We indicate the dimension of the feature map for each layer.

B Model complexity

Method	# of parameters	FLOPS
SASNet [10]	38.90M	451.41G
E-PFSNet	34.78M	457.90G
PFSNet	34.04M	405.11G
Image enhancer	0.74M	52.79G

Table A: Model complexity.

As shown in Table A, we compare FLOPS and the number of parameters with input resolution 589×868 which is the mean resolution of ShanghaiTech PartA. Compared to the previous work based also on FPN [10], our PFSNet is 12.5% and 10.26% lighter in terms of the number of learnable parameters and FLOPS, respectively. Image enhancer is light-weight model. Also, E-PFSNet has fewer learnable parameters than SASNet [10].

C Qualitative results

C.1 Selection

As illustrated in Fig. C, we visualize levels of features that received the most attention in our PFSNet and demonstrates that PFSNet is capable of selecting features according to local pedestrian sizes. The resolution of which increases from S_5 to S_1 . The lower resolution feature level with rich contextual information performs helpful to the prediction of large scale heads while the rich detail information in higher resolution feature level is better for the prediction of small scale heads. In Fig C, the lower resolution feature level tends to be selected in the closer area from the camera and the higher resolution feature level tends to be selected in the far area from the camera, according to perspective. Our E-PFSNet adapts dynamic receptive fields by selecting appropriate features softly according to the sizes and densities of pedestrians.

To visualize the level of feature pixel-wise, we select the level of feature by the following algorithm in each pixel.

- Multiply dynamic feature vector and weight scalar of the counter corresponding to the pixel.
- Select the channel of the vector which has the max value.
- Select the level of feature that received the most attention within the channel.

After selecting the level of feature in every pixel, we mask GT density map to it.

C.2 Enhanced images

Fig. D and Fig. E visualizes enhanced image from the original image which is visualized left side of it. Image enhancer flattens the brightness and contrast of each image while emphasizing the people area. For example, in the Fig. D image enhancer darkens the bright areas because the light source is reflected and brighten relatively dark areas. As illustrated in Fig. E, image enhancer also brightens dark photos taken at late afternoon to night. We find that image enhancer tend to remove shadows.

D Failure cases

In this section, we analyze four worst and second-worst failure cases according to counting errors in ShanghaiTech PartA and ShanghaiTech PartB datasets. In Fig. F(a), the crowd in the center of the image is so far away that it occupies a very small area of the image, but there are a lot of people. As shown in its predicted density map, the counting value in this area is very low. In particular, the feature level selection result shows that a higher resolution feature level should be selected in this area, but on the contrary, a lower resolution feature level is selected. In Fig. F(b), human statues and trees occupy a large part, but as shown in the predicted density map, the model count incorrectly on non-human objects and sky areas. In Fig. F(c), despite the various human head sizes, the model selects the highest resolution feature level as the most appropriate feature in the whole area. In Fig. F(d), the model performs appropriate feature level selection in opposite to other images.

From the fact that all four failure cases predict lower count than GT count and fail to feature level selection, we demonstrate that our dynamic feature selection tries to detect

the people in the crowded but occupied small image area which is a hard problem in crowd counting task. To fully perform crowd counting, research to solve this problem is still needed in the future.

References

- [1] Qingyu Song, Changan Wang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Jian Wu, and Jiayi Ma. To choose or to fuse? scale selection for crowd counting. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2021.



Figure C: Visualization of levels of features that received the most attention in PFSNet. These examples suggest that PFSNet is capable of adaptively select appropriate features according to sizes and densities of pedestrians.

