Supplementary material: Learning To Count Everything

Viresh Ranjan[1] Udbhav Sharma[1] Thu Nguyen[2] Minh Hoai[1,2]
[1]Stony Brook University, USA
[2]VinAI Research, Hanoi, Vietnam

Adaptation Loss	MAE	RMSE
No Adaptation	24.32	70.94
Min-Count	24.01	70.25
Perturbation	23.90	69.22
Min-Count + Perturbation	23.75	69.07

Table 1: Evaluating the usefulness of the Test Time Adaptation on Validation set of FSC147. No Adaptation refers to the FamNet version trained on the training set without any test time adaptation. As expected, this results in the worst performance. Using either Min-Count or Perturbation loss for adaptation leads to better results than the No Adaptation case. Using both Perturbation and Min-Count losses leads to the best results.

1. Overview

In this Supplementary submission, we first present ablation study on the usefulness of the Min-Count and Perturbation losses proposed for the test time adaptation in Sec. 2. Next, we provide details related to the FamNet architecture in Sec. 3. Additional images from the FSC-147 dataset are shown in Sec. 4. We present additional qualitative results on images from the validation and test splits of FSC-147 dataset in Sec. 5.

2. Ablation Study on the Test Time Adaptation

In Table 1, we analyze the usefulness of the Min-Count and Perturbation losses for the test time adaptation. We train the FamNet on our train set and evaluate it on our validation set. No Adaptation refers to the FamNet version trained on the training set without any test time adaptation. As expected, not doing any test time adaptation leads to the worse results. Using either Min-Count or Perturbation loss for adaptation leads to improved results. Using both Perturbation and Min-Count losses leads to the best results.

3. Architecture Details for FamNet

As described in the main paper, FamNet architecture consists of two key modules: 1) multi-scale feature extraction module 2) density prediction module. The feature extraction module consists of the first four blocks from a pre-trained ResNet-50 backbone(the parameters of these blocks are frozen during training).

The density prediction module has the following architecture: Conv7-196, Upsampling-2, Conv5-128, Upsampling-2, Conv3-64, Upsampling-2, Conv1-32, Conv1-1. Here, ConvX-Y implies a convolution layer having Y filters with $X \times X$ kernel size. Upsampling-2 refers to the bilinear interpolation layer which upsamples the input to twice its size. Upsampling-2 layer does not have any learnable parameters. We use ReLU nonlinearity after each convolution layer.

The density prediction module takes as input 6 correlation maps (2 feature blocks $\times$ 3 scales) and predicts 6 intermediate density maps, one for each of the correlation maps. Final density map is obtained by doing mean pooling across the 6 intermediate density maps.

4. Additional Images from the FSC-147 Dataset

In Fig. 1, we present few representative images from our FSC-147 dataset. We present images from the following visual categories: grapes, boats, marbles, lipstick, alcohol bottles, Go game, chair, beads, zebras, coffee beans, cashew nuts, potatoes, kidney beans, bottle caps and watermelon. We also show the dot annotations and the exemplars for all the images. The dot annotations and the exemplars are shown in red and blue respectively. As can be seen from the images, the number of objects in the images varies widely, some images contain a dozen of objects while some contain thousands. Some of the images in the dataset may also contain large number of distractor objects, as shown by the first and last images.

5. Qualitative Results on FSC-147 Dataset

Next, we present qualitative results on our FSC-147 dataset obtained using FamNet. For this experiment, FamNet is trained on the training set of our dataset, and

we present the predicted density maps on few images from the validation and test sets. We perform test time adaptation using three exemplars, as described in the main paper.

In Fig. 2, we present the results on few images from the validation set. We show the query image along with the exemplars shown by red bounding boxes, the groundtruth density map and predicted density map obtained by FamNet after test time adaptation. The first four query images are success cases for FamNet, while the fifth one is a failure case. The fifth image is an extremely dense image with the ground truth count of over 900. Furthermore, the object of interest is small in size. As a result, FamNet performs poorly on this query image.

In Fig. 3, we present the qualitative results on few images from the test set. We show the query image along with the exemplars shown by red bounding boxes, the groundtruth density map and predicted density map obtained by FamNet after test time adaptation. The first three query images are success cases for FamNet, while the last two are failure cases. The fifth image is rather challenging since there is large variation in the scale of the object of interest because of perspective distortion.

