Robust Crowd Counting via Image Enhancement and Dynamic Feature Selection

Nayeong Kim
kimnay@postech.ac.kr
Department of CSE
POSTECH, Korea

Suha Kwak
suha.kwak@postech.ac.kr
Graduate School of AI
POSTECH, Korea

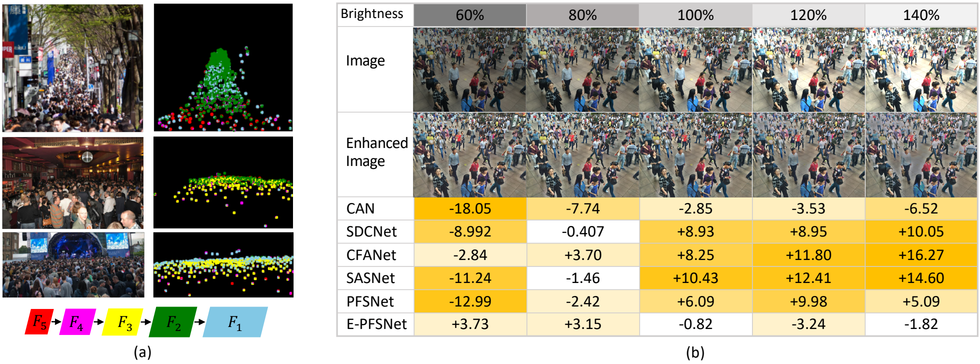
Abstract

In spite of their remarkable success in many vision tasks, convolutional neural networks (CNNs) often has trouble counting people in crowded scenes due to the following reasons. First, ordinary CNNs with fixed receptive fields are inadequate to handle diverse sizes and densities of people. Second, CNNs for counting are sensitive to brightness and contrast changes of input image. This paper proposes a new CNN for crowd counting that resolves these two issues. First, we develop a new counting network called pyramid feature selection network (PFSNet) that adapts its receptive fields dynamically to local crowd densities of the input image. Second, we introduce a light-weight and effective image enhancement network, which manipulates input image to normalize its condition and make it more counting-friendly, leading to robust and improved crowd counting. The concatenation of the two networks, dubbed E-PFSNet, achieves the state of the art on three public benchmarks for crowd counting. Also, it outperforms previous arts in terms of robustness against changes in image conditions as well as counting accuracy.

1 Introduction

Crowd counting is the task of counting the number of people in an image, and has attracted increasing attention since it is essential to tackle timely problems such as visual surveillance and communicable disease control. As this task assumes realistic crowd scenes where a significantly large number of people are densely distributed, it has been typically formulated as regression or classification problems that aim to predict the number of people directly while bypassing explicit pedestrian detection. Recently, convolutional neural networks (CNNs) have been widely adopted for the direct prediction of the count.

In spite of their great potential, the common architecture and training strategy of CNNs however have trouble counting people in real world scenarios because of the following two issues. One of them, which is relatively well-known, is that sizes and densities of people could vary significantly even within a single image due to different camera poses and perspective distortions as shown in Fig. 1(a); hence ordinary CNNs with fixed receptive fields are not optimal. The other issue, which is one of the challenges in general computer vision applications, is that CNNs for counting tend to be sensitive to image conditions and it has



Brightness	60%	80%	100%	120%	140%
Image					
Enhanced Image					
CAN	-18.05	-7.74	-2.85	-3.53	-6.52
SDCNet	-8.992	-0.407	+8.93	+8.95	+10.05
CFANet	-2.84	+3.70	+8.25	+11.80	+16.27
SASNet	-11.24	-1.46	+10.43	+12.41	+14.60
PFSNet	-12.99	-2.42	+6.09	+9.98	+5.09
E-PFSNet	+3.73	+3.15	-0.82	-3.24	-1.82

Figure 1: Two challenges of crowd counting. (a) *Diversity in sizes and densities of pedestrians.* The left side images show that sizes and densities of pedestrians could vary significantly even in a single image. The right side visualizes the level of feature that received the most attention in our PFSNet, and demonstrates that PFSNet is capable of selecting features according to local pedestrian sizes. (b) *Sensitivity to variations of image conditions.* The table shows disparities between predictions of counting models and ground-truth counts (*i.e.*, prediction minus ground-truth). Except E-PFSNet incorporating the image enhancer, all tested models are largely affected by the brightness of the input image. The second row shows that the enhanced image is robust to changes in brightness. These models are also sensitive to brightness and contrast in entire datasets except E-PFSNet.

never been discussed in literature of crowd counting. In particular, we have found that their predictions are affected by brightness and contrast of input image as reported in Fig. 1(b); these results suggest that the counting networks would not be well-generalized to images taken at different time or by different cameras.

In this paper, we present a new crowd counting model that resolves these two issues. To address the first one, we design a new crowd counting CNN, called *pyramid feature selection network* (PFSNet), that adapts its receptive fields dynamically to local crowd densities of input image. Specifically, PFSNet computes multiple feature maps of diverse receptive fields through a feature pyramid network (FPN) [27], and selects features appropriate for handling local crowd densities through an attention module. The use of FPN allows to compute rich features of various receptive fields effectively. Also, compared to previous work using FPN with the same motivation [44], PFSNet is more efficient in size and computation.

