HA-CCN: Hierarchical Attention-based Crowd Counting Network

Vishwanath A. Sindagi, *Student Member, IEEE* and Vishal M. Patel, *Senior Member, IEEE*

***Abstract*—Single image-based crowd counting has recently witnessed increased focus, but many leading methods are far from optimal, especially in highly congested scenes. In this paper, we present Hierarchical Attention-based Crowd Counting Network (HA-CCN) that employs attention mechanisms at various levels to selectively enhance the features of the network. The proposed method, which is based on the VGG16 network, consists of a spatial attention module (SAM) and a set of global attention modules (GAM). SAM enhances low-level features in the network by infusing spatial segmentation information, whereas the GAM focuses on enhancing channel-wise information in the higher level layers. The proposed method is a single-step training framework, simple to implement and achieves state-of-the-art results on different datasets.**

Furthermore, we extend the proposed counting network by introducing a novel set-up to adapt the network to different scenes and datasets via weak supervision using image-level labels. This new set up reduces the burden of acquiring labour intensive point-wise annotations for new datasets while improving the cross-dataset performance.

***Index Terms*—crowd counting, weakly supervised learning, crowd analytics**

I. Introduction

THE task of crowd counting is riddled with various challenges such as perspective distortion, extreme scale variations, heavy occlusion, illumination changes, clutter, non-uniform distribution of people, etc. Due to these issues, crowd counting and density estimation is a very difficult problem, especially in highly congested scenes. Several recent convolutional neural network (CNN) based methods for counting [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12] have attempted to address one or more of these issues by adding more robustness to scale variations by proposing different techniques such as multi-column networks [2], intelligent selection of regressors suited for a particular crowd scenario [3] and incorporating global, local context information into the counting network [4], *etc*. Methods such as [3, 4, 8] achieve significantly lower errors compared to the earlier approaches, however, they are complex to train due to the presence of multiple learning stages. For instance, Switching-CNN [3] involves different stages such as pre-training, differential training, switch training and coupled training. Similarly, CP-CNN [4] requires that their local and global estimators to be trained separately, followed by end-to-end training of their density map estimator. Although the most recent methods such as [6, 7, 10] achieve better results while being efficient, there is still considerable room for further improvements.

V. A. Sindagi and V. M. Patel are with the Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD, 21218 USA e-mail: (vishwanathsindagi@jhu.edu, vpatel36@jhu.edul).

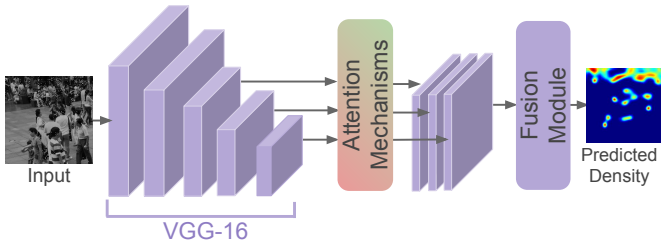


Fig. 1. Overview of the proposed Hierarchical Attention-based Crowd Counting Network (HA-CCN).

In this paper, we propose to improve the counting performance by explicitly modeling spatial pixel-wise attention and global attention into the counting network. Considering that crowd images have large variations in head sizes, it is essential to leverage multi-scale information by employing feature maps from different conv layers of the VGG16 network [13]. Several works such as [14, 15, 16, 17] have demonstrated that different sized objects are captured by different layers in a deep network. Hence, an obvious approach would be to design a multi-scale counting network [18, 19] that concatenates feature maps from different layers of the VGG16 network. However, earlier layers in a deep network capture primitive features and do not learn semantic awareness. Due to this, naive concatenation of feature maps from different layers of the network is not necessarily an optimal approach to address the issue of large scale variations in crowd images.

To address this issue, we introduce a spatial attention module (SAM) in the network, that is designed to infuse semantic awareness into the feature maps. This module takes the feature maps from lower layers as input, and learns to perform foreground-background segmentation. Furthermore, it uses this learned segmentation map to enhance the lower layer feature maps by selectively attending to specific spatial locations in this lower layer. Furthermore, we also attempt to augment channel-wise information in the higher level layers by employing a set of global attention modules. These modules selectively enhance important channels while suppressing the unnecessary ones. Fig. 1 provides an overview of the proposed attention-based feature concatenation for multi-scale crowd counting.

In addition to improving the count performance, another major issue in the crowd counting research community is the poor generalization performance of the existing networks. This is due to the fact that CNN-based methods are highly data-driven and suffer from inherent dataset bias. Hence, they cannot be applied directly to new scenes without further fine-tuning. A simple solution to this would be to train the

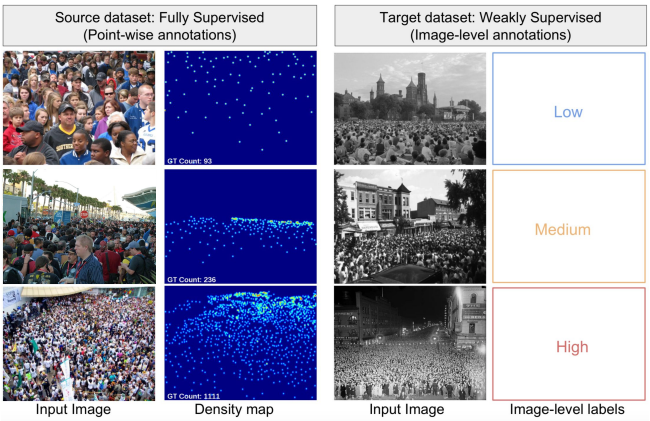


Fig. 2. Target dataset adaptation. (a) Source dataset with point-wise annotations is used to train the counting network. (b) Target dataset with only image-level annotations is used to fine-tune the pre-trained counting network.

model on the target dataset in a fully-supervised fashion, which requires expensive ground-truth annotations. Several earlier works such as [1, 5] address this issue by proposing different semi-supervised or unsupervised fine-tuning methods in addition to their novel network designs. For instance, Zhang *et al.*[1] presented a cross-scene counting approach where they use perspective maps to retrieve candidate scenes from source dataset that are similar to the target set, which are then used to fine-tune the network. However, perspective maps may not be always available. Additionally, it is dependent on the assumption that their pre-trained model provides good estimates of count in the target patches. Liu *et al.*[5] proposed a self-supervised method based on image ranking to adapt to different datasets. While it achieves better generalization performance, their method is still limited since they use only unlabeled data.

To address this generalization issue, we take a different approach as compared to earlier attempts ([1, 5]) by proposing a novel weakly supervised learning setup. We leverage image-level labels, which are much easier to obtain as compared to point-wise annotations [1], in a weakly supervised fashion for fine-tuning networks to newer datasets/scenes. To achieve this weak supervision, we use the idea of image-level labeling of crowd images into different density levels by Sindagi *et al.*[4] and Fu *et al.*[20]. While these methods [4, 20] employ image-level labels in conjunction to point-wise annotations to train their networks, we propose to use only image-level labels in the weakly supervised setup while adapting to new datasets, thereby avoiding the labour intensive point-wise annotation process. Fig. 2 illustrates the different types of annotations used for training the network. Fig. 2(a) represents samples from a source dataset, which consists of images and corresponding point-wise ground-truth annotations. The source dataset is used to pre-train the counting network. Fig. 2(b) represents samples from the target set to which we intend to adapt the pre-trained counting network. The pre-trained network is then fine-tuned on the target dataset using image-level labels via the proposed weakly supervised approach. During testing, the estimated density map from the fine-tuned network is compared with the the ground-truth using standard metrics as described later.

[1]Crowd counting datasets are usually provided with point-wise (x,y) location annotations, which are converted to pixel-wise density maps.

To summarize, the following are our key contributions in this paper:

- A new network design that employs attention mechanisms at various levels for selectively enhancing the features from different layers of VGG16 network to increase the effectiveness of multi-scale concatenation.
- A novel setup to adapt existing crowd counting models to new scenes and datasets via weakly supervised learning. To the best of our knowledge, this is the first attempt to perform weak supervision using image-level labels for crowd counting.

In the following sections, we discuss related work (Section II) and the proposed method in detail (Section III). Details of experiments and results of the proposed method along with comparison on different datasets are provided in Section V, followed by conclusions in Section VI.

II. Related work

Crowd Counting. Early approaches for crowd counting are based on hand-crafted representations and different regression techniques [21, 22, 23, 24, 25, 26, 27]. A comprehensive survey of these early methods can be found in [23, 28, 29]. Recent focus in the crowd counting community has been towards exploiting the advances in CNN-based methods and in this attempt, methods such as [1, 2, 3, 3, 4, 30, 31, 32, 33, 34, 35, 36, 36, 37, 38, 39, 40] have demonstrated significant improvements over the traditional methods. Majority of the existing work is focused on addressing the problem of large scale variations in crowd images through different techniques such as multi-resolution network [33], multi-column networks [2], selective regression [3], context-aware counting [38].

