Supplementary Material for Learning to Count in the Crowd from Limited Labeled Data

Vishwanath A. Sindagi[1], Rajeev Yasarla[1], Deepak Sam Babu[2], R. Venkatesh Babu[2], and Vishal M. Patel[1]

[1] Johns Hopkins University, Baltimore MD 21218, USA
[2] Indian Institute of Science, Bangalore 560012, India
{vishwanathsindagi,ryasarl1,vpatel36}@jhu.edu
{deepaksam,venky}@iisc.ac.in

Due to limited space in the main paper, we present additional details about the proposed method and experiments in the supplementary.

1 Encoder and Decoder Architecture

Here, we provide details of the encoder and decoder architecture for all the experiments.

Encoder: In the main paper, we conducted experiments with 4 different networks for the encoder: For semi-supervised experiments, we used Res50, Res101 and VGG16. For learning from synthetic data we used Res101-SFCN [4] .Following are the details:

(i) Res50: First 3 layers of Res50 are used as the encoder.
(ii) Res101: First 3 layers of Res101 are used as the encoder.
(iii) VGG16: First 10 layers of VGG16 are used as the encoder.
(iv) Res101-SFCN: We use the network exactly as described in [4]. In this network, the layers until final dilated conv layer are considered as a part of the encoder.

For all the above networks, the features of the final encoder layer are forwarded through a 1×1 conv layer to reduce the dimensionality to 64 channels. The output of this 1×1 conv is the feature embedding in the latent space which is used in GP modeling. Since the train crop size is 256×256, the intermediate feature maps in the latent space is of dimension $64 \times 32 \times 32$.

Decoder: We use the same decoder in all the semi-supervised learning experiments. The decoder consists of 2 conv-relu layers. The first one is a 3×3 conv layer, that takes in 64 channels and outputs 64 channels. The final layer is a a 1×1 layer that takes in 64 channels and outputs 1 channel which is the density map. The final conv layer is followed by an bilinear-upsampling layer that upsamples the output density to the resolution of the input image.

In case of learning from the synthetic data, since we use the same network as in [4], all the layers after the dilated conv layers are used as decoder.

2 Dataset Details

In this section, we provide details of the different datasets used for evaluating the proposed method in the main paper.

ShanghaiTech [6]:This dataset contains 1198 annotated images with a total of 330,165 people. This dataset consists of two parts: Part A with 482 images and Part B with 716 images. Both parts are further divided into training and test datasets with training set of Part A containing 300 images and that of Part B containing 400 images. Rest of the images are used as test set.

UCF-QNRF [2]: UCF-QNRF is a large crowd counting dataset with 1535 high-resolution images and 1.25 million head annotations. There are 1201 training images and 334 test images. It contains extremely congested scenes where the maximum count of an image can reach 12865.

WorldExpo [5]: The WorldExpo'10 dataset was introduced by Zhang *et al.*. [5] and it contains 3,980 annotated frames from 1,132 video sequences captured by 108 surveillance cameras. The frames are divided into training and test sets. The training set contains 3,380 frames and the test set contains 600 frames from five different scenes with 120 frames per scene. They also provided Region of Interest (ROI) map for each of the five scenes.

UCSD [1]: The UCSD dataset crowd counting dataset consists of 2000 frames from a single scene. These scenes contain relatively sparse crowds with the number of people ranging from 11 to 46 per frame. A region of interest (ROI) is provided for the scene in the dataset. Of the 2000 frames, frames 601 through 1400 are used for training while the remaining frames are held out for testing.

GCC [4]:GTA V Crowd Counting Dataset (GCC) is a large-scale synthetic dataset based on an electronic game, which consists of 15,212 crowd images. GCC provides three evaluation strategies (random splitting, cross-camera,and cross-location evaluation).

3 Hyper-parameter λ_{un}

In this section, we study the effect of λ_{un} on the overall performance. λ_{un} weighs the unsupervised loss function in the Eq. 12 of main paper. For this study, we use the ShanghaiTech A dataset, due to its wide variety of scenes and diversity in the count. We conducted this experiment for the 5% data setting where 5% of the data was used as labeled data and rest was used as unlabeled data. We used Res50 encoder. Note that we perform the evaluation on the held-out validation set (and not on the test set). The results for different values of λ_{un} are shown in Table 1.

Table 1. Effect of λ_{un} on ShanghaiTech Part-A val set.

λ_{un}	MAE	MSE
0.0	102	175
0.2	100	162
0.4	89	149
0.6	85	140
0.8	88	147
1.0	92	156

We observed that the performance peaks when the value of λ_{un} is 0.6. $\lambda_{un} = 0$ corresponds to only labeled data. This is the baseline performance. As we increase λ_{un}, we observe that the error improves. However, for $\lambda_{un} > 0.6$, we see a small drop. This is because the network would not have learned to optimal level at the initial stages of training. Due to this the pseud-GT will be erroneous, and hence, using high weight for unsupervised at initial stages prohibits the network from reaching optimal performance.

