# Supplementary: Learning from Synthetic Data for Crowd Counting in the Wild

This file provides some additional information from three perspective: dataset, supervised and domain adaptation methods, which correspond to the Section 3, 4 and 5 in the paper.

## 1. GCC Dataset

### 1.1. Exemplars of GCC Dataset

For a deeper understanding GCC dataset, some typical crowd scenes are shown in Fig. 1.

Figure 1. The exemplars of synthetic crowd scenes from the proposed GCC dataset.

### 1.2. Information Provided by GCC

For each scene, the complete camera parameters in the virtual world are provided: position coordinates, height, pitch/yaw angle and field of view. In addition, we also provide the Region of Interest (ROI) for placing person models, which is represented by a polygon region. According to the area of ROI, we assign a capacity label from 9 levels for each scene. Based on aforementioned parameters, all scenes in GCC dataset can be easily reproduced.

For one specific crowd image, in addition to coordinates of head locations, we also provide its capturing time in 24h, weather condition and binary crowd segmentation map.

### 1.3. 100 Locations in GTA5 World

Fig. 2 demonstrates the position of each location in GTA5 world. In general, our locations are mainly concentrated in the urban area.

Figure 2. The demonstration of selected 100 locations in GTA5 world.

## 2. Supervised Crowd Counting

### 2.1. Configuration Details of the Proposed Networks in this Paper

Table 1 explains the configurations of FCN, SFCN and SFCN†. In the table, "k(3,3)-c256-s1-d2" represents the convolutional operation with kernel size of $3 \times 3$, 256 output channels, stride size of 1 and dilation rate of 2. Note that we modify the stride size to 1 in conv4_x of ResNet-101 backbone, which makes conv4_x output the feature maps with 1/8 size of the input image. Other architecture settings fully follow the original VGG-16 and ResNet-101.

Table 1. The network architectures of FCN, SFCN and SFCN†.

| FCN | SFCN | SFCN† |
|---|---|---|
| **VGG-16 backbone** | | **ResNet-101 backbone** |
| conv1: [k(3,3)-c64-s1] $\times$ 2 | | conv1: k(7,7)-c64-s2 |
| ... | | ... |
| conv3: [k(3,3)-c512-s1] $\times$ 3 | | conv4_x: $\begin{bmatrix} k(1,1)-c256-s1 \\ k(3,3)-c256-s1 \\ k(1,1)-c1024-s1 \end{bmatrix} \times 23$ |
| - | **Dilation Convolution** | |
| - | k(3,3)-c512-s1-d2 | |
| - | k(3,3)-c512-s1-d2 | |
| - | k(3,3)-c512-s1-d2 | |
| - | k(3,3)-c256-s1-d2 | |
| - | k(3,3)-c128-s1-d2 | |
| - | k(3,3)-c64-s1-d2 | |
| - | **Spatial Encoder** | |
| - | down: k(1,9)-c64-s1 | |
| - | up: k(1,9)-c64-s1 | |
| - | left-to-right: k(9,1)-c64-s1 | |
| - | right-to-left: k(9,1)-c64-s1 | |
| **Regression Layer** | | |
| k(1,1)-c1-s1 | | |
| upsample layer: $\times 8$ | | |

### 2.2. Performance of SFCN on GCC

Table 2 lists the results on GCC dataset. The models are evaluated using standard Mean Absolute Error (MAE) and Mean Squared Error. In the table, "Average" denotes the average value of each class.

Table 2. Results of SFCN on GCC dataset (MAE/MSE).

Performance of SFCN in each class

| Method | **Average** | 0∼10 | 0∼25 | 0∼50 | 0∼100 | 0∼300 | 0∼600 | 0∼1k | 0∼2k | 0∼4k |
|---|---|---|---|---|---|---|---|---|---|---|
| random | **28.7/46.2** | 6.5/8.8 | 8.5/14.2 | 6.8/10.2 | 5.7/8.7 | 11.5/16.1 | 20.8/27.9 | 32.9/46.9 | 52.1/91.5 | 113.8/191.8 |
| cross-camera | **47.0/73.0** | 13.7/23.3 | 14.7/18.3 | 10.3/13.6 | 11.1/14.0 | 17.6/27.5 | 21.8/29.1 | 57.3/73.4 | 96.2/165.0 | 180.6/293.3 |
| cross-location | **58.4/87.2** | 4.7/4.9 | 7.8/13.5 | 11.0/13.2 | 11.4/13.3 | 17.2/24.5 | 20.9/28.3 | 18.6/26.3 | 138.3/232.3 | 295.8/428.6 |