Figure 1: **Few annotated images from the proposed FSC-147 dataset**. Dot and box annotations are shown in red and blue respectively. The number of objects in each image varies widely, some images contain a dozen of objects while some contains thousands.

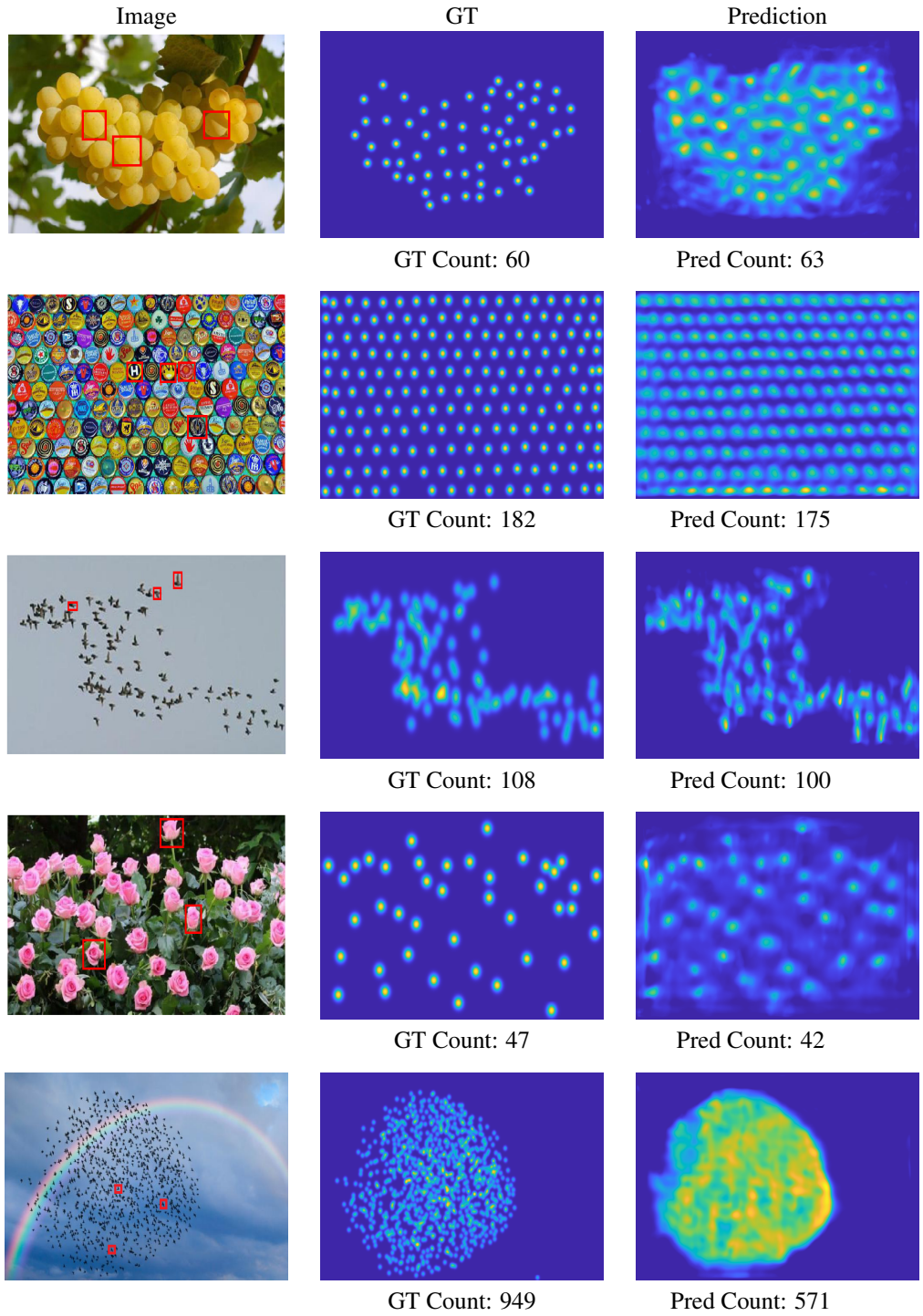

Figure 2: **Predicted density maps and counts of FamNet on the Validation Set of FSC-147 dataset**. Shown are query images, groundtruth maps and predicted density maps. FamNet performs well on the first four test cases, but fails on the last one.