Second, we introduce a light-weight and effective image enhancement network, which is attached in front of PFSNet to address the second issue. This network is learned to normalize brightness and contrast of input image so that the effects of these conditions are reduced and the entire counting framework becomes insensitive to them consequently. Further, it not only normalizes but also enhances input image to improve the accuracy of PFSNet since it is also trained along with PFSNet in an end-to-end manner to minimize a counting loss.

Our final model is the combination of the image enhancer and PFSNet, which we call *E-PFSNet*. Our model is evaluated and compared with previous work on six public benchmarks for crowd counting [17, 18, 42, 51, 56], where it achieves the state of the art in almost every dataset and evaluation metric. We also demonstrate that E-PFSNet is robust against variations of image conditions, *i.e.*, brightness and contrast, as intended. The main contribution of this paper is three-fold:

- We propose a new crowd counting network dubbed PFSNet. It can adapt its receptive fields to local crowd densities dynamically, and is more effective and efficient than the existing model based on the same motivation [44].

- To the best of our knowledge, this work is the first to reveal the sensitivity of crowd counting models to brightness and contrast of input image. We also present an image enhancement network as a solution to this issue, and it improves performance and robustness of PFSNet substantially.
- The combination of the image enhancer and PFSNet, dubbed E-PFSNet, achieves the state of the art on three public benchmarks for crowd counting. It also outperforms existing models in terms of robustness against variations of image conditions.

2 Related work

2.1 Crowd Counting

Crowd counting has been tackled by detection-based approaches [13, 24, 26, 57], regression-based approaches [4, 5, 6, 7, 28, 39, 40, 50], and density map estimation approach [23] which is widely used recently. In addition, the side effect of imperfect ground-truth has been alleviated for improved crowd counting [2, 48] and several works targeted the problem of discrepancy between predicted density maps and point annotations [1, 34, 35, 49]. Recently two large-scale congested crowd counting and localization datasets are released [42, 51].

In this section, we focus particularly on multi-scale CNN models that have been proposed to address diverse scales and densities of people like our method. These models can be categorized into two classes: Multi-column and multi-level models.

Multi-column CNNs. Since MCNN [56] proposes a multi-column CNN for crowd counting which extracts multi scale objects, multi-branch network methods have poured out. SANet [3] proposes a stacked multi-branch block to reflect a variety of receptive fields. DADNet [15] studies a multi-dilated convolution to reflect a wider spatial context and a deformable convolution for a high quality density map. These methods try to address the scale variation problem by adopting multi-columns which have different receptive fields, but introduce surplus features.

Multi-level CNNs. Methods in this category detect multi-scale objects using the intrinsic layers of the backbone network. SaCNN [55] employs a single-column network and combine the feature maps from different layers to obtain multi-scale representation. ANF [54] introduces an encoder-decoder network with conditional random fields (CRFs) to aggregate multi-scale features. TEDNet [21] hierarchically aggregates multi-scale features at different encoding stages with multiple decoding paths. SASNet [44] automatically learns the internal level-scale correspondence using FPN [27] without extra annotations or scale estimation strategies, and generates final predictions with weighted sum of level-wise predictions. Followed by these methods, we adopts FPN to compute multi-scale features. Our goal is to generate dynamic feature which dynamically adapts its receptive fields by applying the selection not to predictions but muli-scale features.

2.2 Recognition-aware Image Enhancement

Vision models trained in controlled environments are easily degraded when input images are distorted by weather conditions, blur, and noise in real-world applications. Recognition-aware image enhancement has been studied as a way of resolving this issue. Instead of

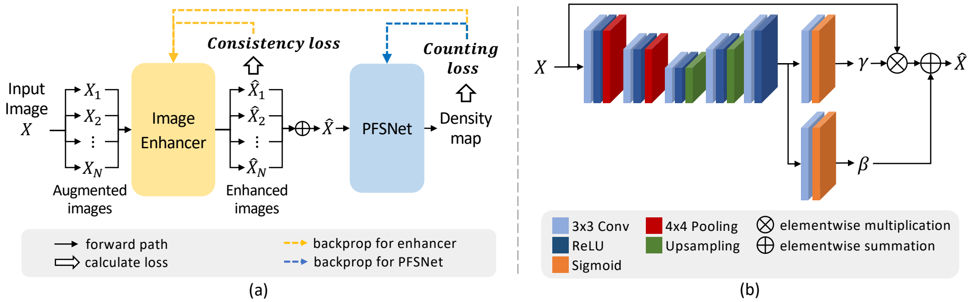


Figure 2: The overall framework of E-PFSNet. (a) An overview of E-PFSNet architecture and training. (b) Detailed architecture of the image enhancement network.

improving perceptual quality of images, it focuses on enhancing images to improve performance of recognition models. Diamond *et al.* [10] developed a differentiable image processing module that is jointly learned with a classifier for blurred and noisy images. Liu *et al.* [29] studied denoising networks for classification and semantic segmentation of noisy images. Gomez *et al.* [14] studied networks that enhance images taken in difficult illumination conditions for robust visual odometry. Son *et al.* [43] proposed a universal enhancement model that can cope with various types of image degradation for diverse vision tasks.

Motivated by these methods, we for the first time reveal the fragility of crowd counting models against image conditions, and propose a counting-aware enhancement network to overcome the limitation.

