DECCNet: Depth Enhanced Crowd Counting

Shuo-Diao Yang Hung-Ting Su Winston H. Hsu Wen-Chin Chen

National Taiwan University

Abstract

Crowd counting which aims to calculate the number of total instances on an image is a classic but crucial task that supports many applications. Most of the prior works are based on the RGB channels on the images and achieve satisfied performance. However, previous approaches suffer from counting highly congested region due to the incomplete and blurry shapes. In this paper, we present an effective crowd counting method, Depth Enhanced Crowd Counting Network (DECCNet), which leverages the estimated depth information with our novel Bidirectional Cross-modal Attention (BCA) mechanism. Utilizing the depth information enables our model to explicitly learn to pay attention to those congested regions on the basis of the depth information. Our BCA mechanism interactively fuses two different input modalities by learning to focus on the informative parts according to each other. In our experiments, we demonstrate that DECCNet outperforms the state-of-the-art on the two largest crowd counting datasets available, including UCF-QNRF, which has the highest crowd density. The visualized result shows that our method can accurately regress dense regions through leveraging depth information. Ablation studies also indicate that each component of our method is beneficial to final prediction.

1. Introduction

Crowd counting has attracted the researchers, thanks to the massive growth of the various types of unmanned cameras. Many applications such as traffic control, election rally or video surveillance are built upon the accurate crowd counting. Recent methods [2, 4, 8, 20, 24, 29, 32, 35, 37] estimate the count with RGB channels on the image by generating density map without explicitly detect accurate position then taking integration of whole predicted density map and achieve promising results. These methods most leverage CNN architecture due to its excellent feature representation learning compared with hand-crafted features. However, these approaches suffer from highly congested regions due to severe occlusion, perspective distortion and non-uniform crowd distribution where the RGB channels of pixels are too small to offer the informative representation to the crowd counting networks, demonstrated in figure 1. Furthermore, we find that those regions are often deeper among the whole image, resulting in the count of pixels of each head is smaller than other, which is hard to estimate than those close-to-camera heads.

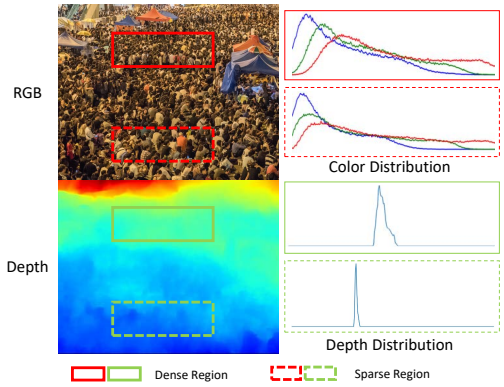

Figure 1. This figure explains the design intuition of our method. The left column is RGB image and its corresponding depth channel. The right column is the distribution of each bounding box. It can be observed that deep regions tend to have denser crowds. In RGB image, the color distribution of the solid box (dense region) and the dashed box (sparse region) are roughly the same, while in the depth image, their corresponding depth value is quite different, providing extra information compared to RGB and can be leveraged as prior to the network.

To tackle this, we introduce an effective Depth Enhanced Crowd Counting Network (DECCNet), which leverages the RGB channels with our estimated depth information that has been widely used and improves many tasks such as object detection [34], image segmentation [15] and 3D scene reconstruction [16]. Figure 1 explains the intuition of our method. Since an image may contain several crowds that distance of each crowd to camera differs, scale variation of people hurts the performance. Furthermore, information which RGB channel offers has insufficient capability to represent such difference. But in the depth space, dis-

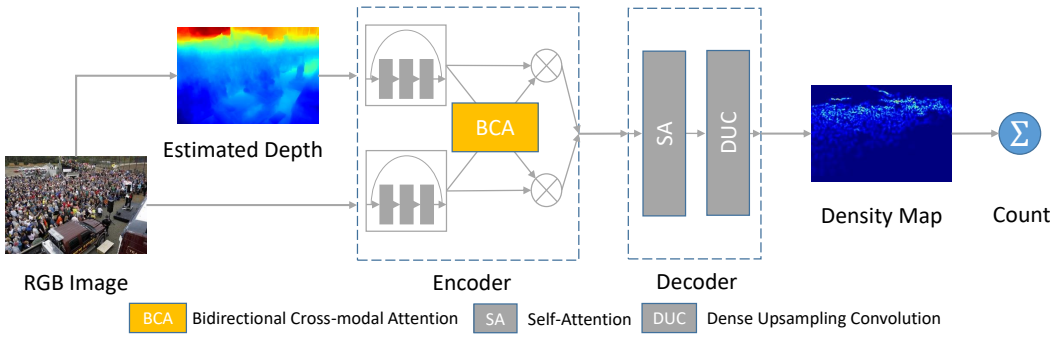

Figure 2. Overall architecture of the proposed method. In addition to the RGB image, we also use the estimated depth as an input to our model. Our proposed component, BCA (bidirectional cross-modal attention), which attends the color and depth modalities according to another (details in section 3.3), is applied to two-stream Resnet encoder, while decoder is composed of a self-attention layer followed by DUC layers. The final count is calculated by taking integration of the predicted density map.

