Attentional Neural Fields for Crowd Counting

Anran Zhang[1]*, Lei Yue[1]*, Jiayi Shen[1], Fan Zhu[4], Xiantong Zhen[4], Xianbin Cao[1,2,3]†, Ling Shao[4]
[1]School of Electronic and Information Engineering, Beihang University, Beijing, China
[2]Key Laboratory of Advanced Technology of Near Space Information System (Beihang University), Ministry of Industry and Information Technology of China, Beijing, China
[3]Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, Beijing, China
[4]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
zhanganran@buaa.edu.cn, yuelei@buaa.edu.cn, shenjiayi@buaa.edu.cn,
fan.zhu@inceptioniai.org, zhenxt@gmail.com, xbcao@buaa.edu.cn, ling.shao@ieee.org

Abstract

Crowd counting has recently generated huge popularity in computer vision, and is extremely challenging due to the huge scale variations of objects. In this paper, we propose the Attentional Neural Field (ANF) for crowd counting via density estimation. Within the encoder-decoder network, we introduce conditional random fields (CRFs) to aggregate multi-scale features, which can build more informative representations. To better model pair-wise potentials in CRFs, we incorperate non-local attention mechanism implemented as inter- and intra-layer attentions to expand the receptive field to the entire image respectively within the same layer and across different layers, which captures long-range dependencies to conquer huge scale variations. The CRFs coupled with the attention mechanism are seamlessly integrated into the encoder-decoder network, establishing an ANF that can be optimized end-to-end by back propagation. We conduct extensive experiments on four public datasets, including ShanghaiTech, WorldEXPO 10, UCF-CC-50 and UCF-QNRF. The results show that our ANF achieves high counting performance, surpassing most previous methods.

1. Introduction

Crowd counting, which aims to predict an accurate number of individuals in a scene, has recently generated great popularity in computer vision due to its extensive real-world applications, such video surveillance and urban planning. However, crowd counting for real-life applications faces many challenges. Some of the most common include occlusion, low image quality/resolution, severe perspective distortion, huge scale variation and the model inefficiency at the inference stage [24, 25]. Prior works [28, 41] have made great attempts to address these issues, and deep learning models have been ubiquitously employed in existing crowd counting methods. For instance, [21] iteratively incorporates the convolutional features and the predicted multi-resolution density maps in various stages, and [28] combines global and local contextual information from multiple estimators.

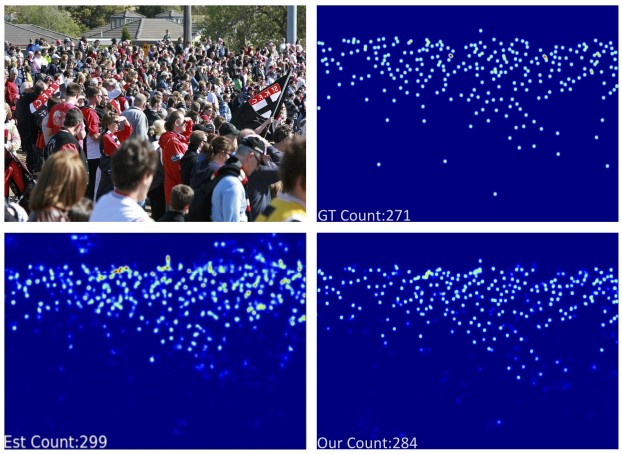


Figure 1. Density estimation results. Top Left: Input image. Top right: Ground truth. Bottom Left: The best previous method (SANet [3]). Bottom Right: ANF. We can easily observe the huge scale variations in the input image, where the spatial scales of pedestrians with large vertical pair-wise distances are diverse.

In most crowd counting tasks, the input data comes from surveillance cameras that are mounted above the crowd as illustrated in Fig. 1. This means that the scale variation normally takes place in regions with large vertical distances on input surveillance images. Thus, how to handle scale variations while utilizing multi-scale features becomes the

*These authors contribute equally.
†Corresponding author.

crux of crowd counting, which is the major focus in this work. Previous research attempts [3, 41] have been made to address the scale variation issue. For instance, the scale aggregation network (SANet) was developed in [3] to address the scale variance problem; however, it largely relies on different scales of convolutional kernels. Despite multi-scale convolutional kernels being used, the obtained feature maps still suffer from significant information loss as the network goes deeper. Feature aggregation across different layers using skip connections is popular in image segmentation tasks, such as UNet [23] and DeepLab [6]. This has shown the great effectiveness of fusing multiple features across convolutional layers in CNNs [20, 42, 38], which, however, remains unexplored for the crowd counting task.