Based on this experiment, we use $\lambda_{un} = 0.6$ for all the experiments.

4 Additional Architecture Ablation

In this section, we conducted additional architecture ablation experiments using two recent crowd counting techniques: CSRNet [3] and Res101-SFCN [4]. WE use the 5% data-setting, where we use 5% of the data as labeled and rest as unlabeled. We evaluated both these methods on ShanghaiTech-A (SH-A) and UCF-QNRF datasets.For CSRNet, we use the layers upto the last dilated conv as the encoder. For the decoder, we use 2 conv layers as described earlier.

The results of this experiment are shown in Table 2. In addition to MAE/MSE, we report Average Gain (AG)[3]. We observed consistent gains in both the cases when we used the proposed GP-based method to leverage unlabeled data.

Table 2. Semi-supervised experiments with recent crowd counting methods. We used 5% of the training data as labeled set, and the rest as unlabeled samples. AG: Average Gain %[3].

Net	$\mathcal{D}_\mathcal{L}$%	SH-A					UCF-QNRF				
		No-GP($\mathcal{D}_\mathcal{L}$-only)		GP($\mathcal{D}_\mathcal{L} + \mathcal{D}_\mathcal{U}$)		AG	No-GP ($\mathcal{D}_\mathcal{L}$-only)		GP ($\mathcal{D}_\mathcal{L} + \mathcal{D}_\mathcal{U}$)		AG
		MAE	MSE	MAE	MSE	%	MAE	MSE	MAE	MSE	%
Res101-SFCN	100	74	114	-	-	-	113	196	-	-	-
	5	128	199	109	160	17	193	323	172	282	12
CSRNet	100	71	112	-	-	-	123	195	-		-
	5	120	200	111	159	14	187	310	171	293	7.0

[3] $AG = \frac{G_{mae}+G_{mse}}{2}$, $G_{mae} = \frac{mae_{(\mathcal{D}_\mathcal{U}+\mathcal{D}_\mathcal{L})} - mae_{(\mathcal{D}_\mathcal{L})}}{mae_{(\mathcal{D}_\mathcal{L})}}$, $G_{mse} = \frac{mse_{(\mathcal{D}_\mathcal{U}+\mathcal{D}_\mathcal{L})} - mse_{(\mathcal{D}_\mathcal{L})}}{mse_{(\mathcal{D}_\mathcal{L})}}$

5 Multiple Trials

In this section, we report the standard-deviations for the experiments with our proposed method corresponding to Table 1 and Table 4 in the main paper. See Table 3 and Table 4. Note that the standard deviations are computed using 5 trials.

Table 3. Results in SSL settings. Reducing labeled data to 5% results in performance drop by a big margin as compared to 100% data. Res50 was used as the encoder network for all the methods. RL: Ranking-Loss. GP: Gaussian-Process. AG: Average Gain %[3].

Method	$\mathcal{D}_{\mathcal{L}}$	$\mathcal{D}_{\mathcal{U}}$	SH-A			SH-B			UCF-QNRF			WExpo		UCSD		
			MAE	MSE	AG	MAE	MSE	AG	MAE	MSE	AG	MAE	AG	MAE	MSE	AG
Ours	5%	95%	102 ± 0.8	172 ± 2.1	16	15.7 ± 0.9	27.9 (± 1.1)	22	160 ± 2.4	275 ± 3.1	10	12.8 ± 0.5	10	2.0 ± 0.05	2.4 ± 0.09	12

Table 4. Results for synthetic-to-real transfer settings. We train the network on synthetic crowd counting dataset (GCC), and leverage the training set of real-world datasets without any labels. We used the same network and training/evaluation protocol as in [4].

Method	SH-A		SH-B		UCF-QNRF		UCF-CC-50		WExpo
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Ours	121 ± 0.6	181 ± 1.6	12.8± 0.3	19.2± 0.9	210± 2.7	351± 4.1	355 ± 4.4	505± 5.9	20.4 ± 0.9

Bibliography

[1] Chan, A.B., Liang, Z.S.J., Vasconcelos, N.: Privacy preserving crowd monitoring: Counting people without people models or tracking. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–7. IEEE (2008)

[2] Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: European Conference on Computer Vision. pp. 544–559. Springer (2018)

[3] Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1091–1100 (2018)

[4] Wang, Q., Gao, J., Lin, W., Yuan, Y.: Learning from synthetic data for crowd counting in the wild. arXiv preprint arXiv:1903.03303 (2019)

[5] Zhang, C., Li, H., Wang, X., Yang, X.: Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 833–841 (2015)

[6] Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y.: 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 (2016)