tance of each pixel from the camera can be leveraged to alleviate such problems, as demonstrated in figure 1. To fuse two input modalities, we also propose a novel bidirectional cross-modal attention (BCA) which attends the color and depth modalities according to another, on the basis of the observation that despite the depth is additionally informative, the depth and the RGB channels provide the signal to each other to encourage the model to pay attention to the desirable regions (figure 7).

The experiments on various benchmarks demonstrate that our DECCNet surpasses the state-of-the-art models. The qualitative results demonstrate that our proposed approach accurately estimates the far and blurry regions where previous state-of-the-art methods suffer from (figure 4, 5 and 6). Our ablation study also shows the significance of BCA by comparing different fusing methods and our method performs better than directly concatenating depth as the fourth channel for input RGB image or without BCA.

To summarize, our main contributions of this paper are as follows:

- In addition to RGB data, we argue the importance of depth information based on the observation in figure 1. To the best of our knowledge, we are the first to discover the advantage of incorporating depth channel to crowd counting. Recent works use only RGB as input and tend to fail in highly congested regions due to scale variation.
- We proposed an effective attention mechanism called BCA (bidirectional cross-modal attention) to progressively fuse two different input modalities and shows that BCA can correctly focus on deep/dense regions.
- In ShanghaiTech dataset, we reach MAE of 58.6 in part A and MAE of 7.1 in part B, and MAE of 107.9 in UCF-QNRF dataset which outperforms the state-of-the-art crowd counting methods.

2. Related work

There is a large number of methods proposed in recent years aimed at crowd counting and single image depth estimation. In this section, we discuss each task respectively.

2.1. Crowd counting

Crowd counting methods can be mainly divided into three categories: detection-based, regression-based and density estimation-based method. Detection-based methods such as [9, 11, 26] use head or body-part detectors to localize each person's position. However such methods often fail on dense scenes due to severe occlusion and low resolution of each person. To tackle this problem, regression-based methods such as [5, 6] learn the mapping from image feature to the number of people in a patch-based fashion. This kind of methods consists of two steps, image feature extraction and regression from the extracted features. Although regression-based methods perform better than detection-based methods, outputting only a number of people lacks spatial information.

More recent works such as [4, 20, 24, 36] are density estimation-based methods that leverage the power of CNN, due to its great feature representation learning capability, to generate density map. [45] propose a multi-column CNN with various receptive-field size to aggregate different scale of people/head. [33] proposes a switch CNN that consists of a switcher and many independent regressors, each of them has different receptive fields. [39] introduces a contextual pyramid CNN which leverages global and local context of crowd image. [24] use VGG backbone and dilated kernel to enlarge the receptive field of convolution, resulting in performance gain in congested regions. [35] generates density maps by leveraging adversarial training, compared to only using L2 loss, resulting in a sharper prediction. [20] proposed a novel loss of combining count, density map estimation and localization in addition to only MSE loss. [4]