To deal with scale variation, long-range dependency has recently been explored in the form of a non-local operation embedded in convolutional neural networks (CNNs) [32]. It computes the response at a specific position by attending to all positions on the feature graph and taking their weighted average in an embedding space. The effectiveness of the non-local operation stems from the fact that it essentially expands the receptive fields in CNNs. Meanwhile, conditional random fields (CRF) [15, 16] as a representative discriminative graphical model have been investigated in conjunction with CNNs for multiple inference tasks, e.g., image segmentation [9] and depth estimation [34].

In this paper, we propose attentional neural fields (ANF) for crowd counting by density estimation. Within the convolutional encoder-decoder network, the ANF integrates conditional random fields and an attention mechanism, which can jointly aggregate multi-scale features and capture long-range dependencies. Different from the previous work [34] in which CRFs are usually appended to the prediction results as a post-processing, our ANF directly applies the random fields to the feature level to aggregate and refine of multi-scale features derived from the CNN inner layers. Furthermore, we introduce an attention mechanism to construct pair-wise potentials in the CRF. The attention is implemented as the inter- and the intra-layers, and can exploit the spatial correlations between feature vectors not only within the same scale but also across different scales. Both intra and inter attentions calculate the weighted sum across the holistic feature map as the response at each pixel, which is beneficial to enlarging receptive fields and building the communication channel across different scales of feature maps. More importantly, attention variables and multi-scale feature variables are jointly estimated through mean-fields updates and the full architecture can be trained in an end-to-end manner.

The proposed ANF seamlessly assembles conditional random fields, the attention mechanism and neural networks, establishing a new compact deep learning model. Its effectiveness is demonstrated by extensive experiments on four public benchmarks, i.e., ShanghaiTech, WorldEXPO'10, UCF-CC-50, and UCF-QNRF. Moreover, the results have shown that the proposed ANF can handle both sparse and dense crowds, which indicates its huge generality for diverse crowd counting tasks.

To summarize, the major contributions of this work are as follows.

- We propose the attentional neural field (ANF) for crowd counting, which leverages the strengths of both convolutional neural networks for feature learning and conditional random fields (CRFs) for fusing multi-scale features.

- We introduce the attention mechanism to model the pair-wise potentials in CRFs. We implement it with inter- and intra-layer attentions to capture long-range dependencies both within the same scale and across different scales, which can, to a large extent, handle huge scale variations.

- The proposed ANF achieves the new state-of-the-art performance on four public benchmark datasets. In particular, on the challenging UCF-QNRF dataset with dense crowds, our method surpasses the best previous method by up to 16.6% in terms of MAE.

2. Related work

Our attentional neural field (ANF) implants the conditional random field (CRF) and the attention mechanism into a convolutional network, establishng a new and compact deep model for crowd counting. We briefly review recent work on crowd counting as well as the relevant work using CRFs and attention.

Crowd Counting. In general, there are three types of crowd counting approaches, including person detection, holistic person quantity regression and density map estimation. While object detection (*person* is one of the commonly studied categories in object detection) has achieved tremendous success, predicting the quantity of people by aggregating all located persons in the crowd is the most straightforward approach for crowd counting. In this aspect, a significant number of previous work have been proposed, from the distant handcrafted low-level features [30, 8, 17] to recent CNN-based approaches. Unfortunately, due to occlusion, undersize and low-quality issues, even the best-performing person detection methods are incapable of achieving satisfactory performance for the crowd counting task. Regression-based methods [4, 5, 14] learn from a holistic crowd image and directly regress on the person quantity from the input crowd image without explicitly localizing each person's individual position.

In recent years, density map estimation-based methods have started to play an increasingly important role in crowd