3 Method

We propose E-PFSNet, a crowd counting network that is robust against variations of image conditions and able to handle diverse crowd densities flexibly. An overview of its architecture and training is illustrated in Fig. 2(a). E-PFSNet consists of two parts, image enhancement network and PFSNet. The image enhancement network manipulates the input image to improve accuracy and robustness of PFSNet, while PFSNet enables effective counting by adapting its receptive fields dynamically according to local crowd densities. The remainder of this section describes architecture details and the training strategy of the two components.

3.1 Image Enhancement Network

As demonstrated in Fig. 1(b), existing counting models are easily degraded by varying brightness and contrast of input image. The reason in the case of regression-based counting models [31, 38, 44] is that they use raw outputs of CNNs as-is as predicted counts although such outputs are affected heavily by the image conditions. Interestingly, the classification-based model [53] is not free from this issue either; we suspect that its ordinal classes (*i.e.*, quantized counts) are highly correlated in particular when they are adjacent, so the ranks of their activations are easily distorted by the image conditions. Indeed, this unexpected fragility of counting models damages their accuracy, robustness, and generalization capability.

The image enhancement network of E-PFSNet alleviates the fragility by normalizing the conditions of input image. Moreover, it enhances the image to improve the performance of

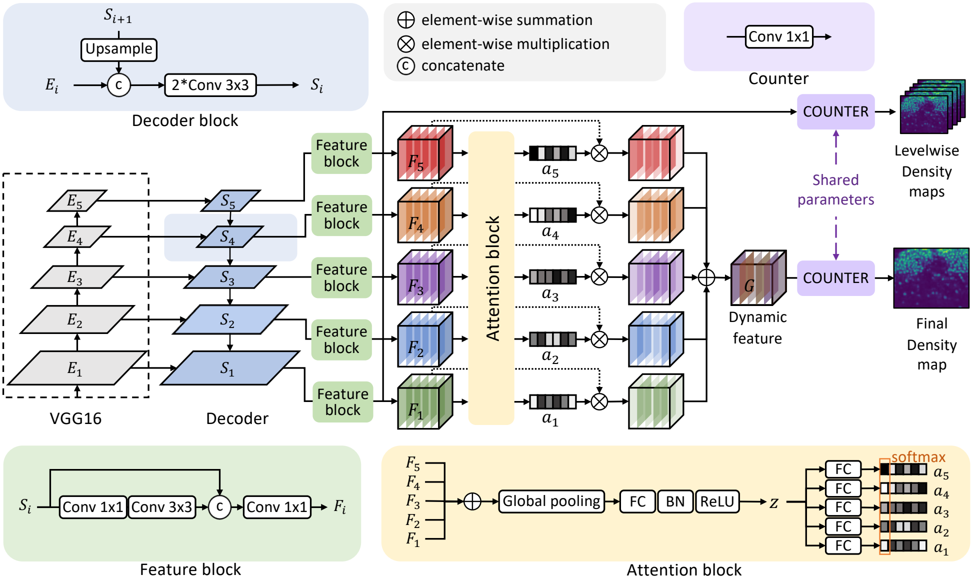


Figure 3: Overall architecture of PFSNet.

the following counting model. As illustrated in Fig. 2(b), the image enhancer manipulates the input image by rescaling and shifting individual pixel intensities. Specifically, it learns to adaptively generate rescaling tensor γ and shift tensor β. The enhanced image $\hat{X}$ is then obtained by scaling and shifting the input image X through γ and β, respectively, and clipping the range of output pixel intensities:

$$\hat{X} = \text{clip}(\gamma \otimes X \oplus \beta), \tag{1}$$

where X, $\hat{X}$, γ, and β are all real tensors of the same size, and $\otimes$ and $\oplus$ indicate element-wise multiplication and summation, respectively.

As shown in Fig. 2(b), the image enhancer has a small encoder-decoder architecture to minimize the computation burden it imposes: The encoder consists of two 3×3 convolution layers and the decoder has three 3×3 convolution layers. Also, the aggressive downsampling and upsampling operations between the layers allow the network to consider large contexts when manipulating pixel intensities. ReLUs follow all convolution layers except for the last layer, which is followed by a sigmoid function. Finally, the ranges of the generated γ and β values are linearly transformed to $[0,4]$ and $[-1,1]$, respectively.

3.2 Pyramid Feature Selection Network

The overall architecture of PFSNet is illustrated in Fig. 3. FPN [27] is adopted as the backbone of PFSNet to compute multi-level feature representations in a hierarchical manner. Then the final feature map with dynamic receptive fields is obtained by aggregating the multi-level features in a channel-wise manner through an attention block. The final feature map is in turn used to predict a crowd density map, which is integrated so as to count the number of people. Two major components of PFSNet are elaborated below.

Feature blocks. A feature block is appended on top of each level of the FPN decoder to generate richer level-wise features, denoted by F_i. The block consists of three convolution layers and a skip connection. In particular, the last convolution layers of all feature blocks have the same number of kernels so that their output features have the same channel dimension. Further, their output features are upsampled to the same spatial resolution.