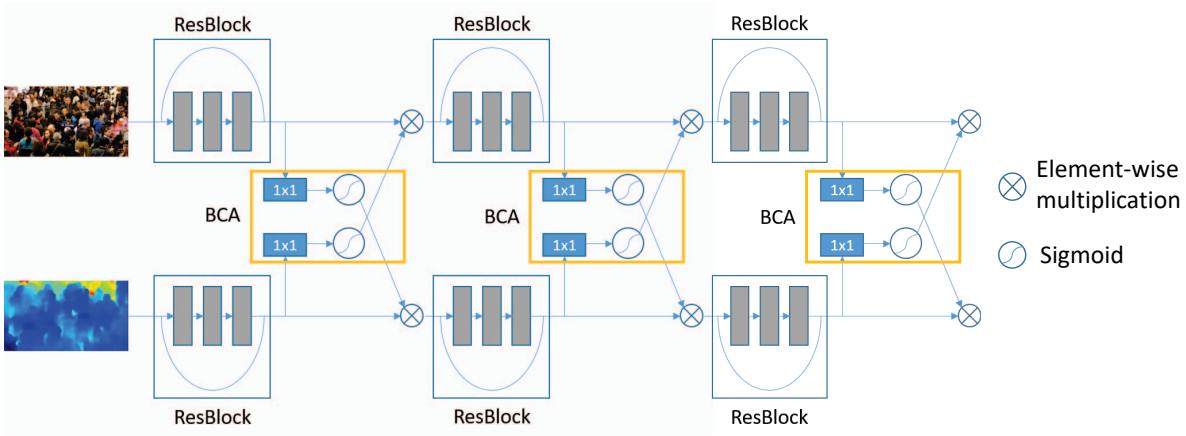

Figure 3. Encoder of DECCNet: Encoder uses the first 3 ResBlocks of ResNet-50 as the main backbone, coupled with proposed bidirectional cross-modal attention mechanism (BCA). Detail of BCA can be found in section 3.3. Two input streams interactively attend to each other after each ResBlock, scaling feature map for each modality. With BCA, two input modalities are progressively fused from the low level feature to high level representations. Overall architecture can be found at table 1.

propose a novel SSIM [42] loss to emphasize local patch consistency, and use an efficient Inception-like [40] network that has only about 1M parameters. [36] propose a perspective-aware network that simultaneously estimates density maps and perspective maps resulting in state-of-the-art performance. However, previous methods suffer from estimation for those high-density regions due to scale variation of the crowd, which can be enhanced by leveraging our proposed depth information.

2.2. Single image depth estimation

There are lots of works [10, 23, 27, 28] proposed to predict depth from a single RGB image. Some datasets such as NYU [31] or KITTI [12] are generated by RGB-D sensor or laser. Furthermore, MegaDepth [25] is a large dataset whose images are collected from the Internet, containing about 130K images. [10] presents a method that first generates a coarse global prediction then refines the predicting locally. [27] formulates single image depth prediction into a conditional random field learning problem, combined with deep CNN. [23] proposes a novel end-to-end method that leverages residual [17] architecture and can run in real time.

For this work, as estimating depth is not our main focus, we do not directly estimate depth from the RGB image. Instead, we use a state-of-the-art pretrained model, MegaDepth [25], to extract depth information from the RGB image. Further end-to-end methods can be further investigated.

3. Method

In this section, we will discuss our proposed DECCNet (figure 3) in detail. Section 3.1 presents the motivation, followed by depth estimation in section 3.2, the network design in section 3.3 and the loss function in section 3.4.

3.1. Motivation

The bottleneck of the crowd counting task where previous methods suffer from is to accurately estimate the count on certain regions where the instances are dense, small and overlapping. Moreover, these instances can dominate the count due to the high density. We observe that the regions where previous approaches can hardly estimate are usually far away from the viewpoint, and argue to fill this gap by encouraging the model to pay attention to the faraway regions. Our framework leverages the encoder-decoder architecture with following novelties: 1) combining depth channel along with RGB as the input of the model and 2) bidirectional cross-modal attention mechanism that effectively fuses two input modalities.

3.2. Depth estimation

Since available crowd counting datasets have only RGB images, generating depth information turns out to be a critical task. We decide to directly use MegaDepth [25] as our main depth extraction model. MegaDepth is a state-of-the-art single image depth estimation model trained on more than 100K examples, including indoor and outdoor scenes, which shows better prediction on crowd counting datasets whose images are usually outdoor scenes, than other models.

Besides, to show the robustness of our concept that incorporating depth information is beneficial to final prediction in spite of the quality of the estimated depth, we also extract depth channel using MonoDepth [14] which is mainly designed for KITTI [12], not general cases.

3.3. Network architecture

Inspired by [18], convolutional encoder-decoder architecture has shown great success in various works [3, 7]. Our proposed network consists of a front-end cross-modal encoder to encode two different input modalities and a back-end density map decoder to progressively generate estimated density map. Encoder is a two-stream ResNet [17] coupled with a novel attention mechanism, bidirectional cross-modal attention (BCA), to fuse RGB and depth inputs. Decoder uses self-attention mechanism [44] to catch long-range dependency followed by Dense Upsampling Convolution (DUC) [41] to generate the high resolution density map. Each component is discussed in the following section.

3.3.1 Cross-modal encoder

As shown in figure 3, we use the first 3 block of Resnet-50 as the main backbone due to Resnet's powerful feature representation. We remove the fourth block of Resnet and fully connected layer since model should accept arbitrary input size and the fourth block will downsample feature map to 1/16 of the original input size which is too small for later upsampling. Instead of directly concatenating RGB and depth channel as input or lately fusing extracted feature of each input modalities, we introduce a novel Bidirectional Cross-modal Attention (BCA) mechanism to interactively affect one stream by the other.