Babu *et al.*[8] proposed an automatically growing CNN to progressively increase the capacity of the network based on the dataset. Shen *et al.*[7] used adversarial loss similar to [4] to attenuate blurry effects in the estimated density maps. Shi *et al.*[6] proposed deep negative correlation based learning of more generalizable features. In another interesting approach, Liu *et al.*[5] proposed to leverage unlabeled data for counting by introducing a learning to rank framework. Li *et al.*[41] proposed CSR-Net, that consists of two components: a front end CNN-based feature extractor and a dilated CNN for the back-end. Ranjan *et al.*[10] proposed a network with two branches that estimates density map in a cascaded fashion. Cao *et al.*[9] proposed an encoder-decoder network with scale aggregation modules. They use a combination of Euclidean loss and a newly introduce local pattern consistency loss to train their network. Idrees *et al.*[42] proposed a new large-scale crowd dataset with 1.25 million annotations, along with a novel loss function to train their dense-net [43] based architecture.

Recently, a few approaches have been proposed that incorporate detection of crowded regions through different

techniques such as attention injective deformable network [44], semantic prior-based residual regression [45], use of auxiliary task such as segmentation [46] and segmentation infusion via inverse attention [47]. Other techniques such as [48] exploit multi-scale features in different ways. For instance, Jiang *et al.*[48, 49, 50] propose a Trellis style encoder decoder, where the multi-scale feature maps in the decoder are combined in an effective way. Similarly, Shi *et al.*and Liu *et al.*[49, 50] exploit multi-scale features by explicitly considering perspective and context information respectively. Zhang *et al.*[51] address the issue of wide area counting by proposing a multi-view fusion CNN. Wang *et al.*[52] presented a new large-scale, diverse synthetic dataset and proposed a SSIM based CycleGAN [53] to adapt the synthetic datasets to real world dataset.

Attention mechanisms. Inspired by the role of attention in human visual perception [54, 55, 56], several works have successfully incorporated attention mechanism to improve the performance of CNNs for a variety of tasks such as image captioning [57, 58], visual question answering [59, 60, 61, 62], pose-estimation [63], classification [64, 65, 66], detection [67], fine-grained recognition [68], sequence to sequence modeling [69] *etc*. Xu *et al.*[62] were among the first to introduce visual attention in image captioning where they use different pooling mechanisms that attend to important and relevant regions in the scene. Zhu et al. [70] employed soft attention to combine image region features for the task of VQA. For the same task, Yang et al. [59] and Xu et al. [62] employed multiple stacked spatial attention models, in which the spatial attention map is successively refined. Chu *et al.*. [63] used attention to guide multi-contextual representation learning for improving pose estimation performance. Wang et al. [71] improved classification performance by incorporating 3D attention maps, generated using hour glass modules, into residual networks. Hu et al. [72] proposed a compact Squeeze-and-Excitation (SE) module to leverage inter-channel relationships. Recently, Woo *et al.*[64] extended the work of [66] by employing spatial and channel-wise attention modules after every layer in the network. In a different application, Zhu *et al.* [69] employed seq2seq in their decoder structure to model temporal video sequences. The attention mechanism at each step is used to help the decoder to decide which frames in the input sequence might be related to the next frame reconstruction. Note that they use attention mechanism to perform input selection. In contrast, we use attention mechanism to select relevant and important features and additionally refine them.

The closest methods to our work are [58, 64, 69]. There are several notable differences as compared to our method. First, these methods employ a sequence of channel-wise and spatial attention after every convolutional layer to refine the feature maps. In contrast, we specifically insert a spatial attention module after the conv3 block. By doing this, we are able to infuse the attention early into forward process and hence, such a module is rendered unnecessary after every block for an application like crowd counting. Second, the spatial attention modules in the existing works are self-supervised. Different from them, we employ a mask guided spatial attention module that is explicitly supervised using foreground/background masks, resulting in faster learning of the spatial attention. Further, the global attention modules are inserted only after conv4 and conv5 blocks. These different modules are carefully added at specific blocks in the network, thereby avoiding unnecessary over-parameterization.

Weak Supervision. Weakly supervised learning has been extensively used for various problems in computer vision such as semantic segmentation [73, 74, 75, 76, 77, 78], object localization [79, 80, 81, 82], saliency detection [83, 84], scene recognition [85, 86] and many more. However, this form of learning has been relatively unexplored for crowd counting. Liu *et al.*[87] proposed a solution based on Bayesian model adaptation of Gaussian processes for transfer learning, which is limited only to the GP model of [88]. Recently, Borstel *et al.*[89] proposed a method to count objects within an image from only region-level count information. Though the problem is defined in a weakly supervised setting, they require local region level count as annotations which is a labour intensive process.

In this work, we introduce a novel weakly-supervised learning setup that employs image-level labels to generate pseudo ground-truth, which is further used to fine-tune the counting network. The use of pseudo ground-truth generation is inspired from semi-supervised learning approaches such as [90, 91, 92, 93, 94, 95]. These approaches typically use labeled data in the dataset to train a predictor. The trained predictor is then used to generate predictions for the unlabeled data. The highly confident predictions are then used as ground-truth to further fine-tune the data. In contrast to this framework, we propose a weakly supervised approach where the pseudo-ground truth is generated from weak image-level labels.

III. Hierarchical attention for crowd counting

As discussed earlier, a natural solution to address scale variation in crowd counting images is to leverage multi-scale features from different layers in the backbone network. However, the layers in the backbone network are learned in a hierarchical manner with the earlier layers capturing primitive features and the subsequent layers capturing higher level concepts. Hence, direct fusion of these multi-scale feature maps might not be the most effective approach. In order to overcome this, we propose Hierarchical Attention-based Crowd Counting Network (HA-CCN) that leverages attention mechanisms to enrich features from different layers of the network for more effective multi-scale fusion.

Fig. 3 provides an overview of the proposed method, which is based on the VGG-16 network. We include a spatial attention module (SAM) and a set of global attention modules (GAM) with the objective of enriching the feature maps at different levels. The base network consists of conv layers (conv1 $\sim$ conv5) from the VGG16 network. The conv3 features are enhanced by passing them through SAM. Similarly, features from conv4 and conv5 are passed through GAMs in order to perform channel-wise enhancement. The enhanced feature maps from conv3 are

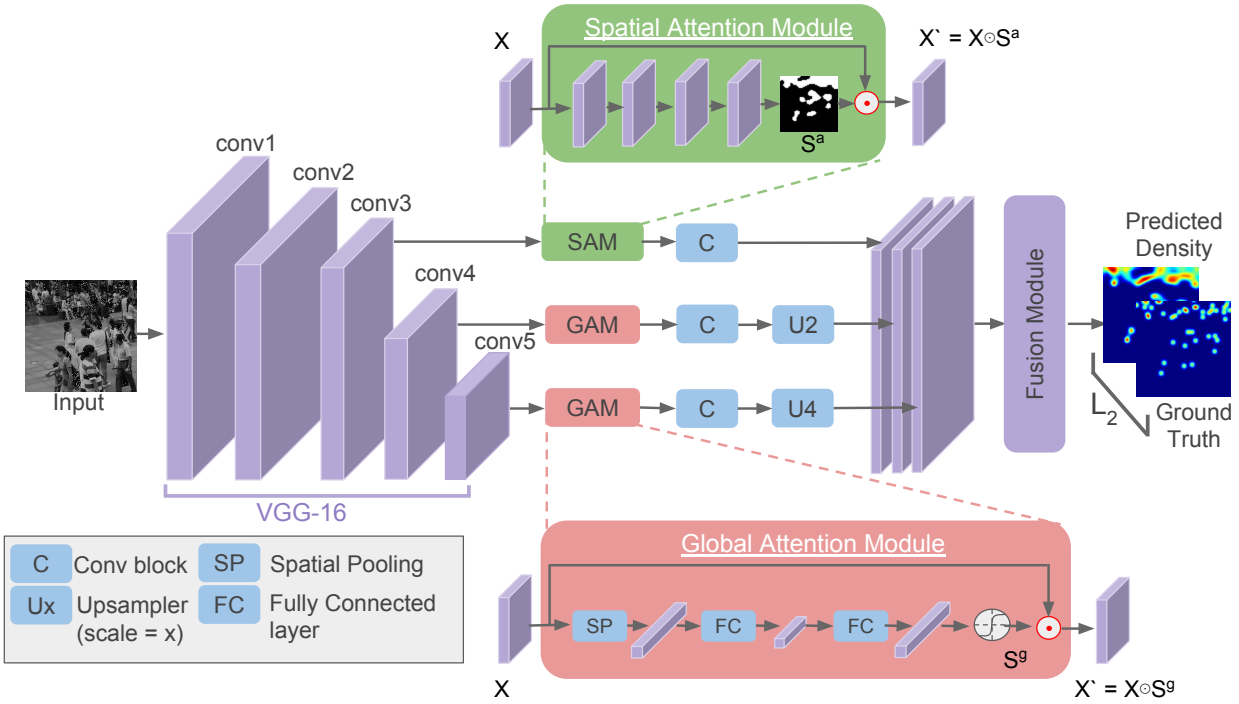


Fig. 3. Overview of the proposed Hierarchical Attention-based Crowd Counting Network (HA-CCN). VGG16 is used as the base network. Feature maps from conv3 are forwarded through a spatial attention module that incorporates pixel-wise segmentation information into the features. Feature maps form higher layers (conv4, conv5) are forwarded through a set of global attention modules that augment the feature maps along the channel dimension.

then forwarded through a conv block which consists of 3 conv layers defined as follows: $\{\text{Conv2d}(256,64,1)^3\text{-ReLU}, \text{Conv2d}(64,64,3)^3\text{-ReLU}, \text{Conv2d}(64,24,1)^3\text{-ReLU}\}$.