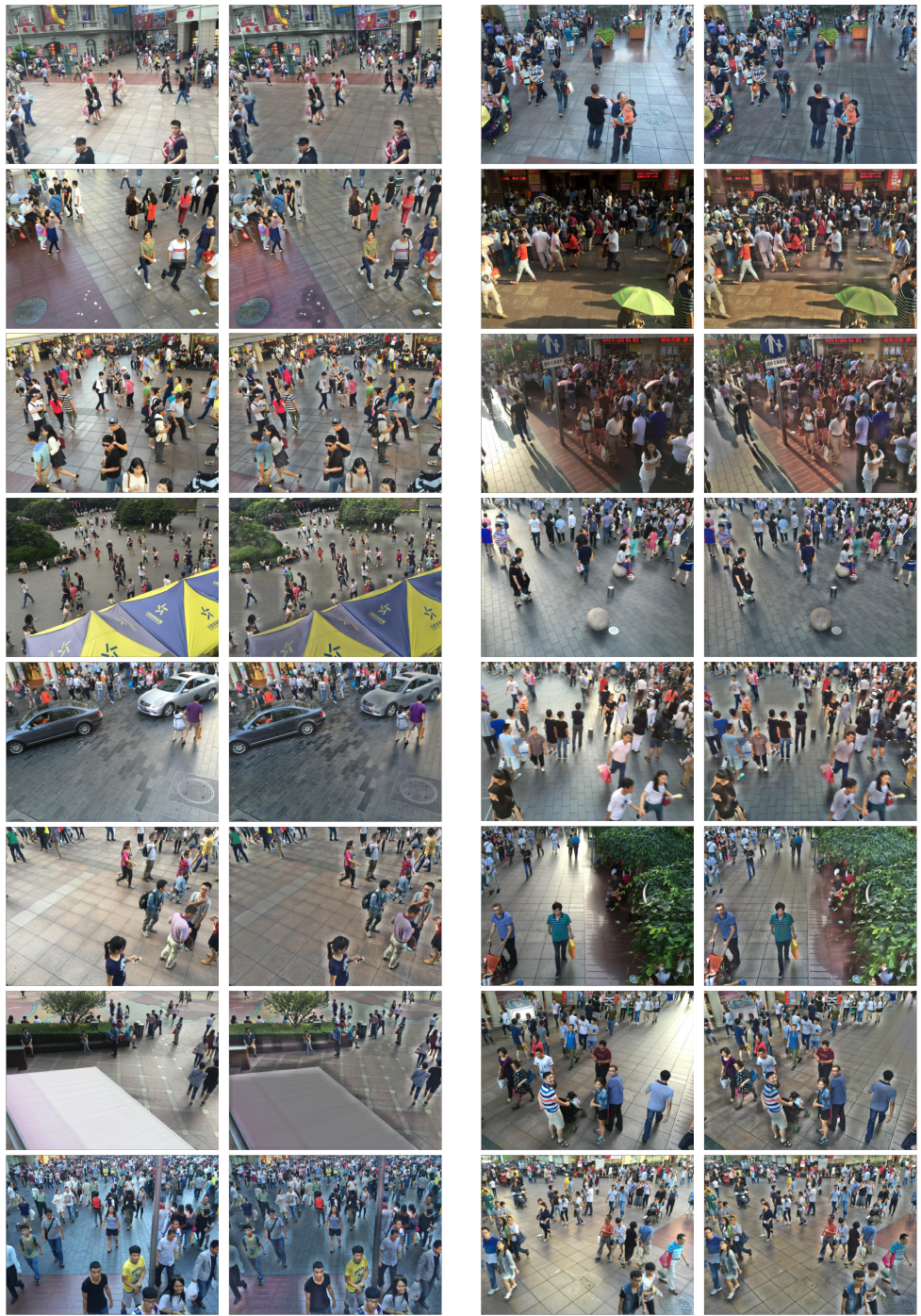


Figure D: Visualization of enhanced image from the original image which is visualized left side of it. These examples are taken during the day, so these are heavily affected by sunlight.



Figure E: Visualization of enhanced image from the original image which is visualized left side of it. These examples are taken at night or late afternoon.

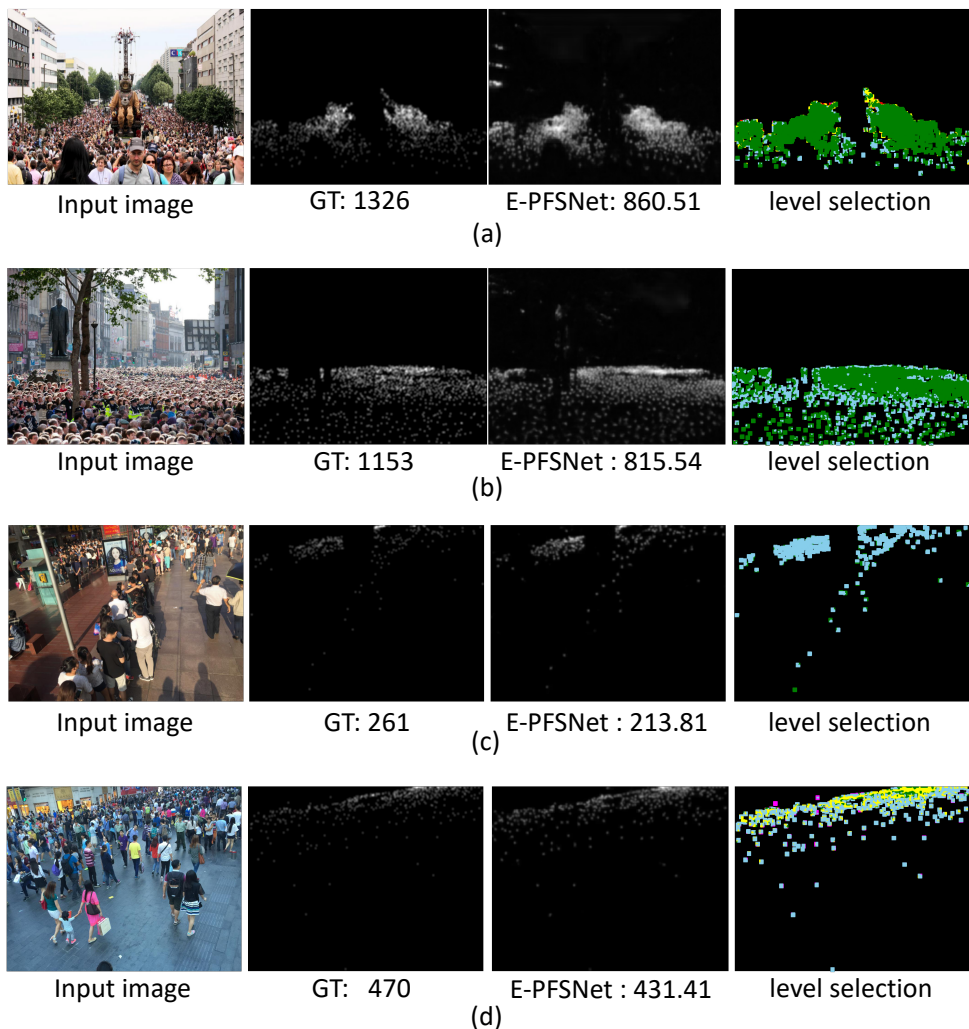


Figure F: Failure cases. (a) and (b) are worst and second worst case in ShanghaiTech PartA, respectively. (c) and (d) are worst and second worst case in ShanghaiTech PartB, respectively.