Suppose there are two stream, RGB stream and depth stream, output of each ResBlock is denoted as:

$$R_i^{\{rgb,d\}}, i = \{1, 2, 3\}. \tag{1}$$

Then Bidirectional Cross-modal Attention is defined as:

$$R_i'^{rgb} = R_i^{rgb} \otimes Sigmoid(Conv(R_i^d)), \tag{2}$$

$$R_i'^{d} = R_i^{d} \otimes Sigmoid(Conv(R_i^{rgb})). \tag{3}$$

$\otimes$ denotes element-wise multiplication with broadcasting and $Conv$ is 1×1 convolution that reduces the number of channels to one to generate an attention map. Then $R_i'^{rgb,d}$ is used as input to next ResBlock or decoder that extracts higher level representations or generate density map.

There are two attention directions, one is depth to RGB, and the other is RGB to depth. Depth to RGB, shown in equation 2, scales the RGB feature map by depth stream, making RGB stream focus on deep regions. On the other hand, RGB to depth, shown in equation 3, use RGB information to enhance depth stream since depth information our model used is an estimation, not an accurate one.

This BCA mechanism is repeated three times for each output of ResBlock, progressively fusing two input modalities from low level feature map to high level representations (see figure 7). Final outputs of each stream are concatenated and then send to decoder. Table 4 shows the effectiveness of BCA. Besides its performance, BCA uses only 1×1 convolution to produce a one-channel attention map for each stream, slightly increasing trainable parameters by about 2.7K of the whole network, but significantly reduces the error (table 4), indicating that this attention mechanism is also an efficient one.

Network Architecture	
RGB	Depth
Cross-modal encoder	
c48k7s2	c48k7s2
ResBlock (f=48, s=3)	ResBlock (f=48, s=3)
Bidirectional cross-modal attention (BCA)	
ResBlock (f=96, s=4)	ResBlock (f=96, s=4)
BCA	
ResBlock (f=192, s=6)	ResBlock (f=192, s=6)
BCA	
Density map decoder	
concat	
c512k1s1	
Self-Attention	
2*c256k3s1, DUC (r=2)	
2*c128k3s1, DUC (r=2)	
2*c64k3s1, DUC (r=2)	
2*c32k3s1	
c1k1s1	
Output	

Table 1. Detailed network architecture of DECCNet. 2*c256k3s1 denotes convolution operation with 256 output channels, kernel size of 3×3, strides 1, repeated 2 times. ResBlock (f=48, s=3) means a standard ResBlock with 48 initial filter, block size of 3. DUC (r=2) means exchanging feature map of $1 \times 1 \times 4$ for $2 \times 2 \times 1$. Details of BCA can be found at figure 3 and section 3.3.

3.3.2 Density map decoder

Density map decoder is composed of a standard 1x1 convolution to reduce the channel sent from the encoder and self-attention mechanism [44] followed by Dense Upsampling Convolution [41]. These components are commonly used in other areas but also boost the performance of our work.

Self-attention mechanism is firstly used on Generative Adversarial Network, which fails to capture geometric or structural patterns [44]. Instead of standard convolution operation only processing information of local neighborhood, self-attention mechanism attends on each pixel guided by global context. By incorporating self-attention mechanism in our crowd counting network, decoder can deal with the long-range dependency which crowd regions on images of-

ten present.

Since the size of the feature map sent from the encoder is only 1/8 of the original image, directly upsampling to original size using bilinear method leads to blurry density maps. Hence, choosing an appropriate upsampling method is important, considering the quality of the predicted density map. Dense Upsampling Convolution [41], also known as pixel shuffler or sub-pixel convolution, upsample the feature map by reassigning each $1*1*r^2$ feature map subregion to $r*r*1$, without any parameterized operation. Since DUC exchanges channel for space, two convolutional layers are applied after each DUC operation to increase the channel for later DUC layers. Experiments show that DUC can generate high quality density map as shown in figure 4.

3.4. Loss functions

Our model leverages two loss functions, euclidean loss and SSIM loss. The former focuses on pixel-wise similarity while the latter emphasizes patch consistency.

3.4.1 Euclidean loss

Instead of estimating a single number of people of the given image, our method regresses each pixel of density map whose size is the same as the input. Thus, Euclidean loss is chosen to force the estimated density map as close as ground truth as possible. Euclidean loss compares the difference between two images, defined as:

$$L_E = \frac{1}{N}\sum_{i=1}^{N}\|P_i - GT_i\|^2, \tag{4}$$

supposed there are N pixels in a given image where P_i is the prediction of our model and GT_i is ground truth density map.

3.4.2 SSIM loss

SSIM [42] is firstly used to measure the structural similarity of images. [4] uses SSIM as loss function to train network and achieves promising result. Compared with Euclidean loss, which encourages only the similarity of each, not a group of pixels. SSIM loss considers the local patch consistency by applying a sliding window on the given image, computing each region's SSIM respectively. Following [4], we use a 11x11 Gaussian kernel with a standard deviation of 1.5 to compute local statistics. SSIM of a given point is defined as:

$$SSIM = \frac{(2\mu_p\mu_{gt} + c_1)(2\sigma_{p,gt} + c_2)}{(\mu_p^2 + \mu_{gt}^2 + c_1)(\sigma_p^2 + \sigma_{gt}^2 + c_2)} \tag{5}$$

where μ_p and μ_{gt} are mean and σ_p and σ_{gt} are standard deviation of prediction map and ground truth density map, $\sigma_{p,gt}$ denotes the covariance, c_1 and c_2 are small value to avoid division by zero.