Similarly, the enhanced features from conv4 and conv5 are forwarded through a conv block and an upsampling layer to scale the feature maps to a size similar to that of conv3 feature maps. The conv block is defined by: $\{\text{Conv2d}(512,64,1)^3\text{-ReLU}, \text{Conv2d}(64,64,3)^3\text{-ReLU}, \text{Conv2d}(64,24,1)^3\text{-ReLU}\}$.

These processed features are concatenated together before being forwarded through the fusion module that consists of a set of conv layers to produce the final density map. These conv layers are defined by: $\{\text{Conv2d}(72,64,1)^3\text{-ReLU}, \text{Conv2d}(64,64,3)^3\text{-ReLU}, \text{Conv2d}(64,1,1)^3\text{-ReLU}\}$. The network is trained by minimizing the Euclidean distance between the predicted density map and the ground truth density map as below:

$$L_d = \frac{1}{N}\sum_{i=1}^{N} \|F_d(X_i,\Theta) - D_i\|_2, \qquad (1)$$

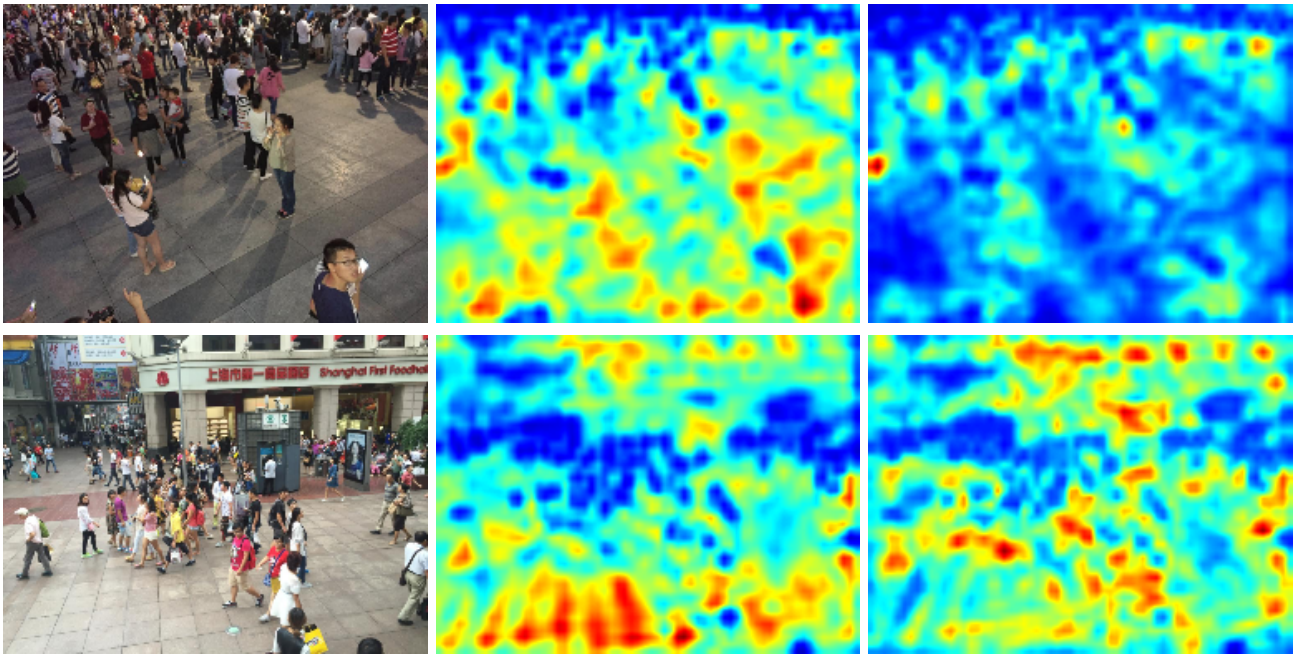
where, N is number of training samples, X_i is the i^{th} input image, $F_d(X_i,\Theta)$ is the estimated density, D_i is the i^{th} ground-truth density and it is calculated by summing a 2D Gaussian kernel centered at every person's location x_g as follows: $D_i(x) = \sum_{x_g \in S} \mathcal{N}(x - x_g, \sigma)$, where σ is scale parameter of 2D Gaussian kernel and S is the set of all points where people are located. The density map generated by the network is $1/4^{th}$ of the input image resolution. Due to its construction, the sum of the density map provides an estimate of the number of people in the input image.

Details of the proposed method and its various components are described in the following sub-sections.

A. *Spatial attention module*

Inspired by the success of spatial attention mechanisms in image captioning, visual question answering and classification [58, 62, 68], we explore its utilization for crowd counting. The goal of spatial attention is to select attentive regions in the feature maps, which are then used to dynamically enhance the feature responses. In contrast to existing work that learn spatial attention in a self-supervised manner, we propose to learn this by explicitly using foreground background segmentation for supervision. Since the goal of the spatial attention module is to focus on relevant regions and foreground regions are necessarily a part of these relevant regions, it is beneficial to use these labels to supervise the module. By explicitly supervising the module, we are able to infuse foreground background information into the network, thereby forcing the network to focus on relevant regions among the foreground. Moreover, these labels are readily available and hence, it does not require additional annotation efforts.

The spatial attention module consists of 4 conv layers with 3×3 filters that takes feature maps from the conv3 layer of the VGG16 network as input (denoted by $X \in \mathbb{R}^{W\times H\times C}$), and produces a segmentation output $S^a \in [0,1]^{W\times H}$. The segmentation map is then used to actuate the low level feature map X via element-wise multiplication: $\hat{X} = X \odot S^a$, where $\hat{X}$ is the actuated low level feature map from conv3. Through this attention mechanism, we are able to incorporate segmentation awareness into the low level feature maps. As illustrated in Fig. 4, the use of segmentation information into the network enriches feature maps by suppressing irrelevant regions and boosting the foreground regions. The actuated feature maps are then forwarded to the fusion block (FM),



(a) (b) (c)

Fig. 4. Visualization of the conv3 feature maps: (a) Input image (b) Before segmentation infusion (c) After segmentation infusion. By infusing segmentation information into the counting network, we are able to suppress background regions. Note that in the density maps, red color indicates high density and blue color indicates low density.

where they are fused with the features from other layers to generate the final density map.

The weights of SAM are learned by minimizing the cross entropy error between the predicted segmentation map and the corresponding ground-truth. Normally, the segmentation task requires pixel-wise annotations. However, in this case existing ground truth density map annotations are thresholded to generate the ground truth segmentation maps, which are then used to train the spatial attention module. Basically, the pixels that contain head regions are labeled as 1 (foreground), and otherwise as 0 (background). Hence, the proposed method does not require any additional labeling. In spite of these annotations being noisy, the use of segmentation information results in considerable gains.

B. Global attention modules

In contrast to the spatial attention module that attends to relevant spatial locations in the feature maps of low-level layers, the global attention module (GAM) is designed to attend to feature maps in the channel-dimension. The global attention module is similar to the channel-wise attention used in earlier work like [58, 64]. Specifically, this module consumes feature maps from the backbone network and learns to compute attention along the channel dimension. The computed attention captures the important channels in the feature maps and hence aids in suppressing information from unnecessary channels. Since this module operates at a global level in terms of spatial dimension, we refer to this attention module as global attention module. It has been demonstrated in [58, 68, 72], that channels capture the presence either different parts of an object or different classes of objects and channel-wise attention is an effective way to boost the correlation between object/object parts and image captions.

Based on these considerations, we employ a set of global attention modules, which take feature maps from the higher conv layers as input and produce a channel-wise attention map, which is then used to actuate the feature maps along the channel dimension. Mathematically, given a feature map input $X \in \mathbb{R}^{W \times H \times C}$, GAM first performs a spatial pooling to produce pooled features $Y \in \mathbb{R}^{1 \times 1 \times C}$ using

$$Y_i = \frac{1}{W \times H} \sum_{w,h} X_i^{wh}, \tag{2}$$

where i is the channel index, and w, h are spatial indices. Y is passed through a set of fully-connected (FC) layers defined by $FC(512, 64) - ReLU - FC(64, 64) - ReLU - FC(64, 512)$[2] and a sigmoid layer to produce channel-wise attention vector $S^g \in \mathbb{R}^{1 \times 1 \times C}$. Finally, S^g is used to actuate the feature maps from the higher layer by performing a element-wise multiplication, *i.e.,*$\hat{X} = X \odot S^g$.