Attention block. From level-wise features, the attention block generates feature-level attention vectors for dynamic feature selection. We integrate information of all level-wise features $F_i \in \mathbb{R}^{w \times h \times C}$ via element-wise summation with resolution $w \times h$ and C channels. Then we embed integrated feature $z \in \mathbb{R}^{(C/r)}$ by applying global average pooling to the summation and reduction of the dimensionality for efficiency with the reduction rate r. In order to receive z as input and compute the attention vector a_i for each level, a fully connected layer exists for each level; the weight vector of the fully connected layer for the i^{th} level is denoted by $u_i \in \mathbb{R}^{C \times (C/r)}$. We apply channel-wise softmax function to attention vectors.

$$a_i^c = \frac{e^{u_i^c z}}{e^{u_1^c z} + e^{u_2^c z} + e^{u_3^c z} + e^{u_4^c z} + e^{u_5^c z}}, \tag{2}$$

where u_i^c is the c^{th} row of u_i and a_i^c is c^{th} element of a_i. The final feature map G is obtained by weighted summation of level-wise features F_i with soft attention vectors a_i:

$$G^c = \sum_{i=1}^{5} a_i^c \cdot F_i^c \tag{3}$$

where G^c and F_i^c is c-th channel of G and F_i, respectively. We call this final feature map *dynamic feature* due to its dynamic receptive fields.

Counter. The counter generates a crowd density map of the input image from the dynamic feature through a 1 × 1 convolution layer. The counter also computes level-wise density maps from features of FPN during the training time to make feature blocks counting-aware more directly, and it helps performance improvement.

3.3 Training

Training of E-PFSNet consists of two consecutive stages. In the first stage, PFSNet is solely trained as in previous work on crowd counting. Then in the second stage, the image enhancer is trained for normalizing image conditions and for improving crowd counting performance of PFSNet at the same time. Each stage is described in detail below.

Stage 1: Learning PFSNet. The counting loss, denoted by $\mathcal{L}_{count}$, is the sum of the Euclidean distances between predicted density maps and ground truth. The predicted density maps are close to ground truth by minimizing the counting loss. Since PFSNet outputs the final density map D_{final} and five level-wise density maps $D_1, \ldots, D_5$, the loss is given by

$$\mathcal{L}_{count} = ||D_{final} - D^{GT}||_2^2 + \frac{\lambda_1}{5} \sum_{i}^{5} ||D_i - D^{GT}||_2^2, \tag{4}$$

where λ_1 is a hyper-parameter.

Stage 2: Learning the image enhancer. To train the image enhancement network, we first generate several augmented versions of an input image by varying the contrast and brightness of the image at random. The image enhancer then learns to generate the same image from the augmented inputs through a consistency loss, for the purpose of normalizing image conditions. Moreover, the network is trained along with PFSNet, which is previously trained and frozen in this stage, to minimize the counting loss in an end-to-end manner so that it generates more counting-friendly images and improves counting performance consequently.

A straightforward form of the consistency loss is the difference between enhanced images of the augmented inputs. However, minimizing such a loss leads to a degenerate solution in which all the rescaling and shift factors are equal to zero thus image information is totally lost; the enhancer should avoid this situation. Instead, our consistency loss is formulated as the summation of the distances between the enhanced images and their mean. Specifically, each enhanced image is vectorized by spatial pyramid pooling (SPP), where the vector form is denoted by $m_i \in \mathbb{R}^d$. The consistency loss, denoted by $\mathcal{L}_{cons}$, is then defined as the average of the Euclidean distances between m_i and their mean vector μ:

$$\mathcal{L}_{cons} = \frac{1}{N}\sum_{i}^{N} ||m_i - \mu||_2^2, \tag{5}$$

where N is the number of augmented images. The total loss for the enhancer is given by

$$\mathcal{L} = \lambda_2 \mathcal{L}_{count} + \mathcal{L}_{cons}, \tag{6}$$

where λ_2 is a hyper-parameter.

4 Experiments

To demonstrate the effectiveness of E-PFSNet on real-world datasets, we conduct extensive experiments on six challenging crowd counting datasets, including ShanghaiTech PartA&B dataset [56], UCF_CC_50 dataset [17], UCF-QNRF dataset [18], JHU-CRWOD++ dataset[42], and NWPU-Crowd dataset [51]. We use a Gaussian kernel with a fixed size of 15 and the sigma of 4 to create the ground-truth density map following [56]. We use mean absolute error (MAE) and root mean squared error (MSE) as evaluation metrics. In the case of NWPU-Crowd dataset, normalized absolute error (NAE) is also used.

4.1 Implementation Details

The first 13 convolutional layers in VGG16-BN [19, 41] that have been pre-trained on ImageNet [9] are used as the backbone of PFSNet. We randomly select eight images for each iteration during training and crop four images with a fixed size from each image. The crop size is 128x128 for ShanghaiTech Part A, PartB, and UCF-CC-50, 224x224 for UCF-QNRF,

Method	Feature selection	Data aug.	Normalization	MAE	MSE
Average predictions	x	x	x	56.44	93.03
Average features	Average	x	x	54.87	88.33
PFSNet (ours)	Attention	x	x	52.49	81.15
A-PFSNet	Attention	✓	x	56.83	95.55
IN-PFSNet	Attention	x	Non-adaptive	54.37	83.47
E-PFSNet (ours)	Attention	x	Adaptive	51.00	80.88

Table 1: Ablation study of dynamic feature selection and image enhancer on ShanghaiTech Part A. The 'Data aug.' column indicates whether train dataset is augmented with brightness/contrast conditions during training PFSNet.