Aggregating each pixel of prediction, SSIM loss is defined as:

$$L_{SSIM} = 1 - \frac{1}{N}\sum SSIM(\mathbf{x}), \tag{6}$$

where $\mathbf{x}$ is the position of the map.

Final loss is calculated by fusing these two loss functions:

$$L = L_E + \lambda L_{SSIM}, \tag{7}$$

where λ is used to balance two loss functions. In this paper we set λ to $1e-3$.

ShanghaiTech Dataset				
	Part A		Part B	
Method	MAE	MSE	MAE	MSE
Zhang *et al.* [43]	181.8	277.7	32.0	49.8
MCNN [45]	110.2	173.2	26.4	41.3
CP-CNN [39]	73.6	106.4	20.1	30.1
ic-CNN [32]	68.5	116.2	10.7	16.0
CSRNet [24]	68.2	115.0	10.6	16.0
SANet [4]	67.0	104.5	8.4	13.6
PACNN [36]	62.4	102.0	7.6	11.8
DECCNet (Ours)	**58.6**	**101.1**	**7.1**	**11.4**

Table 2. The performance comparison between DECCNet and other methods evaluated on both parts of ShanghaiTech dataset. Average count of people in part A is higher than part B. Lower MAE/MSE represents better performance. Our method outperforms all previous methods by a large margin.

4. Experiments

In this section, we first discuss the datasets used to evaluate our method in section 4.1, implementation details in section 4.2, followed by main result in section 4.3. Ablation studies and visualization are discussed in section 4.4 and section 4.5 respectively.

4.1. Datasets

There are several crowd counting datasets, ranging from sparse to dense crowds. We experiment our method on two largest and well-annotated crowd counting datasets: ShanghaiTech and UCF-QNRF.

ShanghaiTech [45]: It contains 1198 images in RGB or greyscale with 330,165 annotations on the center of each head, divided into two parts A and B. Part A has 482 images, 300 for training and 182 for testing. Part B has 716 images, 400 for training and 316 for testing. The main differences between part A and B is that part A is collected from Internet whose average count of people is 501 while

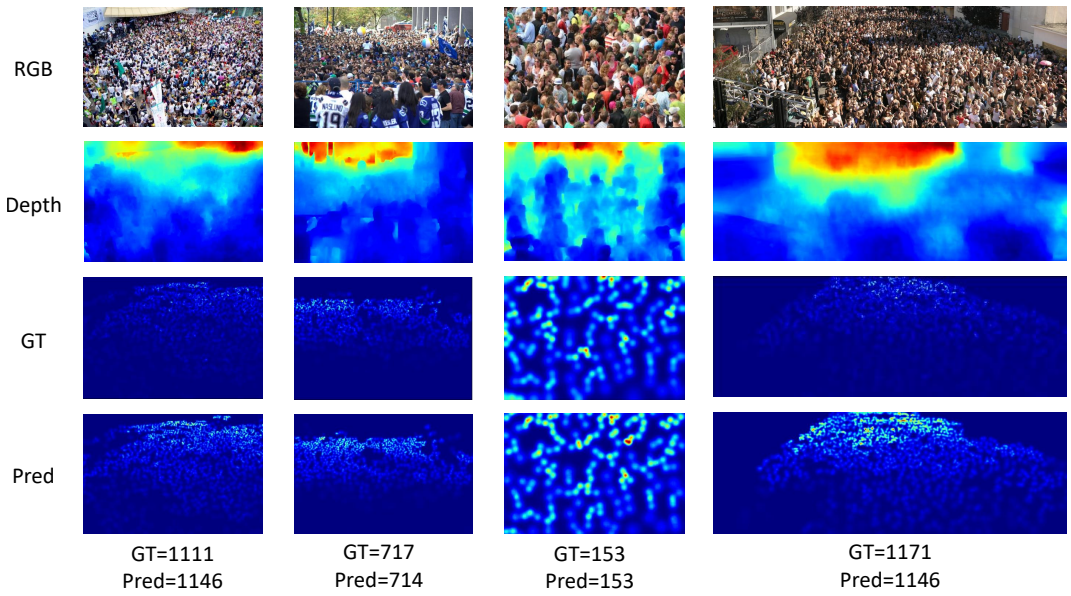

Figure 4. Visualization of predicted density map and final count of our method. First row is the original RGB image, second row is depth generated from pretrained model, third row is ground truth density map and fourth row is our model's prediction. It's obvious that with the help of depth information, those congested regions can be accurately estimated.