Performance of SFCN at different time periods

| Method | **Average** | 0∼3 | 3∼6 | 6∼9 | 9∼12 | 12∼15 | 15∼18 | 18∼21 | 21∼24 |
|---|---|---|---|---|---|---|---|---|---|
| random | **41.4/96.7** | 54.5/110.4 | 49.5/135.5 | 29.1/72.2 | 29.6/76.5 | 33.4/64.2 | 34.2/80.2 | 47.2/87.7 | 54.1/146.7 |
| cross-camera | **63.7/147.4** | 77.9/192.0 | 72.1/222.8 | 52.6/113.9 | 41.6/101.8 | 70.7/144.0 | 54.8/136.2 | 78.5/147.9 | 60.9/121.1 |
| cross-location | **97.8/228.4** | 104.7/216.4 | 138.8/308.2 | 62.6/164.7 | 81.3/209.8 | 77.8/174.7 | 94.7/235.9 | 122.6/250.2 | 100.1/267.2 |

Performance of SFCN under different weathers

| Method | **Average** | Clear | Clouds | Rain | Foggy | Thunder | Overcast | Extra Sunny |
|---|---|---|---|---|---|---|---|---|
| random | **40.8/92.5** | 35.1/84.0 | 36.0/64.8 | 43.7/83.4 | 58.2/167.6 | 45.2/86.8 | 34.2/84.9 | 33.5/76.3 |
| cross-camera | **68.3/155.6** | 54.4/130.9 | 62.5/122.5 | 87.8/208.3 | 70.2/163.2 | 73.4/163.1 | 71.1/172.1 | 58.5/129.1 |
| cross-location | **106.8/246.2** | 76.1/185.2 | 88.7/196.0 | 128.2/286.8 | 160.2/413.1 | 117.7/232.8 | 84.8/193.2 | 92.1/216.4 |

From the performance of the three aspects (random, cross-camera and cross-location splitting), both MAE and MSE are increased, which means the difficulty of three tasks is rising in turn. From the first table, the performance of small scenes is

better than that of large scenes. The main reason is: the count ranges of the latter are far greater than that of the former, which causes that the former's errors become larger. The second table shows that the daytime scenes are easier to count the number of people than the night scenes. Similarly, from the third table, we also find the clear, cloud, overcast and extra sunny scenes are easier than the rain, foggy and thunder scenes.

# 3. Crowd Counting via Domain Adaptation

## 3.1. Scene Regularization in Domain Adaptation

In the paper, we introduce Scene Regularization (SR) to select the proper images to avoid negative adaptation. This is not an elaborate selection but a coarse data filter. Here, Table 3 shows the concrete filter condition for adaptation to the five real datasets.

Table 3. Filter condition on five real datasets.

| Target Dataset | level | time | weather | count range | ratio range |
|---|---|---|---|---|---|
| SHT A | 4,5,6,7,8 | 6:00∼19:59 | 0,1,3,5,6 | 25∼4000 | 0.5∼1 |
| SHT B | 1,2,3,4,5 | 6:00∼19:59 | 0,1,5,6 | 10∼600 | 0.3∼1 |
| UCF_CC_50 | 5,6,7,8 | 8:00∼17:59 | 0,1,5,6 | 400∼4000 | 0.6∼1 |
| UCF-QNRF | 4,5,6,7,8 | 5:00∼20:59 | 0,1,5,6 | 400∼4000 | 0.6∼1 |
| WorldExpo'10 | 2,3,4,5,6 | 6:00∼18:59 | 0,1,5,6 | 0∼1000 | 0∼1 |

In Table 3, ratio range means that the numbers of people in selected images should be in a specific range. For example, during adaptation to SHT A, there is a candidate image with level 0∼4000, containing 800 people. According to the ratio range of 0.5∼1, since 800 is not in 2000∼4000 (namely 0.5*4000 ∼1*4000), the image can not be selected. In other words, ratio range is a restriction in terms of congestion.