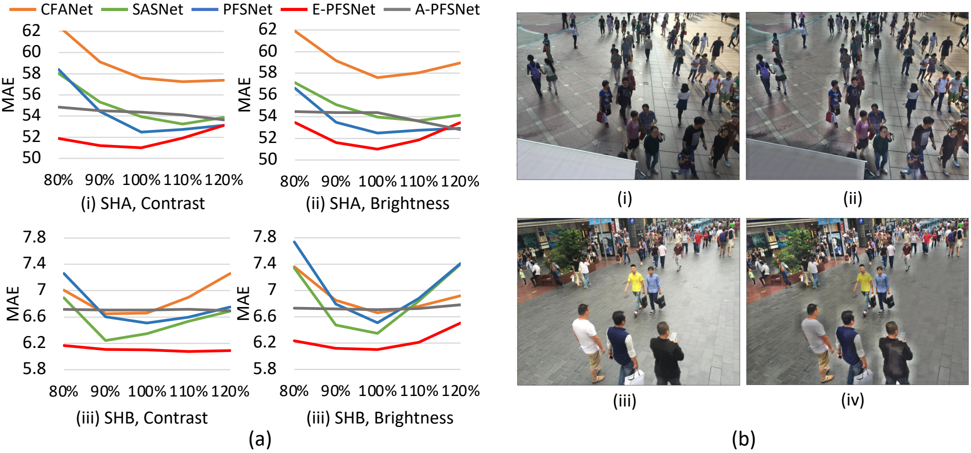


Figure 4: (a) The graph of MAE according to image condition change. We denote the PFSNet trained using augmentation but without enhancer as A-PFSNet. SHA and SHB denote ShanghaiTech PartA and PartB dataset, respectively (b) visualization of enhanced images. (ii) and (iv) are enhanced images from the original image (i) and (iii) respectively.

and 256x256 for JHU-CROWD++. We also apply random horizontal flipping. The batch size of each iteration is 32 for ShanghaiTech PartA, PartB, and UCF-CC-50, 64 for UCF-QNRF, 48 for JHU-CROWD++, and 160 for NWPU-Crowd. We optimize the model using Adam optimizer [22]. The dimensionality reduction rate of the attention block is set to 4. Balance weight λ_1 and λ_2 are set to 2. During training the image enhancer, λ_1 is set to 0 and we make ten copies of the input image and randomly change the brightness and contrast of images between 50% to 150%.

4.2 Ablation Studies

Effectiveness of dynamic feature selection. In this part, two experiments were conducted to screen out the effects of the dynamic feature. First, we compute the final density map by averaging level-wise density maps, which is described as ‘Average predictions’ in Table 1. Compared to ‘Average predictions’, PFSNet reduces the MAE by 7.0% and MSE by 12.8%. This result means that reflecting the dynamic receptive field at the feature level is more effective for counting than reflecting at the density map level. Second, instead of using feature selection, we average level-wise features to generate the dynamic feature, which is described as ‘Average features’ in Table 1. Compared to ‘Average features’, PFSNet reduces MAE by 4.3% and MSE by 8.1%. Our feature selection strategy is effective and the dynamic feature contains richer counting information than level-wise features. PFSNet outperforms the previous work based also on FPN [44] in ShanghaiTech PartA, UCF_CC_50, and UCF-QNRF datasets. PFSNet is even 12.5% and 10.24% lighter, in terms of the number of learnable parameters and FLOPS, respectively. Detailed values of model complexity are in the supplementary material.

Effectiveness of image enhancement for recognition. This section evaluates the effectiveness of the image enhancer in two aspects: robustness and performance improvement. As shown in Fig. 4(a), existing crowd counting models are sensitive to changes in brightness and contrast conditions. However, the image enhancer makes PFSNet robust to brightness and contrast conditions and achieves improved performance. The bottom 3 rows in Table 1

Method	ShanghaiTech PartA		ShanghaiTech PartB		UCF_CC_50		UCF_QNRF		JHU-CROWD++	
	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
CAN [31]	62.3	100.0	7.8	12.2	212.2	243.7	107.0	183.0	101.1	314.0
TEDNet [21]	64.2	109.1	8.2	12.8	249.4	354.5	113	188	-	-
BL [34]	62.8	101.8	7.7	12.7	229.3	308.2	88.7	154.8	75.0	299.9
S-DCNet [53]	58.3	95.0	6.7	10.7	204.2	301.3	104.4	176.1	-	-
SANet+SPANet [8]	59.4	92.5	6.5	9.9	232.6	311.7	-	-	-	-
PaDNet [46]	59.2	98.1	8.1	12.2	185.8	278.3	96.5	170.2	-	-
DUBNet [37]	64.6	106.8	7.7	12.5	243.8	329.3	105.6	180.5	-	-
HYGNN [33]	60.2	94.5	7.5	12.7	184.4	270.1	100.8	185.3	-	-
SDANet [36]	63.6	101.8	7.8	10.2	227.6	316.4	-	-	-	-
SOFA-Net [11]	57.5	92.12	6.80	10.38	185	281	96.2	158.7	-	-
ASNet [20]	57.78	90.13	-	-	174.84	251.63	91.59	159.71	-	-
ADSCNet [2]	55.4	97.7	6.4	11.3	198.4	267.3	**71.3**	**132.5**	-	-
LibraNet [30]	55.9	97.1	7.3	11.3	181.2	262.2	88.1	143.7	-	-
AMRNet [32]	61.59	98.36	7.02	11.00	184.0	265.8	86.6	152.2	-	-
AMSNet [16]	56.7	93.4	6.7	10.2	208.4	297.3	101.8	163.2	-	-
M-SFANet [45]	59.69	95.66	6.76	11.89	162.33	276.76	85.60	151.23	-	-
CG-DRCN-VGG16 [42]	64.0	98.4	8.5	14.4	-	-	112.2	176.3	82.3	328.0
CG-DRCN-Res101 [42]	60.2	94.0	7.5	12.1	-	-	95.5	164.3	71.0	278.6
MNA [48]	61.9	99.6	7.4	11.3	-	-	85.8	150.6	67.7	258.5
DM-Count [49]	59.7	95.7	7.4	11.8	211.0	291.5	85.6	148.3	-	-
CFANet [38]	56.1	89.6	6.5	10.2	203.6	287.3	89.0	152.3	-	-
UOT [35]	58.1	95.9	6.5	10.2	-	-	83.3	142.3	**60.5**	**252.7**
TopoCount [1]	61.2	104.4	7.8	13.7	184.1	258.3	89	159	60.9	267.4
SASNet [44]	53.59	88.38	6.35	9.9	161.4	234.46	85.2	147.3	-	-
PFSNet (ours)	52.49	81.15	6.509	9.82	155.05	228.48	80.87	138.96	61.2	257.8
E-PFSNet (ours)	**51.00**	**80.88**	**6.10**	**9.56**	**137.40**	**210.80**	79.91	138.34	60.6	**252.7**