UCF-QNRF Dataset		
Method	MAE	MSE
Idrees [19]	315.0	508.0
MCNN [45]	277.0	426.0
Switch-CNN [33]	228.0	445.0
CMTL [38]	252.0	514.0
CL [20]	132.0	191.0
DECCNet (Ours)	**107.9**	**179.0**

Table 3. This table shows the performance comparison between our method and the state-of-the-arts evaluated on UCF-QNRF dataset, which is the most challengeable one. Our DECCNet reaches MAE of 107.9 which is 18% lower than previous methods.

Ablation Studies				
	Part A		Part B	
Method	MAE	MSE	MAE	MSE
DECCNet	**58.6**	**101.1**	**7.1**	**11.4**
RGB-D fused	63.9	106.3	8.7	14.8
MonoDepth	61.5	97.2	8.6	14.4
w/o depth	60.8	102.9	9.3	15.1
w/o BCA	62.4	104.6	9.1	14.1
w/o self-attn	62.0	102.2	8.3	13.9
w/o DUC	58.9	101.9	8.7	14.6

Table 4. Ablation studies conducted on ShanghaiTech dataset. MonoDepth denotes replacing MegaDepth with MonoDepth. RGB-D fused denotes directly using depth information as 4^{th} channel of RGB image, which results in a one stream network. This table shows the significance of each component of our network. Dropping any of them leads to performance decrease.

part B is collected from busy streets of metropolitan areas in Shanghai, China, which has an average count of 123.

UCF-QNRF [20]: This is the largest and the newest real-world crowd counting dataset. Collected from the Internet, this dataset contains highly congested crowd image with higher resolution (2013×2092 on average), denser crowds (815 people per image on average) and finer annotations. Training set consists of 1201 images while testing set has 334 images.

4.2. Implementation details

First, we need to generate ground truth density map as training target. Since annotations in the aforementioned datasets have only coordinate of each person, We can generate a binary map as:

$$H_{i,j} = \begin{cases} 1 & if\ (i,j)\ is\ annotated \\ 0 & else \end{cases} \tag{8}$$

Then convolve H with geometric-adaptive Gaussian kernel [45] or fixed Gaussian kernel. Part A of ShanghaiTech is generated by applying geometric-adaptive Gaussian kernel and others are generated by a fixed kernel.

Since images' sizes are different, for those images bigger than 1024×768, we resize their height to 768, then randomly crop a patch of 256×256 during training. To prevent

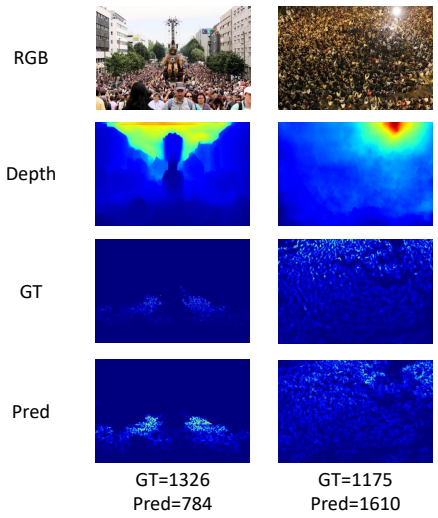

Figure 5. Failure cases. First row is the original RGB image, second row is depth generated from pretrained model, third row is ground truth density map and fourth row is our model's prediction. Depth prediction of these images is noisy and inaccurate. Our method depends on two input, RGB and depth. If one is incomplete or erroneous, performance will degrade.

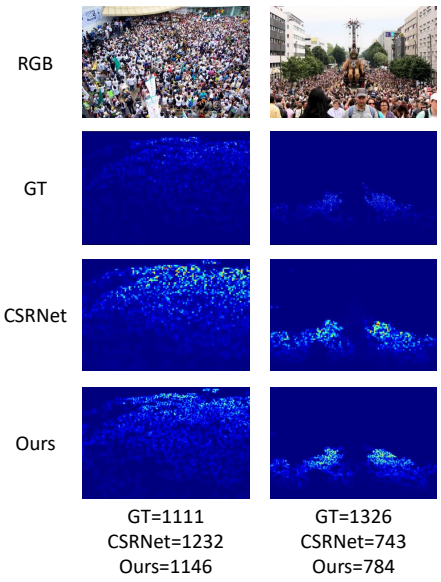

Figure 6. Comparison versus state-of-the-art. First row is the original image, second row is ground truth density map, third row is the prediction of CSRNet and fourth row is our prediction. In highly congested regions, our DECCNet generates more accurate crowd distribution and final count.

overfitting, each example is horizontally flipped with probability 0.5 and converted to grey scale with probability 0.1 if the original image is RGB to simulate the presence of grey scale images. Batch normalization [21] and Relu [30] are applied after each convolution layer. Our network is trained from scratch without any pretrained weight. Weights on ResBlocks are initialized with Xavier [13] initialization and the rest are initialized with Gaussian distribution using zero mean and 0.01 standard deviation. Adam [22] optimizer is used because it shows faster convergence on this task, learning rate is set to $1e-4$ with $1e-7$ weight decay and batch size is set to 16. All of the experiments are implemented by TensorFlow [1] framework.