IV. Weak supervision via image-level labels

As discussed earlier, existing methods [1, 2] recognize the inability of these networks to generalize well to different datasets. Their solutions to improve the cross-dataset performance is through fine-tuning in either a fully-supervised or semi-supervised fashion. In contrast to these approaches, we propose a weakly supervised setup to train the counting networks on the new datasets with just image-level labels. Such a setup will simplify the training process as it does not

[2] FC_{N_i,N_o} denotes fully connected layer (with N_i input elements, N_o output elements)

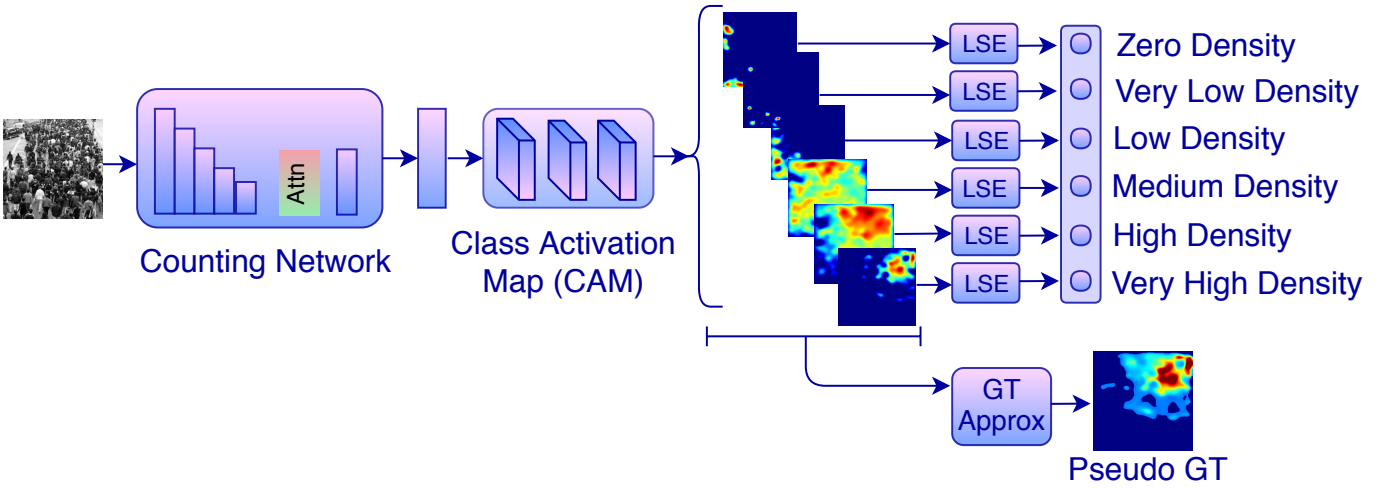


Fig. 5. Overview of the proposed weakly supervised learning for adapting counting network to new datasets. A class activation map (CAM) module is learned to produce class-wise score maps using image-level labels, which are further used to estimate pseudo ground-truth density maps for target set images.

require point-wise annotations which are labour intensive and expensive to obtain.

The idea of performing weakly supervised crowd counting is largely inspired by the success of recent CNN-based weakly supervised semantic segmentation methods [73, 74, 75] that typically fit the problem into Multiple-Instance Learning framework [96]. In their setup, every image is considered to have at least one pixel corresponding to image class label, and the segmentation task is formulated as inferring the pixels belonging to the object class. These methods usually employ class activation mappings to perform weak supervision. However, crowd counting is a regression problem and cannot be directly fit into such a framework. To overcome this issue, crowd counting is transformed into a crowd-density classification task, i.e., instead of counting the number of people in an image, this task is reformulated into categorizing the image into one of the six classes: $\mathcal{C}$ = {*zero density, very-low density, low density, medium density, high density, very-high density*}. This reformulation is based on the intuition that it is easier to label an image as containing large or few number of people as compared to the exact count. Sindagi *et al.*[4] used a similar concept for leveraging image context. In this work, the labels are used to reformulate the counting problem into a classification task for weakly supervised learning.

Fig. 5 illustrates the proposed weakly supervised approach for adapting to new target scenes or datasets. Similar to semantic segmentation where a pre-trained CNN is used, we use counting network (HA-CCN) described in Section III that is pre-trained on the source dataset. A class activation map module (CAM), consisting of 4 conv layers is added before the fusion module in the counting network. This module is defined as: {Conv2d(72,64,3)-ReLU, Conv2d(64,64,3)-ReLU, Conv2d(64,32,3)-ReLU}. Conv2d(32,6,3)[3]

This sub-network takes in features from the counting network and processes them to produce output with $|\mathcal{C}|$ feature planes, one for each class. That is, the output of CAM is pixel-wise scores for each class and is denoted by $S^c_{i,j}$ at pixel location (i, j) for each class $c \epsilon \mathcal{C}$. Since point-wise labels are not available for the target set, the pixel-wise scores for each class are mapped to a single image-level classification score using an aggregation function F_{agg} such that $s^c = F_{agg}(S^c_{i,j})$. This class-wise (s^c) score is then maximized for the right class label. Different aggregation functions such as Global Average Pooling (GAP) and Global Max Pooling (GMP) [97] have been used in the literature. In case of GAP, all pixels in the score map are assigned the same weights even if they do not belong to image's class label. GMP addresses this by assigning weight to the pixel that contributes most to the score, however the training is slow [75]. Hence, smooth version and convex approximation of the max function is chosen for F_{agg}, called Log-Sum-Exp (LSE) which is defined as:

$$S^c = \frac{1}{r} \log \left[\frac{1}{wh} \sum_{i,j} \left(r S^c_{i,j} \right) \right], \tag{3}$$

where, S^c denotes aggregated score for class c, $S^c_{i,j}$ is pixel-level score at location (i, j) for class c, r is a hyper-parameter that controls the smoothness of approximation, w, h are width and height of the score map. A soft-max function is applied to the aggregated class scores. The CAM module is trained using the standard binary cross entropy loss function. Parameters of the counting network are kept fixed during this training. The class-wise score maps obtained from the above procedure indicate regions/pixels in the image that belong to a particular density level and hence can be viewed similar to class activation maps [98] (see Fig. 6). These class-wise maps are then used to approximate the pseudo ground-truth density maps for the target set using:

$$D_{pseudo}(i, j) = \sum_{c \epsilon \mathcal{C}} n(c) \widetilde{S}^c_{i,j}, \tag{4}$$

where, $\widetilde{S}^c_{i,j}$ are obtained by normalizing $S^c_{i,j}$ and $n(c)$ is the average count for class c in the source dataset. The pseudo ground-truth maps (as seen in Fig. 6) are not as sharp as actual ground-truth maps, however, they provide coarse regional density that is better as compared to just image-level labels.

[3] *Conv2d(N_i,N_o,k)* denotes 2d convolutional layer (with N_i input channels, N_o output channels, $k \times k$ filter size)

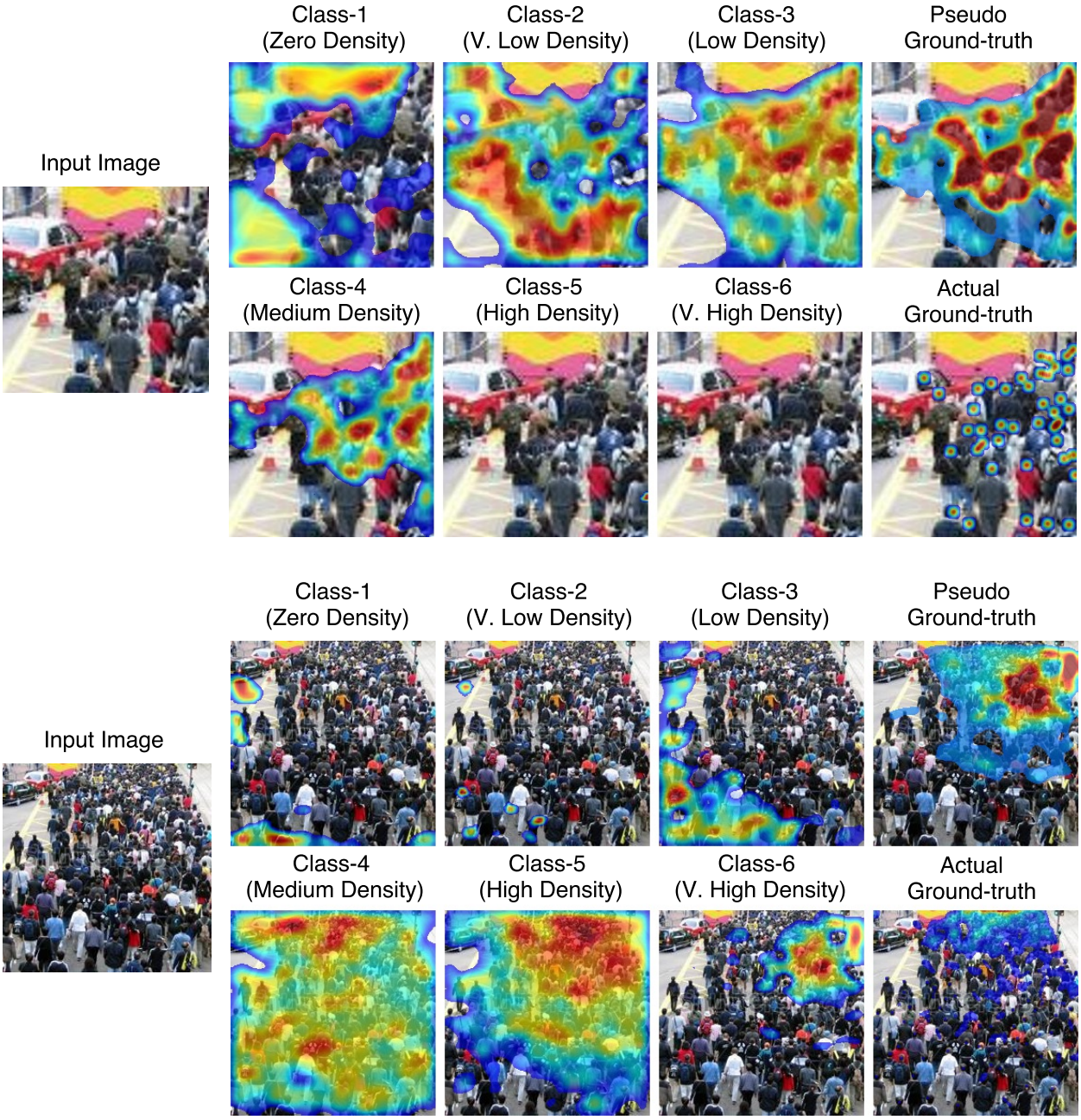


Fig. 6. Example of class-wise score maps overlaid on input images. It can be observed that the CAM module is able to accurately identify regions corresponding to different density levels in an image. We also illustrate pseudo ground-truth estimated using image-level labels. Note that in the density maps, red color indicates high density and blue color indicates low density.