Table 2: Comparisons with state-of-the-art methods.

compares normalizing strategies in the aspect of performance improvement. First, we simply normalize the input image using an instance normalization layer [47] and train PFSNet, which is denoted as 'IN-PFSNet'. Differ to the image enhancer, the instance normalization removes contrast condition by rescaling and shifting image to same mean and standard deviation. From the results that IN-PFSNet performs worse than PFSNet, instance normalization degrades counting performance. Second, we augment train data by randomly changing the brightness and contrast of an input image and train PFSNet without the image enhancer, which is indicated as 'A-PFSNet'. From the result that A-PFSNet performed similarly to PFSNet, the performance improvement by the image enhancer is not a simple augmentation effect. As illustrated in Fig 4(b), the enhanced image roughly divides the human area, although only dot annotation is provided. The image enhancer reflects the recognition process of PFSNet to improve counting accuracy.

4.3 Comparisons with State of the Art

Table 2 reports the results of five challenging datasets and Table 3 report the results of NWPU-Crowd dataset. Bold numbers indicate the best performance, and underlined numbers indicate the second best.

ShanghaiTech Dataset. ShanghaiTech dataset consists of two parts: ShanghaiTech PartA and ShanghaiTech PartB. PartA contains highly congested scenes than PartB, while PartB is gathered from a busy street and contains relatively sparse scenes. Our E-PFSNet achieves the best performance on both PartA and PartB.

UCF_CC_50 Dataset. UCF_CC_50 is a tiny crowd counting dataset with only 50 images in extremely congested scenes with heavy background noise. The number of head annotations within an image varies from 96 to 4633. To evaluate model performance, We perform 5-fold cross-validation following in [17]. Our E-PFSNet outperforms the previous state-of-the-art

Method	Backbone	Overall			Scene Level (only MAE)		Luminance (only MAE)	
		MAE	MSE	NAE	Avg.	$S_0 \sim S_4$	Avg.	$L_0 \sim L_2$
MCNN [56]	From scratch	232.5	714.6	1.063	1171.9	356.0 / 72.1 / 103.5/ /509.5/ /4818.2	220.9	472.9 / 230.1 / 181.6
SANet [3]	From scratch	190.6	491.4	0.991	716.3	432.0 / 65.0 / 104.2 / 385.1 / 2595.4	153.8	254.2 / 192.3 / 169.7
CSRNet [25]	VGG-16	121.3	387.8	0.604	522.7	176.0 / 35.8 / 59.8 / 285.8 / **2055.8**	112.0	232.4 / 121.0 / 95.5
CAN [31]	VGG-16	106.3	386.5	0.295	612.2	82.6 / 14.7 / 46.6 / 269.7 / 2647.0	102.1	222.1 / 104.9 / 82.3
SCAR [17]	VGG-16	110.0	495.3	0.288	718.3	122.9 / 16.7 / 46.0 / 241.7 / 3164.3	102.3	223.7 / 112.7 / 73.9
BL [34]	VGG-19	105.4	454.2	0.203	750.5	66.5 / 8.7 / 41.2 / 249.9 / 3386.4	115.8	293.4 / 102.7 / 68.0
SFCN+ [52]	ResNet-101	105.7	424.1	0.254	712.7	54.2 / 14.8 / 44.4 / 249.6 / 3200.5	106.8	245.9 / 103.4 /78.8
MNA [48]	VGG-19	96.9	534.2	0.223	608.1	218.7 / 10.7 / 35.2 / **203.3** / 2572.5	93.2	214.0/ 99.6 / 60.0
DM-Count [49]	VGG-19	88.4	388.6	**0.169**	**498.0**	146.7 / **7.6** / **31.2** / 228.7 / 2075.8	**88.0**	**203.6** / 88.1 / 61.2
UOT [35]	VGG-19	**87.8**	387.5	0.185	566.5	80.7 / 7.9 / 36.3 / 212.0 / 2495.4	95.2	240.3 / **86.4** / **54.9**
TopoCount [1]	VGG-16	107.8	438.5	-	-	-	-	-
E-PFSNet (ours)	VGG-16	93.5	**369.9**	0.234	588.6	**38.4** / 12.1 / 42.7 / 230.4 / 2619.3	101.6	253.9 / 90.9 / 62.9

Table 3: Comparisons with state-of-the-art methods on the NWPU-Crowd test set. NWPU-Crod divides test set into following fine-grained subsets: $S_0 \sim< S_4$ respectively indicates five categories according to the different number range: 0, (0, 100], ..., ≥ 5000. There are three more subsets based on images' average luminance values in the YUV color space, which are, $L_0 \sim L_2$ respectively denotes three luminance levels on the test set: [0, 0.25], (0.25, 0.5], and (0.5, 0.75].

method [44] with a 14.9% relative improvement.