To evaluate performance, previous works [4, 20, 24, 29] use MAE (Mean Absolute Error) and MSE (Mean Square Error) as evaluation metrics, defined as follows:

$$MAE = \frac{1}{M}\sum_{i=1}^{M}|Pred_i - Cnt_i| \tag{9}$$

$$MSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}|Pred_i - Cnt_i|^2} \tag{10}$$

where M is total images, $Pred_i$ is prediction count integrated from predicted density map and Cnt_i is ground truth count.

Different from [4, 24, 35], which use the patch-based testing scheme. We empirically find out that directly input the whole image during testing to our network resulting in better performance.

4.3. Results

Table 2 shows the result of our method versus the state-of-the-art evaluated on ShanghaiTech dataset. Our DECCNet (last row) achieves the lowest MAE of 58.6 in harder part A with higher density and significantly outperforms the state-of-the-art, with 6% better, indicating our proposed method, as mentioned in section 3.3, can deal with high density regions. In part B with a relatively lower density of crowds, our approach also reaches a new state-of-the-art result, with 7% better. Figure 4 and figure 5 present the density estimation examples of part A of ShanghaiTech.

Table 3 compares the result of DECCNet versus several methods evaluated on UCF-QNRF dataset, which is the largest, newest and most challengeable dataset available. In this dataset our method achieves the lowest MAE of 107.9, outperforming other methods by a large margin.

4.4. Ablation studies

To verify the effectiveness of each component of our architecture, we conduct ablation studies on both part of ShanghaiTech dataset, containing several different settings: 1) replace MegaDepth with MonoDepth, 2) remove depth information from input, resulting in a one stream Resnet encoder, where decoder remains the same, 3) remove BCA,

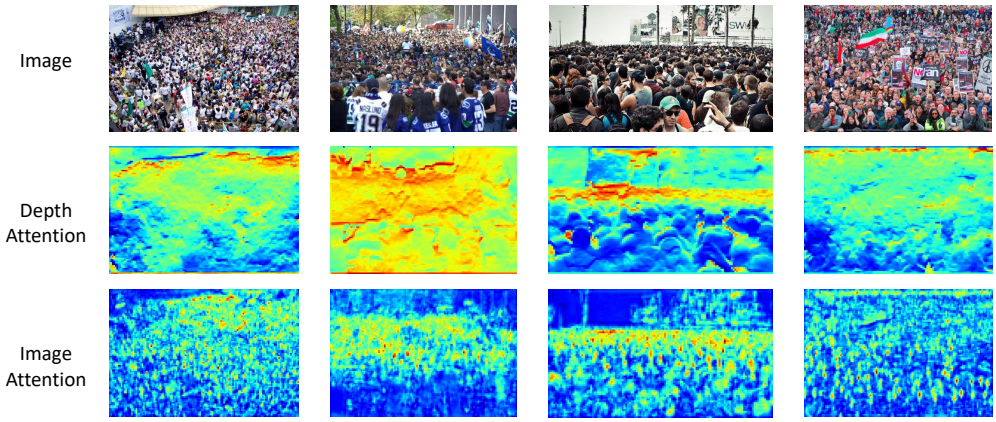

Figure 7. Attention map of the proposed novel BCA component. The first row is the original image, the second row is the attention map of depth stream after third ResBlock, and the third row is the attention map of RGB stream after third ResBlock. We can see that in the second row, the proposed BCA component can attend on those deep regions, while the third row demonstrates the information flow from RGB to depth stream, guiding it to attend on high density regions.

so that two streams cannot affect each other in middle layers, 4) remove self-attention, which is used to catch long-range dependency and 5) remove DUC, where bilinear upsampling is used instead.

As shown in table 4, dropping each component of our network leads to performance decrease. For example, directly combining depth as the fourth channel leads to a significant performance drop, possibly due to the distribution difference between two modalities (figure 1). And if BCA component is removed from the encoder, MAE/MSE of both part of ShanghaiTech dataset also increases, indicating the importance of designing a proper fusing method. Besides, if we replace the depth estimation model with MonoDepth, our model can still reach a reasonable performance, indicating the generalization of the proposed method.