These pseudo ground-truth density maps are used to supervise the counting network on the target dataset. During fine-tuning, weights of the VGG-16 network are fixed and only the weights of the later conv layers are updated. This ensures that the resulting estimated density maps are sharper since the feature maps extracted from VGG-16 preserve details, while the later layers adapt to the newer dataset.

Although the network is trained using image-level labels, it learns to generate density maps for the target set as well. Hence, during inference, test image from the target set is forwarded through the network to estimate the density map. The performance of the proposed weakly supervised technique is measured using standard count error metrics (MAE/MSE).

V. Experiments and Results

A. Hierarchical attention-based counting

In this section, we first describe the training and implementation specifics followed by a detailed ablation study to understand the effects of different components in the proposed counting network. Finally, we compare results of the proposed method against several recent approaches on 3 publicly available datasets (ShanghaiTech [2], UCF-QNRF [42], UCF_CROWD_50 [24]).

1) Training and implementation details: The network is trained end-to-end using the Adam optimizer with a learning

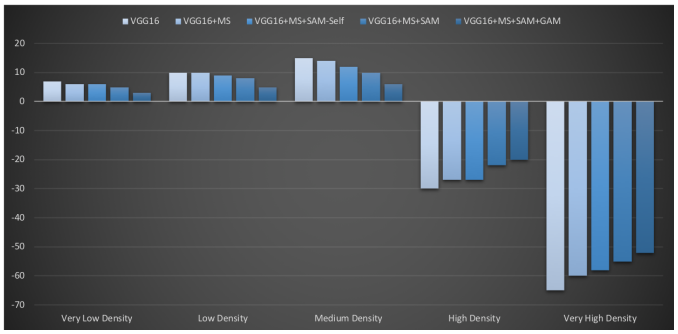


Fig. 7. Ablation study: MAE for different configurations at different density levels.

TABLE I
RESULTS OF THE ABLATION STUDY ON SHANGHAITECH PART A AND PART B DATASETS.

	Part A		Part B	
Configuration	MAE	MSE	MAE	MSE
VGG16	78.3	120.1	18.3	22.9
VGG16+MS	72.1	115.5	15.6	20.6
VGG16+MS+SAM (Self-sup)	69.5	108.2	12.3	20.1
VGG16+MS+SAM	65.1	103.5	10.6	19.6
VGG16+MS+SAM+GAM (HA-CCN)	**62.9**	**94.9**	**8.1**	**13.4**

rate of 0.00005 and a momentum of 0.9 on a single NVIDIA GPU Titan Xp. 10 % of the training set is set aside for validation purpose. The final training dataset is formed by cropping patches of size 224×224 from 9 random locations from each image. Further data augmentation is performed by randomly flipping the images (horizontally) and adding random noise. Similar to earlier work, the count performance is measured using mean absolute error (MAE) and mean squared error (MSE) given by:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|, \quad MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}|y_i - y_i'|^2},$$

where N is the number of test samples, y_i is the ground-truth count and y_i' is the estimated count corresponding to the i^{th} sample.

Since the network is fully convolutional, entire test image is forwarded through the network during inference. This results in faster inference as compared to the existing methods (such as Switching-CNN [3], IG-CNN [8], CP-CNN [4], SA-Net [9]) which involve patch-based testing.

2) Architecture ablation: To understand the effectiveness of various modules present in the network, we perform experiments with the different settings using ShanghaiTech dataset (Part A and Part B). This dataset consists of 2 parts with Part A containing 482 images and Part B containing 716 images and a total of 330,165 head annotations. Both parts have training and test subsets.

The ablation study consisted of evaluating 3 baselines in addition to the proposed method:
(i)*VGG16*: VGG16 network with an additional conv block at the end.
(ii)*VGG16+MS*: VGG16 with multi-scale feature map concatenation and a fusion module at the end.
(iIi)*VGG16+MS+SAM (Self-sup)*: VGG16 with spatial attention module (self-supervised) for conv3 layer and multi-scale feature map concatenation, followed by a fusion module at the end.
(iv)*VGG16+MS+SAM*: VGG16 with spatial attention module for conv3 layer and multi-scale feature map concatenation, followed by a fusion module at the end.
(v)*VGG16+MS+SAM+GAM (HA-CCN)*: proposed method.

The results of these experiments are tabulated in Table I. It can be observed that the naive approach of performing multi-scale feature concatenation does not necessarily yield the most optimal performance. The use of SAM infuses segmentation information in to the feature maps of conv3 layer in the base network, resulting in considerable reduction of the count error as compared to the naive approach. The use of global attention results in further improvement, thus showing significance of incorporating channel-wise importance in the network.

Additionally, it can also be noted that the explicitly supervised SAM results in better performance as compared to the self-supervised spatial attention.

Fig. 7 shows a plot of the mean absolute error for different configurations in the ablation study at different density levels. It can be observed that the proposed HA-CCN network achieves best error among all the density levels.

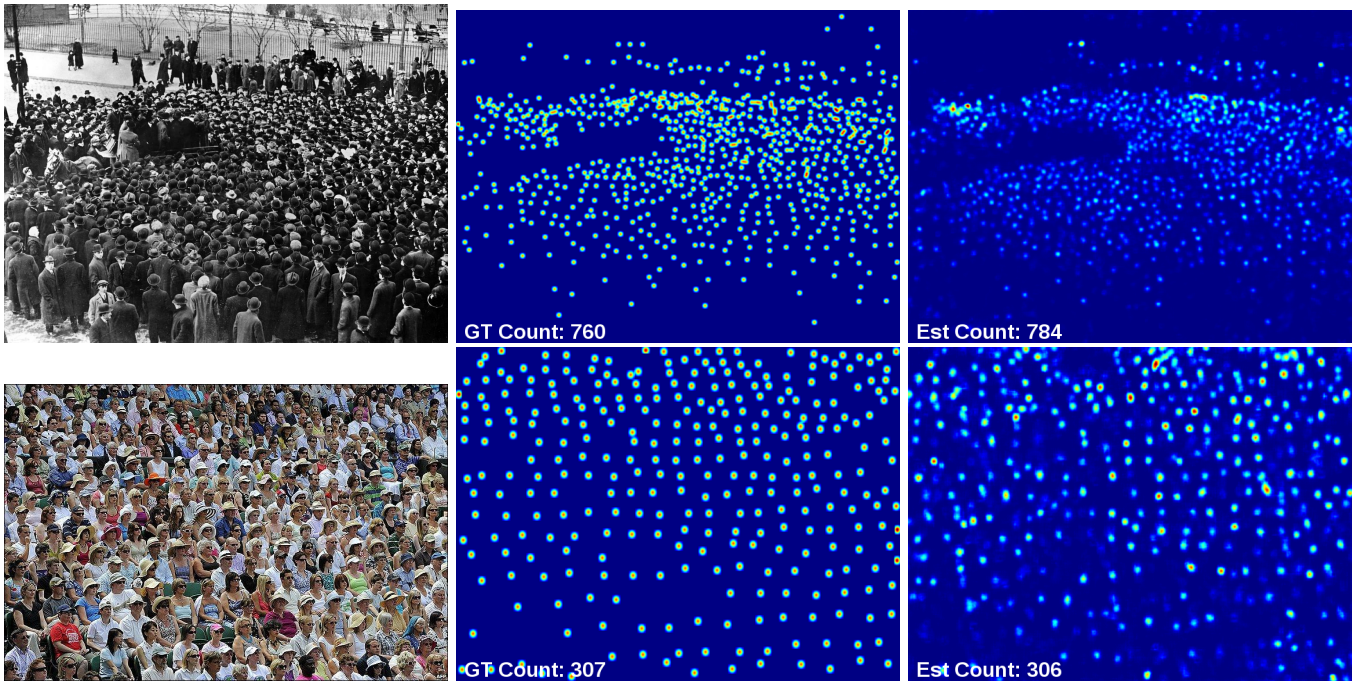
3) Comparison with recent methods: In this section, we discuss the results of the proposed method as compared with recent approaches on 3 different datasets: ShanghaiTech [2], UCF_CROWD_50 [24] and UCF-QNRF [42]. As discussed

TABLE II
COMPARISON OF RESULTS ON THE SHANGHAITECH [2] AND UCF_CROWD_50 [99] DATASETS. TOP TWO METHODS ARE HIGHLIGHTED USING UNDERLINE AND BOLD FONTS RESPECTIVELY. * INDICATES PATCH-BASED TESTING.