UCF-QNRF Dataset. UCF-QNRF dataset is a large-scale crowd counting dataset which contains extremely congested scenes where the maximum count of an image can reach 12865. We limit the maximum size of images to 1920 pixels due to the availability of high-resolution images. Our E-PFSNet achieves the second-best performance with 79.91 MAE and 138.34 MSE, even with the missing information introduced by the downsampling.

JHU-CROWD++ Dataset. JHU-CROWD++ is a large-scale and congested crowd counting and localization dataset under diverse scenarios and environmental conditions, such as different weathers and illumination, consisting 4,372 images. The number of head and box annotation varies from 0 to 25,791. The high-resolution images are resized to 2048 pixels with the original aspect ratio and we do not use box annotation in our experiments. We achieve 60.6 MAE, 252.7 MSE which is compatible performance of state-of-the-art.

NWPU-Crowd Dataset. NWPU-Crowd dataset is a large-scale and congested crowd counting and localization dataset, consisting 5,109 images with 351 negative sample (scenes without people), which are similar to congested crowd scenes in terms of texture features. The ground truth counts for test images are not opened, and the results on the test set must be obtained by submitting to the evaluation server. The high-resolution images are resized to 2048 pixels with the original aspect ratio and we do not use box annotation in our experiments. We achieve 93.5 MAE which is best score in the models which use VGG-16 as backbone network and 369.9 MSE which is state-of-the-art.

5 Conclusion

In this paper, we propose a new CNN for crowd counting E-PFSNet which is robust against variations of images and able to handle diverse crowd densities flexibly. PFSNet adapts its receptive fields dynamically to local crowd densities of the input image by selecting features. The image enhancer alleviates the fragility by normalizing the condition of the input image. E-PFSNet achieves the state of the art on three public benchmarks for crowd counting and also outperforms existing models in terms of robustness against variation of image conditions.

Acknowledgements

This work was supported by the IITP grant and R&D program for Advanced Integrated-intelligence for IDentification funded by Ministry of Science and ICT, Korea (No.2019-0-01906 Artificial Intelligence Graduate School Program–POSTECH, NRF-2018M3E3A1057 306–50%, IITP-2020-0-00842–50%).

References

[1] Shahira Abousamra, Minh Hoai, Dimitris Samaras, and Chao Chen. Localization in the crowd with topological constraints. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[2] Shuai Bai, Zhiqun He, Yu Qiao, Hanzhe Hu, Wei Wu, and Junjie Yan. Adaptive dilated network with self-correction supervision for counting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4594–4603, 2020.

[3] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *Proc. European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[4] Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 545–551. IEEE, 2009.

[5] Antoni B Chan, Zhang-Sheng John Liang, and Nuno Vasconcelos. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7. IEEE, 2008.

[6] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1135–1144, 2017.

[7] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Change Loy. Cumulative attribute space for age and crowd density estimation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2467–2474, 2013.

[8] Zhi-Qi Cheng, Jun-Xiu Li, Qi Dai, Xiao Wu, and Alexander G Hauptmann. Learning spatial awareness to improve crowd counting. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 6152–6161, 2019.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: a large-scale hierarchical image database. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[10] Steven Diamond, Vincent Sitzmann, Stephen P. Boyd, Gordon Wetzstein, and Felix Heide. Dirty pixels: Optimizing image classification architectures for raw sensor data. *arXiv preprint arXiv:1701.06487*, 2017.

[11] Haoran Duan, Shidong Wang, and Yu Guan. Sofa-net: Second-order and first-order attention network for crowd counting. *Proc. British Machine Vision Conference (BMVC)*, 2020.

[12] Junyu Gao, Qi Wang, and Yuan Yuan. Scar: Spatial-/channel-wise attention regression networks for crowd counting. *Neurocomputing*, 363:1–8, 2019.

[13] Weina Ge and Robert T Collins. Marked point processes for crowd counting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2913–2920. IEEE, 2009.

[14] Ruben Gomez-Ojeda, Zichao Zhang, Javier Gonzalez-Jimenez, and Davide Scaramuzza. Learning-based image enhancement for visual odometry in challenging hdr environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 805–811. IEEE, 2018.

[15] Dan Guo, Kun Li, Zheng-Jun Zha, and Meng Wang. Dadnet: Dilated-attention-deformable convnet for crowd counting. In *Proc. ACM Multimedia Conference (ACMMM)*, pages 1823–1832, 2019.

[16] Yutao Hu, Xiaolong Jiang, Xuhui Liu, Baochang Zhang, Jungong Han, Xianbin Cao, and David Doermann. Nas-count: Counting-by-density with neural architecture search. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[17] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2547–2554, 2013.

[18] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proc. European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.

[19] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. International Conference on Machine Learning (ICML)*, 2015.