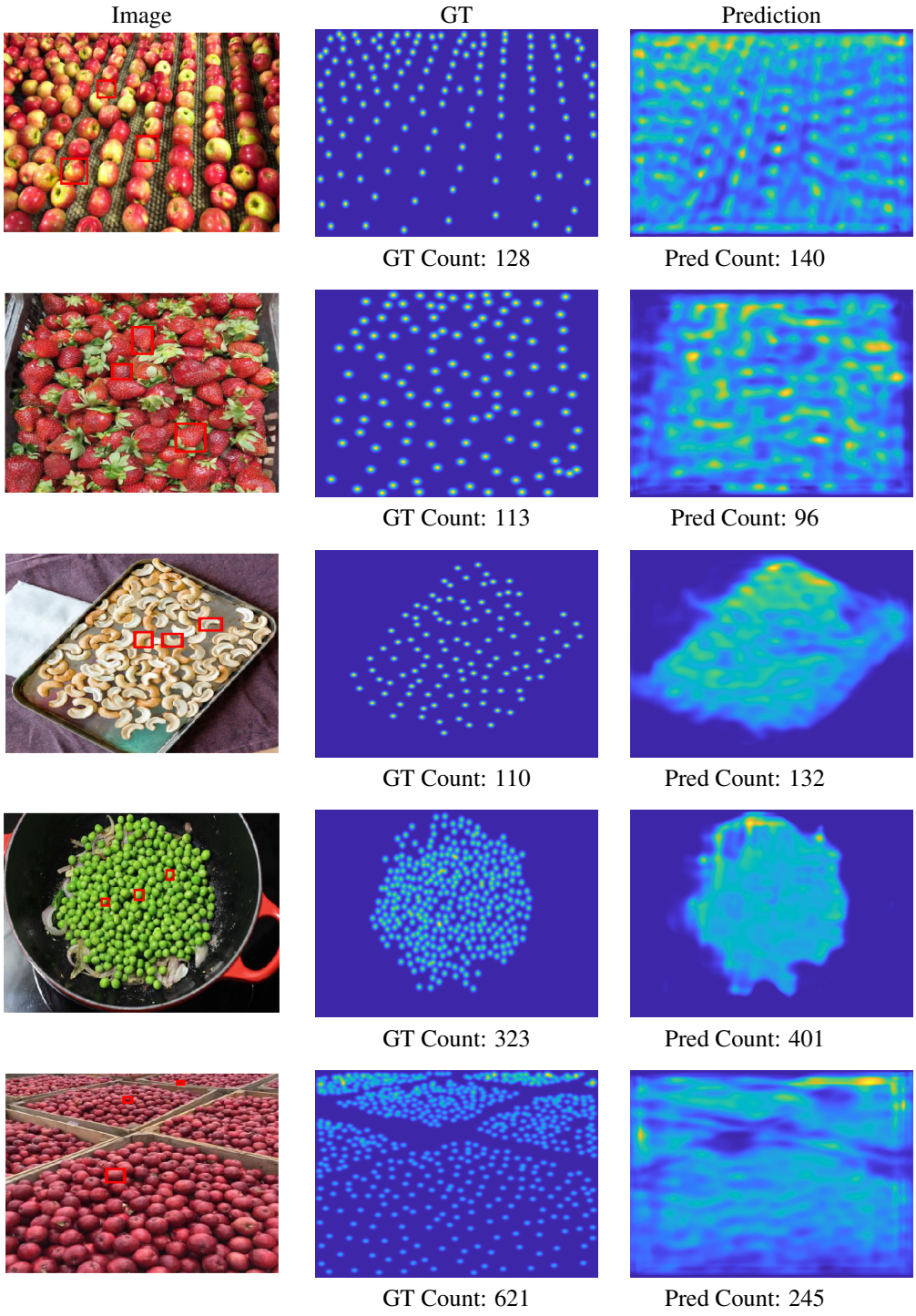

Figure 3: **Predicted density maps and counts of FamNet on the Test Set of FSC-147 dataset**. Shown are query images, groundtruth maps and predicted density maps. FamNet performs well on the first three test cases, but fails on the last two.