	ShTech-A		ShTech-B		UCF-CROWD	
Method	MAE	MSE	MAE	MSE	MAE	MSE
Switch-CNN [3]* (CVPR '17)	90.4	135.0	21.6	33.4	318.1	439.2
CP-CNN [4]* (ICCV '17)	73.6	106.4	20.1	30.1	295.8	**320.9**
IG-CNN [8]* (CVPR '18)	72.5	118.2	13.6	21.1	291.4	349.4
ACSCP [7] (CVPR '18)	75.7	102.7	17.2	27.4	291.0	404.6
CSRNet [41] (CVPR '18)	68.2	115.0	10.6	16.0	266.1	397.5
ic-CNN [10] (ECCV '18)	69.8	117.3	10.7	16.0	260.9	365.5
SA-Net [9]* (ECCV '18)	67.0	104.5	8.4	13.6	**258.5**	**334.9**
IA-DCCN [47]* (AVSS '19)	66.9	108.4	10.2	16.0	264.2	394.4
ADCrowdNet [44] (CVPR '19)	63.2	98.9	**8.2**	15.7	266.4	358.0
RReg [45] (CVPR '19)	**63.1**	**96.2**	8.7	**13.5**	-	-
HA-CCN (ours)	**62.9**	**94.9**	**8.1**	**13.4**	**256.2**	348.4

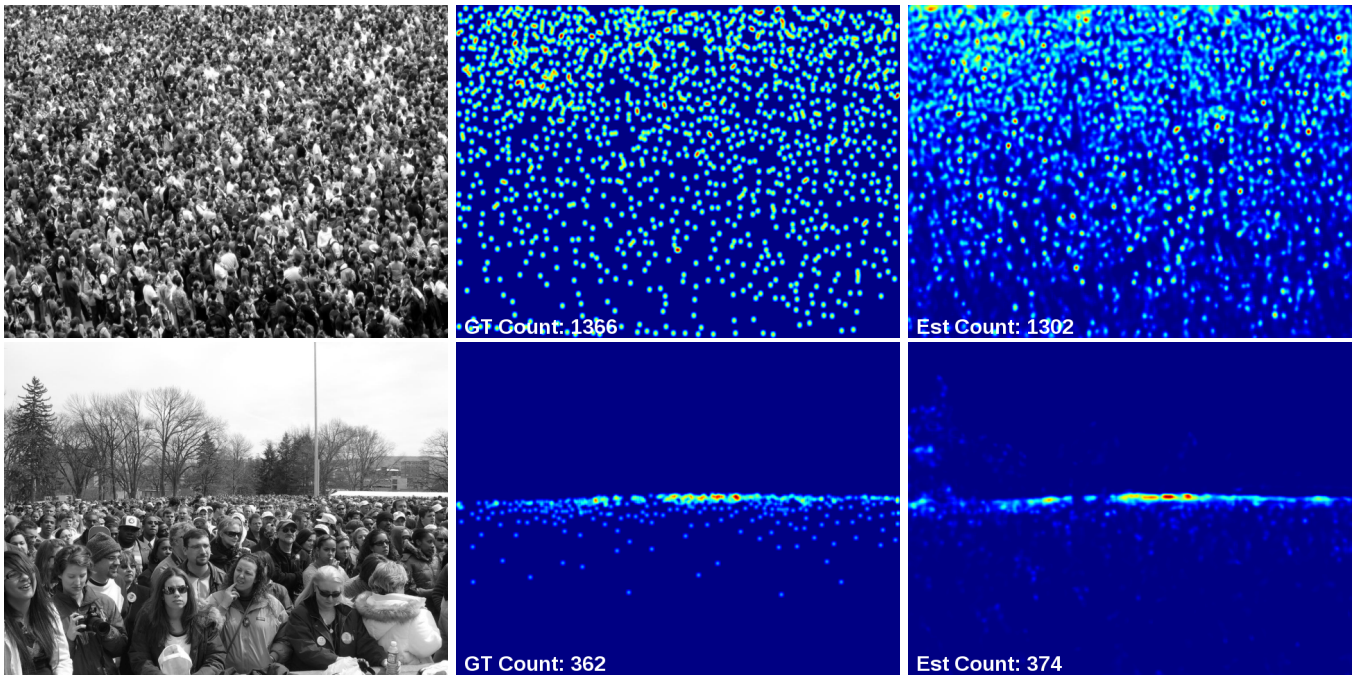
TABLE III
COMPARISON OF RESULTS ON THE UCF-QNRF DATASTET [42]. TOP TWO METHODS ARE HIGHLIGHTED USING UNDERLINE AND BOLD FONTS RESPECTIVELY.

Method	MAE	MSE
Idrees *et al.*[24] (CVPR '13)	315.0	508.0
Zhang *et al.*[1] (CVPR '15)	277.0	426.0
CMTL *et al.*[34] (AVSS '17)	252.0	514.0
Switching-CNN [3] (CVPR '17)	228.0	445.0
Idrees *et al.*[42] (ECCV '18)	132.0	191.0
IA-DCCN *et al.*[47] (AVSS '19)	**125.3**	**185.7**
HA-CCN (ours)	**118.1**	**180.4**



(a) (b) (c)

Fig. 8. Sample results of the proposed method on ShanghaiTech [2] *(a)* Input. *(b)* Ground truth *(c)* Estimated density map.



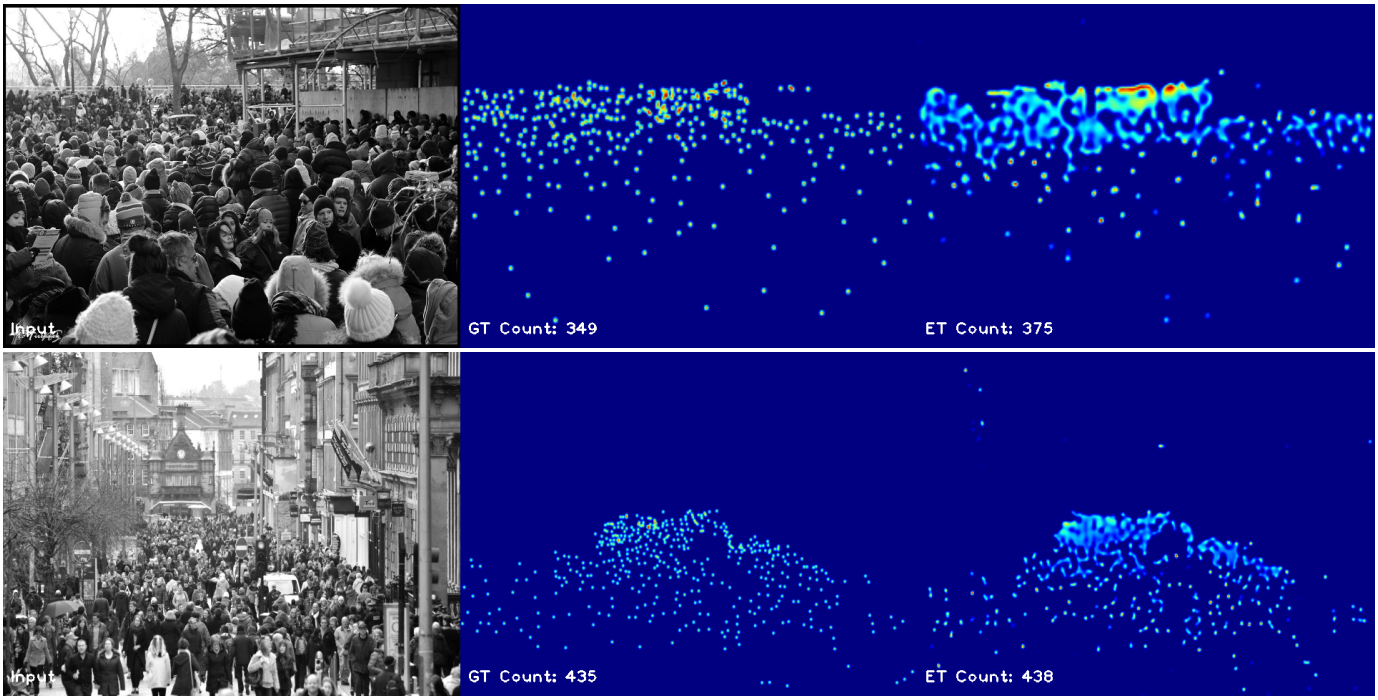
(a) (b) (c)

Fig. 9. Sample results of the proposed method on UCF_CROWD_50 [24]. *(a)* Input. *(b)* Ground truth *(c)* Estimated density map.

earlier, ShanghaiTech has 2 parts with a total of 1198 images. The UCF_CC_50 dataset [24] contains 50 annotated images of different resolutions and aspect ratios. Following the standard protocol discussed in [24], a 5-fold cross-validation is performed for evaluating the proposed method. UCF-QNRF [42] is a more recent dataset that contains 1,535 high quality images with a total of 1.25 million annotations. The training and test sets consists of 1201 and 334 images respectively.

Table II shows the results of the proposed method on the ShanghaiTech and UCF_CROWD_50 datasets as compared with several recent approaches: Switching-CNN [3], CP-CNN [4], IG-CNN [8], D-ConvNet [6], Liu *et al.*[5], CSRNet [41], ic-CNN [10], SA-Net [9], ADCrowdNet [44] and Residual Regression [45]. It can be observed that the proposed method outperforms all existing methods.