[20] Xiaoheng Jiang, Li Zhang, Mingliang Xu, Tianzhu Zhang, Pei Lv, Bing Zhou, Xin Yang, and Yanwei Pang. Attention scaling for crowd counting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4706–4715, 2020.

[21] Xiaolong Jiang, Zehao Xiao, Baochang Zhang, Xiantong Zhen, Xianbin Cao, David Doermann, and Ling Shao. Crowd counting and density estimation by trellis encoder-decoder networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6133–6142, 2019.

[22] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[23] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. volume 23, pages 1324–1332, 2010.

[24] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *IEEE International conference on pattern recognition*, pages 1–4. IEEE, 2008.

[25] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1091–1100, 2018.

[26] Sheng-Fuu Lin, Jaw-Yeh Chen, and Hung-Xin Chao. Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6):645–654, 2001.

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] Bo Liu and Nuno Vasconcelos. Bayesian model adaptation for crowd counts. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 4175–4183, 2015.

[29] Ding Liu, Bihan Wen, Xianming Liu, Zhangyang Wang, and Thomas S Huang. When image denoising meets high-level vision tasks: A deep learning approach. 2017.

[30] Liang Liu, Hao Lu, Hongwei Zou, Haipeng Xiong, Zhiguo Cao, and Chunhua Shen. Weighing counts: Sequential crowd counting by reinforcement learning. In *Proc. European Conference on Computer Vision (ECCV)*, pages 164–181. Springer, 2020.

[31] Weizhe Liu, Mathieu Salzmann, and Pascal Fua. Context-aware crowd counting. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5099–5108, 2019.

[32] Xiyang Liu, Jie Yang, and Wenrui Ding. Adaptive mixture regression network with local counting map for crowd counting. In *Proc. European Conference on Computer Vision (ECCV)*, 2020.

[33] Ao Luo, Fan Yang, Xin Li, Dong Nie, Zhicheng Jiao, Shangchen Zhou, and Hong Cheng. Hybrid graph neural networks for crowd counting. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11693–11700, 2020.

[34] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 6142–6151, 2019.

[35] Zhiheng Ma, Xing Wei, Xiaopeng Hong, Hui Lin, Yunfeng Qiu, and Yihong Gong. Learning to count via unbalanced optimal transport. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, volume 35, pages 2319–2327, 2021.

[36] Yunqi Miao, Zijia Lin, Guiguang Ding, and Jungong Han. Shallow feature based dense attention network for crowd counting. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11765–11772, 2020.

[37] Min-hwan Oh, Peder Olsen, and Karthikeyan Natesan Ramamurthy. Crowd counting with decomposed uncertainty. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 11799–11806, 2020.

[38] Liangzi Rong and Chunping Li. Coarse-and fine-grained attention network with background-aware loss for crowd density map estimation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3675–3684, 2021.

[39] David Ryan, Simon Denman, Clinton Fookes, and Sridha Sridharan. Crowd counting using multiple local features. In *2009 Digital Image Computing: Techniques and Applications*, pages 81–88. IEEE, 2009.

[40] Chong Shang, Haizhou Ai, and Bo Bai. End-to-end crowd counting via joint learning local and global count. In *IEEE International Conference on Image Processing (ICIP)*, pages 1215–1219. IEEE, 2016.

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations (ICLR)*, 2015.

[42] Vishwanath Sindagi, Rajeev Yasarla, and Vishal MM Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.

[43] Taeyoung Son, Juwon Kang, Namyup Kim, Sunghyun Cho, and Suha Kwak. Urie: Universal image enhancement for visual recognition in the wild. In *Proc. European Conference on Computer Vision (ECCV)*, pages 749–765. Springer, 2020.

[44] 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.

[45] Pongpisit Thanasutives, Ken-ichi Fukui, Masayuki Numao, and Boonserm Kijsirikul. Encoder-decoder based convolutional neural networks with multi-scale-aware modules for crowd counting. In *IEEE International conference on pattern recognition*, 2020.

[46] Yukun Tian, Yiming Lei, Junping Zhang, and James Z Wang. Padnet: Pan-density crowd counting. *IEEE Transactions on Image Processing (TIP)*, 29:2714–2727, 2019.

[47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[48] Jia Wan and Antoni Chan. Modeling noisy annotations for crowd counting. In *Proc. Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[49] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai. Distribution matching for crowd counting. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2020.

[50] Chuan Wang, Hua Zhang, Liang Yang, Si Liu, and Xiaochun Cao. Deep people counting in extremely dense crowds. In *Proc. ACM Multimedia Conference (ACMMM)*, pages 1299–1302, 2015.

[51] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting and localization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI).*

[52] Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8198–8207, 2019.

[53] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, and Chunhua Shen. From open set to closed set: Counting objects by spatial divide-and-conquer. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 8362–8371, 2019.

[54] Anran Zhang, Lei Yue, Jiayi Shen, Fan Zhu, Xiantong Zhen, Xianbin Cao, and Ling Shao. Attentional neural fields for crowd counting. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, pages 5714–5723, 2019.

[55] Lu Zhang, Miaojing Shi, and Qiaobo Chen. Crowd counting via scale-adaptive convolutional neural network. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1113–1121. IEEE, 2018.

[56] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 589–597, 2016.

[57] Tao Zhao and Ramakant Nevatia. Bayesian human segmentation in crowded situations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–459. IEEE, 2003.