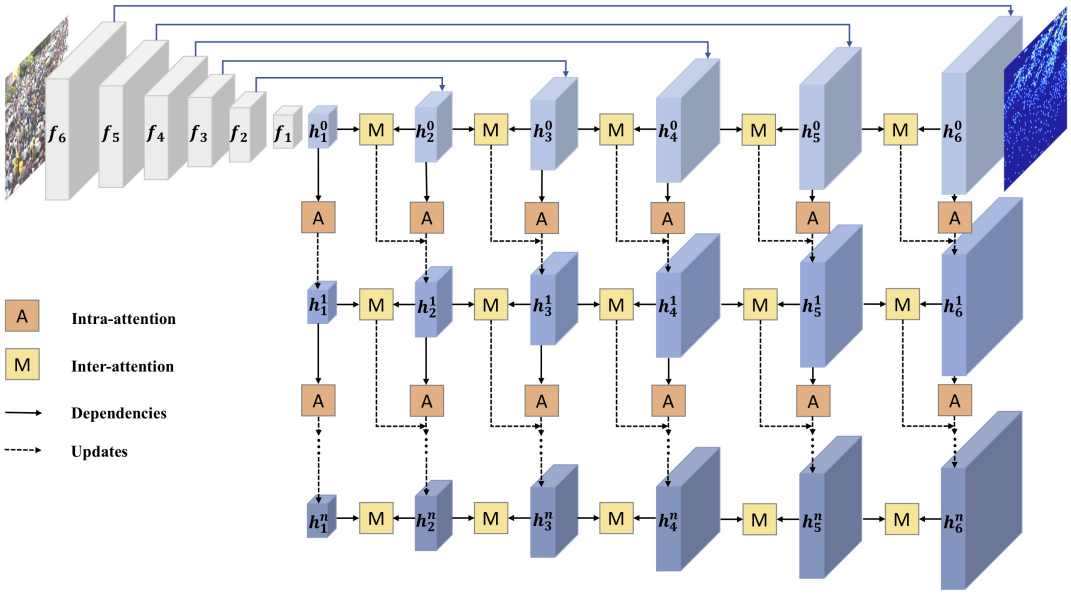


Figure 2. Architecture overview of the proposed attentionl neural network (ANF) for crowd counting. The network incorporates conditional random fields in conjunction with attention mechanism into the encoder-decoder network. Blocks A and M represent the intra and inter attentions. The input arrows indicate the dependencies among the estimated variables in the message passing. The output arrows denote the updates involving the attentional neural fields.

counting [28, 19, 21, 13, 33]. Compared with person detection and holistic person quantity regression, density map estimation is capable of capturing richer spatial information in the crowd image. A significant number of density map estimation approaches target to at capturing the multi-scale spatial information in order to improve the performance. For example, Sindag et al. [28] proposed a contextual pyramid CNN that utilizes various estimators to capture both global and local contextual information, which is integrated with high-dimensional feature maps extracted from a multi-column CNN by a Fusion-CNN. Li et al. [19] replaced the pooling operation with dilated kernels to fuse multi-scale contextual information. Ranjan et al. [21] proposed a multi-stage method for generating high-resolution density map by combining low-resolution density map of the previous stage along with extracted features. Particularly, scale variation is one of the most critical issues for density map estimation. To address this issue, Boominathan et al. [1] employed a multi-column architecture to capture the scale variations with multiple receptive fields in each column, where multi-column features were fused by the convolutional layer for crowd density regression. The proposed ANF is also in the family of crowd counting by density estimation. However, our ANF differs from the above work in two major aspects: 1) it jointly learns the CNN with CRFs, which offers a powerful tool to fuse multi-scale features in the CNN; 2) it effectively incorporates the attention mechanism to model the pair-wise potentials in the fields, which is implemented as both an inter-layer model and intra-layer attentions to fully capture the long-range dependencies.

Conditional Random Fields. As a powerful graphical model, conditional random fields have been successfully applied in various pixel-level labeling tasks, including semantic segmentation [43], depth estimation [34], pose estimation [7], etc. Recent work [43, 34] has shown that CRFs are jointly learnable with CNN frameworks. The most similar work to ours is [34], where a CRF is adopted to combine multi-scale information derived from multiple intermediate layers of a CNN. In contrast to those previous work, our ANF does not produce any side output, and integrates the attention module of both inter and intra layers to model the potentials in CRFs, which can improve the performance of the density map estimation.

Attention. Attention mechanism [29] has recently been incorporated into deep learning for performance augmentation, which has achieved great success in a large variety of vision tasks, e.g., image captioning [35, 37], image question answering [36], image classification [31], face alignment [39, 18] and video analysis [10]. Different from the spatial attention and its variations, our attention is more related to the non-local attention mechanism. Both our attention model and the spatial attention model perform a soft selection on variables and adaptive feature scales. However, the spatial attention model usually computes the response of each position across the channel without exploring the spatial dependencies, while our attention model considers the spatial dependencies not only in the same feature map but also across different feature maps. To the best of our knowledge, this is the first attempt that jointly learns a non-local CNN with a CRF for the crowd counting.

3. Crowd Counting by Density Estimation

Our attentional neural field (ANF) integrates conditional random fields (CRFs) and an attention mechanism into a convolutional encoder-decoder framework, combining their respective strengths. We start with the problem statement of crowd counting by density estimation and provide necessary preliminaries on CRFs and attention mechanisms.

3.1. Problem Statement and Preliminaries

Consider a training set $\mathbf{T} = \{(\mathbf{X}_i, \mathbf{D}_i)\}_{i=1}^{N}$, where $\mathbf{X}_i$ denotes the input RGB image, $\mathbf{D}_i$ denotes its corresponding real-valued crowd density map, and N is number of training samples. The task of crowd counting is essentially to find a non-linear mapping from the input image $\mathbf{X}$ to the density map $\mathbf{D}$ based on which we can compute the crowd counts. Our attentional neural field is built upon the encode-decoder architecture. The encoder consists of six residual convolutional blocks. Each convolutional block downsamples the feature map by a factor of 2 and outputs the feature map with the same number of channels. The decoder has six convolutional layers without pooling operations.