4.5. Visualization

In addition to quantitative result, we also present qualitative result on figure 4, 5, 6 and 7. Figure 4 shows the results that are most accurate among ShanghaiTech dataset. Congested regions on those images are far away from the camera, which are deep as well. Furthermore, depth prediction of these images are roughly correct, which is beneficial to the model.

Figure 5 shows the failure cases of our method. Although the appearances on the predicted density map are roughly the same as ground truth, final counts are still inaccurate. Specifically, extracted depth information of those images is incomplete or noisy, leading to performance degradation. In other words, our method quite depends on the quality of the depth channel.

Figure 6 shows the comparison versus state-of-the-art. CSRNet [24] tends to fail on those congested regions while our method produces finer estimation.

Figure 7 presents the effectiveness of BCA component. Since BCA is a bidirectional attention mechanism and takes place in each of three ResBlocks, we decide to visualize the attention map of each stream after third ResBlock, where high level feature map contains more semantic and global meaning of input. Designing intention of this component is to interactively enhance RGB stream by depth stream, and vice versa. Depth attention, as shown in the second column of figure 7, successfully attends on hard cases, most of which are deep and congested. RGB attention, shown in the third column, guides the other stream to focus on high density region to a certain degree.

5. Conclusion

In this paper, we introduce a novel depth enhanced crowd counting network to accurately estimate crowd density of the given image, especially highly congested regions. Leveraging depth information along with RGB data provides an extra capability for our model to pay attention to deep regions. We demonstrate our method with state-of-the-art and reach the best performance. Experiments show the necessity of BCA and the depth channel. Since depth information we used is an estimation, if accurate depth is available, our performance can be further improved.

6. Acknowledgement

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Grant MOST 108-2634-F-002-004, FIH Mobile Limited, and Qualcomm Technologies, Inc., under Grant NAT-410477. We also benefit from the NVIDIA grants and the DGX-1 AI Supercomputer.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Deepak Babu Sam, Neeraj N Sajjan, R Venkatesh Babu, and Mukundhan Srinivasan. Divide and grow: capturing huge diversity in crowd images with incrementally growing cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3618–3626, 2018.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

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

[5] Antoni B Chan and Nuno Vasconcelos. Bayesian poisson regression for crowd counting. In *2009 IEEE 12th international conference on computer vision*, pages 545–551. IEEE, 2009.

[6] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *BMVC*, volume 1, page 3, 2012.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 801–818, 2018.

[8] Diptodip Deb and Jonathan Ventura. An aggregated multicolumn dilated convolution network for perspective-free counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 195–204, 2018.

[9] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2012.

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014.

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

[12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[13] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[14] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. *arXiv:1806.01260*, 2018.

[15] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *European Conference on Computer Vision*, pages 345–360. Springer, 2014.

[16] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE international conference on Robotics and automation (ICRA)*, pages 1524–1531. IEEE, 2014.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[19] Haroon Idrees, Imran Saleemi, Cody Seibert, and Mubarak Shah. Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2547–2554, 2013.

[20] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 532–546, 2018.

[21] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 239–248. IEEE, 2016.

[24] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1091–1100, 2018.

[25] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[26] Zhe Lin and Larry S Davis. Shape-based human detection and segmentation via hierarchical part-template matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(4):604–618, 2010.

[27] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5162–5170, 2015.

[28] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE transactions on pattern analysis and machine intelligence*, 38(10):2024–2039, 2016.

[29] Xialei Liu, Joost van de Weijer, and Andrew D Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7661–7669, 2018.

[30] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[31] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[32] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 270–285, 2018.

[33] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4031–4039. IEEE, 2017.

[34] Max Schwarz, Anton Milan, Arul Selvam Periyasamy, and Sven Behnke. Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, 37(4-5):437–451, 2018.

[35] Zan Shen, Yi Xu, Bingbing Ni, Minsi Wang, Jianguo Hu, and Xiaokang Yang. Crowd counting via adversarial cross-scale consistency pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5245–5254, 2018.

[36] Miaojing Shi, Zhaohui Yang, Chao Xu, and Qijun Chen. Revisiting Perspective Information for Efficient Crowd Counting. *arXiv e-prints*, page arXiv:1807.01989, Jul 2018.

[37] Zenglin Shi, Le Zhang, Yun Liu, Xiaofeng Cao, Yangdong Ye, Ming-Ming Cheng, and Guoyan Zheng. Crowd counting with deep negative correlation learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5382–5390, 2018.

[38] Vishwanath A Sindagi and Vishal M Patel. Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting. In *2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2017.

[39] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1861–1870, 2017.

[40] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[41] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1451–1460. IEEE, 2018.

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[43] Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 833–841, 2015.

[44] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[45] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 589–597, 2016.