Other explanations of Arabic numerals in the table is listed as follows:

**Level Categories** 0: 0∼10, 1: 0∼25, 2: 0∼50, 3: 0∼100, 4: 0∼300, 5: 0∼600, 6: 0∼1k, 7: 0∼2k and 8: 0∼4k.

**Weather Categories** 0: clear, 1: clouds, 2: rain, 3: foggy, 4: thunder, 5: overcast and 6: extra sunny.

## 3.2. Visualization Comparison of Cycle GAN and SE Cycle GAN

Fig. 3, 4 and 5 demonstrate the translated images from GCC to the five real-world datasets. "Src" and "Tgt" represent the source domain (synthetic data) and target domain (real-world data). The top column shows the results of the original Cycle GAN and the bottom is the results of the proposed SE Cycle GAN.

We compare some obvious differences between Cycle GAN and SE Cycle GAN (ours) and mark them up with rectangular boxes. To be specific, ours can produce more consistent image than the original Cycle GAN in the green boxes. As for the red boxes, Cycle GAN loses more texture features than ours. For the purple boxes, we find that Cycle GAN produces some abnormal color values, but SE Cycle GAN performs better than it. For the regions covered by blue boxes, SE Cycle GAN maintains the contrast of the original image than Cycle GAN in a even better fashion.

In general, from a visualization results, the proposed SE Cycle GAN generates more high-quality crowd scenes than the original Cycle GAN.


Figure 3. The exemplars of translated images.


Figure 4. The exemplars of translated images.

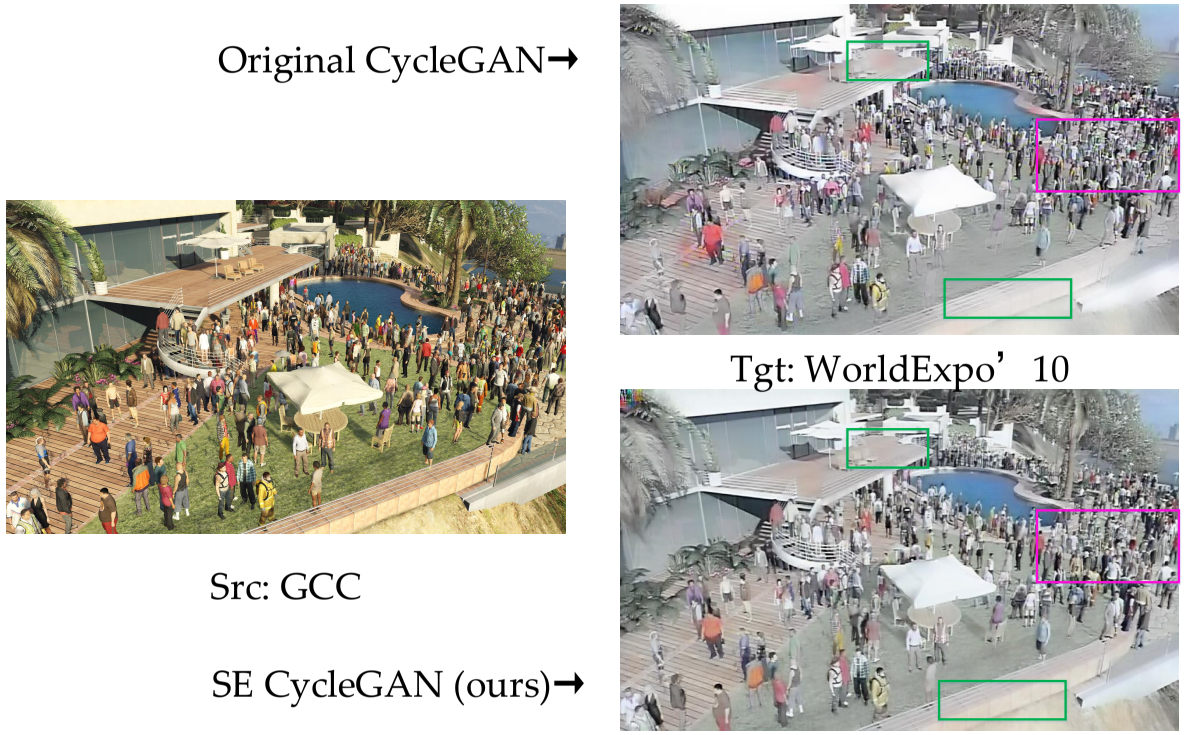


Figure 5. The exemplars of translated images.