Table III shows the comparison of results on the recently released large-scale UCF-QNRF [42] dataset. The proposed method is compared against five different approaches: Idrees



(a) (b) (c)

Fig. 10. Sample results of the proposed method on UCF-QNRF dataset [42]. *(a)* Input. *(b)* Ground truth *(c):* Estimated density map.

TABLE IV
CROSS DATASET PERFORMANCE. S: MODEL IS TRAINED ON TARGET SET, NS: MODEL IS TRAINED ON SOURCE AND TESTED ON TARGET SET. C: DROP IN PERFORMANCE BETWEEN S AND NS.

Method	Target Set				
	ShanghaiTech B		UCF_CROWD_50		WExpo '10
	MAE (S/NS/C)	MSE (S/NS/C)	MAE (S/NS/C)	MSE (S/NS/C)	MAE (S/NS/C)
MCNN [2]	26.4/39.6/13.2	41.3/102.5/61.2	377.6/397.7/20.1	509.1/624.1/115.0	11.6/25.2/13.6
Switch CNN [3]	21.6/59.4/37.8	33.4/130.7/97.3	318.1/1117.5/799.4	439.2/1315.4/876.2	9.4/31.1/21.7
D-ConvNet [6]	18.7/49.1/30.4	26.0/99.2/73.2	288.4/364.0/75.6	404.7/545.8/141.1	-
HA-CCN (ours)	8.1/29.1/21.0	13.4/74.1/60.1	256.2/339.8/83.6	348.4/463.2/114.8	8.5/22.0/13.5

et al.[24], MCNN [2], CMTL [34], Switching-CNN [3] and Idrees *et al.*[42]. It can be observed that the proposed method is able to achieve state-of-the-art results on this complex dataset. Fig. 8, 9 and 10 illustrate the qualitative results for sample images from the ShanghaiTech, UCF_CROWD_50 and UCF-QNFRF datasets respectively.

B. Cross dataset performance

We compare the generalization abilities of the proposed method with that of recent methods (MCNN [2], Switching-CNN [3], D-ConvNet [6]) by testing the network (trained on ShanghaiTech A dataset) on target datasets such as ShanghaiTech B, UCF_CROWD_50 and WorldExpo '10 [1]. The results are presented in Table IV. Note that the other networks are also trained on ShanghaiTech A dataset. The cross-dataset performance is measured using the overall count error (MAE/MSE) and the drop in performance. The drop in performance is the difference between the error of the model trained on the target set and that of the model trained on source set, when tested on target set. It can be observed that the proposed method is relatively more robust to change in dataset distribution as compared to the other methods.

Although the proposed method demonstrates better cross-dataset performance as compared to existing methods, there is considerable gap in the performance as compared to when the network is fully supervised on the target set. We address this issue via the weakly supervised technique described in Section IV.

C. Weakly supervised counting

In this section, we present the experiment details and results of weak supervision setup.

Training. First, a source training set is created that is based on the ShanghaiTech A dataset. The other datasets (ShanghaiTech B, UCF_CROWD_50 and WorldExpo) are used as the target sets. ShanghaiTech A is chosen for creating the source training set since it contains large variations in density, scale and appearance of people across images. The training set is created by cropping multi-scale patches of size 224×224 from 9 random locations. The multi-scale patch extractions increases diversity of the source dataset in terms of count and field of

view. Image-level labels for the source dataset are assigned based on the count in each image in the source set.

The target training set is created by cropping multi-scale patches from 9 random locations from each image. The image-level labels for the target set are obtained based on the count in each image. To compensate for the fact that count values from the target set are used to obtain the image-level labels (which is not practically feasible since the target set is not supposed to have point-wise or count annotations), label noise is added for 15% of the training samples. That is, we randomly changed the labels of 15% of the samples with the neighboring classes. This process of adding label noise simulates human labeling error.

The crowd counting network is first trained on the diverse source dataset using full-supervision by minimizing the loss function described in (1), followed by addition of the CAM module. Weights of the counting network are fixed and the CAM module is trained on the diverse source dataset by minimizing the binary cross entropy between image-level labels and aggregated class scores. This is followed by fine-tuning of the CAM module on the target samples using image-level labels. The class-wise maps from the CAM module are used to generate the pseudo ground-truth density maps for the target samples which are then used to fine-tune the counting network.

Discussion. The results of adapting pre-trained counting model using weak supervision and selective fine-tuning for three target datasets (ShanghaiTech B, UCF_CROWD_50 and WorldExpo) are shown in Table V. For WorldExpo, we average the MAE error over all the five scenes. For weak supervision, following configurations with three different aggregation functions are evaluated:
(1)HA-CCN+W-A: Global Average Pooling (GAP)
(2)HA-CCN+W-M: Global Max Pooling (GMP)
(3)HA-CCN+W-L: Log-Sum-Exponential (LSE)

It can be observed that the proposed WSL setup results in significant improvements in the generalization performance of the network. Among the three aggregation functions for weakly supervised learning, LSE outperforms the other two functions. The results obtained using WSL are comparable to many recent fully supervised techniques such as Hydra-CNN [33], MCNN [2], Walach *et al.*[32], Switching-CNN [3], thus demonstrating the significance of the proposed weak supervision technique.

TABLE V
RESULTS FOR WEAKLY SUPERVISED EXPERIMENTS

Method	Target Set				
	ShanghaiTech B		UCF_CROWD_50		WExpo '10
	MAE	MSE	MAE	MSE	MAE
HA-CCN - NS	29.1	74.1	339.8	463.2	22.0
HA-CCN + W-A	23.1	50.6	320.6	430.6	17.5
HA-CCN + W-M	22.5	51.2	322.2	428.1	17.7
HA-CCN + W-L	21.5	46.1	315.1	420.3	15.9

VI. CONCLUSIONS

In this work, we presented a crowd counting network that consists of different attention mechanisms at various levels in the network. Specifically, the proposed network involves two sets of attention modules: spatial attention and global attention module. The spatial attention module incorporates pixel level attention through a way of foreground background segmentation into the features of the earlier layers of the network. The global attention module incorporates channel-wise importance into the network. Furthermore, we also presented a novel weakly supervised setup to adapt counting models to different datasets using image-level labels. Extensive experiments performed on challenging datasets and comparison with recent state-of-the-art approaches demonstrated the significant improvements achieved by the proposed method.

In the future, we will explore better ways of incorporating features from different layers and extend the current framework to other backbone networks. Additionally, we will explore other forms of weakly supervised and semi-supervised learning approaches to further improve the cross-dataset performance.

ACKNOWLEDGMENT

This work was supported by US Office of Naval Research (ONR) Grant YIP N00014-16-1-3134.

REFERENCES

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

[2] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 589–597.

[3] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] V. A. Sindagi and V. M. Patel, "Generating high-quality crowd density maps using contextual pyramid cnns," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[5] X. Liu, J. van de Weijer, and A. D. Bagdanov, "Leveraging unlabeled data for crowd counting by learning to rank," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[6] Z. Shi, L. Zhang, Y. Liu, X. Cao, Y. Ye, M.-M. Cheng, and G. Zheng, "Crowd counting with deep negative correlation learning," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, and X. Yang, "Crowd counting via adversarial cross-scale consistency pursuit," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[8] D. Babu Sam, N. N. Sajjan, R. Venkatesh Babu, and M. 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, 2018, pp. 3618–3626.

[9] X. Cao, Z. Wang, Y. Zhao, and F. Su, "Scale aggregation network for accurate and efficient crowd counting," in *European Conference on Computer Vision*. Springer, 2018, pp. 757–773.

[10] V. Ranjan, H. Le, and M. Hoai, "Iterative crowd counting," in *European Conference on Computer Vision*. Springer, 2018, pp. 278–293.

[11] D. B. Sam and R. V. Babu, "Top-down feedback for crowd counting convolutional neural network," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[12] D. B. Sam, N. N. Sajjan, H. Maurya, and R. V. Babu, "Almost unsupervised learning for dense crowd counting," in *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

[13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[14] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie, "Feature pyramid networks for object detection."

[15] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *European Conference on Computer Vision*. Springer, 2016, pp. 354–370.

[16] V. Sindagi and V. Patel, "Dafe-fd: Density aware feature enrichment for face detection," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 2185–2195.

[17] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," *arXiv preprint arXiv:1904.01649*, 2019.

[18] L. Wang, V. Sindagi, and V. Patel, "High-quality facial photo-sketch synthesis using multi-adversarial networks," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 2018, pp. 83–90.

[19] H. Zhang, V. Sindagi, and V. M. Patel, "Multi-scale single image dehazing using perceptual pyramid deep network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 902–911.

[20] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," *Engineering Applications of Artificial Intelligence*, vol. 43, pp. 81–88, 2015.

[21] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[22] D. Ryan, S. Denman, C. Fookes, and S. Sridharan, "Crowd counting using multiple local features," in *Digital Image Computing: Techniques and Applications, 2009. DICTA'09.* IEEE, 2009, pp. 81–88.

[23] K. Chen, C. C. Loy, S. Gong, and T. Xiang, "Feature mining for localised crowd counting." in *European Conference on Computer Vision*, 2012.

[24] H. Idrees, I. Saleemi, C. Seibert, and M. Shah, "Multi-source multi-scale counting in extremely dense crowd images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2547–2554.

[25] V. Lempitsky and A. Zisserman, "Learning to count objects in images," in *Advances in Neural Information Processing Systems*, 2010, pp. 1324–1332.

[26] V.-Q. Pham, T. Kozakaya, O. Yamaguchi, and R. Okada, "Count forest: Co-voting uncertain number of targets using random forest for crowd density estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3253–3261.

[27] B. Xu and G. Qiu, "Crowd density estimation based on rich features and random projection forest," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2016, pp. 1–8.

[28] C. C. Loy, K. Chen, S. Gong, and T. Xiang, "Crowd counting and profiling: Methodology and evaluation," in *Modeling, Simulation and Visual Analysis of Crowds*. Springer, 2013, pp. 347–382.

[29] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, and S. Yan, "Crowded scene analysis: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 3, pp. 367–386, 2015.

[30] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1299–1302.

[31] C. Arteta, V. Lempitsky, and A. Zisserman, "Counting in the wild," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–498.

[32] E. Walach and L. Wolf, "Learning to count with cnn boosting," in *European Conference on Computer Vision*. Springer, 2016, pp. 660–676.

[33] D. Onoro-Rubio and R. J. López-Sastre, "Towards perspective-free object counting with deep learning," in *European Conference on Computer Vision*. Springer, 2016, pp. 615–629.

[34] V. A. Sindagi and V. M. Patel, "Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting," in *Advanced Video and Signal Based Surveillance (AVSS), 2017 IEEE International Conference on*. IEEE, 2017.

[35] L. Boominathan, S. S. Kruthiventi, and R. V. Babu, "Crowdnet: A deep convolutional network for dense crowd counting," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 640–644.

[36] D. Oñoro-Rubio, R. J. López-Sastre, and M. Niepert, "Learning short-cut connections for object counting," in *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, 2018.

[37] K. X. Q. Q. X. Z. Ze Wang, Zehao Xiao and X. Cao, "In defense of single-column networks for crowd counting," in *British Machine Vision Conference 2018, BMVC 2018,*

Northumbria University, Newcastle, UK, September 3-6, 2018, 2018.

[38] V. A. Sindagi and V. M. Patel, "A survey of recent advances in cnn-based single image crowd counting and density estimation," *Pattern Recognition Letters*, 2017.

[39] J. Liu, C. Gao, D. Meng, and A. G. Hauptmann, "Decidenet: Counting varying density crowds through attention guided detection and density estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5197–5206.

[40] Z. Wang, Z. Xiao, K. Xie, Q. Qiu, X. Zhen, and X. Cao, "In defense of single-column networks for crowd counting," *arXiv preprint arXiv:1808.06133*, 2018.

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

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

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks." in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[44] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, "Adcrowdnet: An attention-injective deformable convolutional network for crowd understanding," *arXiv preprint arXiv:1811.11968*, 2018.

[45] J. Wan, W. Luo, B. Wu, A. B. Chan, and W. Liu, "Residual regression with semantic prior for crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4036–4045.

[46] M. Zhao, J. Zhang, C. Zhang, and W. Zhang, "Leveraging heterogeneous auxiliary tasks to assist crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 736–12 745.

[47] V. Sindagi and V. Patel, "Inverse attention guided deep crowd counting network," *arXiv preprint*, 2019.

[48] X. Jiang, Z. Xiao, B. Zhang, X. Zhen, X. Cao, D. Doermann, and L. Shao, "Crowd counting and density estimation by trellis encoder-decoder network," *arXiv preprint arXiv:1903.00853*, 2019.

[49] M. Shi, Z. Yang, C. Xu, and Q. Chen, "Revisiting perspective information for efficient crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7279–7288.

[50] W. Liu, M. Salzmann, and P. Fua, "Context-aware crowd counting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5099–5108.

[51] Q. Zhang and A. B. Chan, "Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8297–8306.

[52] Q. Wang, J. Gao, W. Lin, and Y. Yuan, "Learning from synthetic data for crowd counting in the wild," *arXiv preprint arXiv:1903.03303*, 2019.

[53] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[54] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 11, pp. 1254–1259, 1998.

[55] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain," *Nature reviews neuroscience*, vol. 3, no. 3, p. 201, 2002.

[56] Y.-Y. Tang, B. K. Hölzel, and M. I. Posner, "The neuroscience of mindfulness meditation," *Nature Reviews Neuroscience*, vol. 16, no. 4, p. 213, 2015.

[57] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu, "Bidirectional attentive fusion with context gating for dense video captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7190–7198.

[58] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 6298–6306.

[59] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 21–29.

[60] K. Chen, J. Wang, L.-C. Chen, H. Gao, W. Xu, and R. Nevatia, "Abc-cnn: An attention based convolutional neural network for visual question answering," *arXiv preprint arXiv:1511.05960*, 2015.

[61] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6077–6086.

[62] H. Xu and K. Saenko, "Ask, attend and answer: Exploring question-guided spatial attention for visual question answering," in *European Conference on Computer Vision*. Springer, 2016, pp. 451–466.

[63] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang, "Multi-context attention for human pose estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1831–1840.

[64] S. Woo, J. Park, J.-Y. Lee, and I. So Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[65] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Advances in Neural Information Processing Systems*, 2018, pp. 9401–9411.

[66] J. Park, S. Woo, J.-Y. Lee, and I. S. Kweon, "Bam: Bottleneck attention module," *arXiv preprint arXiv:1807.06514*, 2018.

[67] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 714–722.

[68] H. Zheng, J. Fu, T. Mei, and J. Luo, "Learning multi-attention convolutional neural network for fine-grained image recognition."

[69] L. Zhu, Z. Xu, and Y. Yang, "Bidirectional multirate reconstruction for temporal modeling in videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2653–2662.

[70] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4995–5004.

[71] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3156–3164.

[72] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[73] F. Saleh, M. S. A. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J. M. Alvarez, "Built-in foreground/background prior for weakly-supervised semantic segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 413–432.

[74] A. Chaudhry, P. K. Dokania, and P. H. Torr, "Discovering class-specific pixels for weakly-supervised semantic segmentation," in *British Machine Vision Conference (BMVC)*, 2017.

[75] P. O. Pinheiro and R. Collobert, "From image-level to pixel-level labeling with convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1713–1721.

[76] A. Roy and S. Todorovic, "Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3529–3538.

[77] S. Hong, D. Yeo, S. Kwak, H. Lee, and B. Han, "Weakly supervised semantic segmentation using web-crawled videos," *arXiv preprint arXiv:1701.00352*, 2017.

[78] G. Papandreou, L.-C. Chen, K. P. Murphy, and A. L. Yuille, "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1742–1750.

[79] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 189–203, 2017.

[80] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free?-weakly-supervised learning with convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[81] V. Kantorov, M. Oquab, M. Cho, and I. Laptev, "Contextlocnet: Context-aware deep network models for weakly supervised localization," in *European Conference on Computer Vision*. Springer, 2016, pp. 350–365.

[82] M. Shi, H. Caesar, and V. Ferrari, "Weakly supervised object localization using things and stuff transfer," *arXiv preprint arXiv:1703.08000*, 2017.

[83] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4048–4056.

[84] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn*, 2017, pp. 136–145.

[85] M. Oquab, L. Bottou, I. Laptev, J. Sivic *et al.*, "Weakly supervised object recognition with convolutional neural networks," in *Proc. of NIPS*, 2014.

[86] Z. Wang, L. Wang, Y. Wang, B. Zhang, and Y. Qiao, "Weakly supervised patchnets: Describing and aggregating local patches for scene recognition," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 2028–2041, 2017.

[87] B. Liu and N. Vasconcelos, "Bayesian model adaptation for crowd counts," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4175–4183.

[88] A. B. Chan and N. Vasconcelos, "Counting people with low-level features and bayesian regression," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 2160–2177, 2012.

[89] M. von Borstel, M. Kandemir, P. Schmidt, M. K. Rao, K. Rajamani, and F. A. Hamprecht, "Gaussian process density counting from weak supervision," in *European Conference on Computer Vision*. Springer, 2016, pp. 365–380.

[90] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Bian, and Y. Yang, "Progressive learning for person re-identification with one example," *IEEE Transactions on Image Processing*, 2019.

[91] X. Liu, W. Liu, T. Mei, and H. Ma, "Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 645–658, 2017.

[92] D. P. Kingma, S. Mohamed, D. J. Rezende, and M. Welling, "Semi-supervised learning with deep generative models," in *Advances in neural information processing systems*, 2014, pp. 3581–3589.

[93] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko, "Semi-supervised learning with ladder networks," in *Advances in neural information processing systems*, 2015, pp. 3546–3554.

[94] X. Dong, L. Zheng, F. Ma, Y. Yang, and D. Meng, "Few-example object detection with model communication," *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[95] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[96] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*, 1998, pp. 570–576.

[97] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2921–2929.

[98] B. Zhou, X. Tang, and X. Wang, "Learning collective crowd behaviors with dynamic pedestrian-agents," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 50–68, 2015.

[99] H. Idrees, K. Soomro, and M. Shah, "Detecting humans in dense crowds using locally-consistent scale prior and global occlusion reasoning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 10, pp. 1986–1998, 2015.