Conditional Random Fields. We denote the feature maps extracted by the encoder of the CNN as $\mathbf{F} = \{\mathbf{F}_s\}_{s=1}^{S}$, where $\mathbf{F}_s$ denotes the feature at scale s. $\mathbf{F}_s$ is composed of a set of feature vectors, $\mathbf{F}_s = \{\mathbf{f}_s^i\}_{i=1}^{P}, \mathbf{f}_s^i \in \mathbb{R}^{C_s}$, P is the number of pixels, and C_s is the number of channels at scale s. The feature maps in the decoder are defined as $\mathbf{H} = \{\mathbf{H}_s\}_{s=1}^{S}$, where similarly $\mathbf{H}_s = \{\mathbf{h}_s^i\}_{i=1}^{P}$ and $\mathbf{h}_s^i \in \mathbb{R}^{C_s}$. To generate high quality density maps for crowd counting, we introduce conditional random fields (CRFs) to fuse multi-scale features, which improves the robustness of feature representation against huge scale variations. Specifically, CRFs learn a set of latent features as hidden variables that form random fields conditioned upon the observation of feature maps $\mathbf{F}$ from the encoder. Those hidden variables are feature maps $\mathbf{H} = \{\mathbf{H}_s\}_{s=1}^{S}$ in the decoder.

Non-local Attention Mechanism. We introduce the attention mechanism to model the pair-wise potential between hidden variables. To be more specific, we implement inter-layer and intra-layer attention models to formulate the pair-wise potential between feature vectors in the same scale and across different scales respectively. The intra-layer attention map at s scale is defined as $\mathbf{A}_s = \{a_s^{i,j}\}_{i=1,j=1}^{N_s \times N_s}$, which describes the similarity between the latent feature vector at the pixel i and pixel j. The inter-layer attention map $\mathbf{M} = \{\mathbf{M}_{s-1,s}\}_{s=2}^{S}$ encodes the relationship between the latent feature vector at neighboring scales, where $\mathbf{M}_{s-1,s} \in \mathbb{R}^{P_{s-1} \times P_s}$, and $m_{s-1,s}^{i,j} = \mathbf{M}_{s-1,s}(i,j)$ in the range of $[0,1]$ is the correlation score between the pixel j at the scale s to the pixel i at the scale $s-1$. Through the inter-layer attention the updating of hidden variables $\mathbf{h_s^i}$ at scale s takes into account the information from the scale $s-1$. To summarize, our implementation of the non-local attention mechanism computes the response at a position by attending interactions between any two positions, and models pair-wise potential within all pixels. It is worth noting that our non-local attention mechanism considers pair-wise potential for both inter-layer interactions and intra-layer interactions. The utilization of attention models in pair-wise potential helps to expand the receptive fields to the whole image which improves the robustness of the framework against scale variation. Moreover, the long-range dependencies among the spatial locations are fully captured by encoding their correlations.

3.2. Attentional Neural Fields

Given the observed multi-scale feature maps $\mathbf{F}$ of image $\mathbf{X}$, the objective is to estimate the latent multi-scale representation $\mathbf{H} = \{\mathbf{H}_s\}_{s=1}^{S}$, the inter-layer attention variables $\mathbf{M} = \{\mathbf{M}_{s-1,s}\}_{s=2}^{S}$ and the intra-layer attention variables $\mathbf{A} = \{\mathbf{A}_s\}_{s=1}^{S}$. We formalize the problem within a conditional random field framework and write the Gibbs distribution as:

$$P(\mathbf{H}, \mathbf{M}, \mathbf{A} | \mathbf{X}, \Theta) = \exp(-E(\mathbf{H}, \mathbf{M}, \mathbf{A}, \mathbf{X}, \Theta))/Z(\mathbf{X}, \Theta), \tag{1}$$

where $\mathbf{\Theta}$ is the set of parameters and E is the energy function. The energy function is defined as:

$$E(\mathbf{H}, \mathbf{M}, \mathbf{A}) = \Phi(\mathbf{H}, \mathbf{F}) + \Psi(\mathbf{H}, \mathbf{M}) + \Xi(\mathbf{H}, \mathbf{A}). \tag{2}$$

The first term in (2) is the classical unary potential relating the latent feature representation $\mathbf{h}_s^i$ to the observed multi-scale feature vector $\mathbf{f}_s^i$, that is:

$$\Phi(\mathbf{H}, \mathbf{F}) = \sum_{s=1}^{S} \sum_{i} \phi(\mathbf{h}_s^i, \mathbf{f}_s^i) = -\sum_{s=1}^{S} \sum_{i} \frac{1}{2} \parallel \mathbf{h}_s^i - \mathbf{f}_s^i \parallel^2 . \tag{3}$$

The second term in (2) models the relationship between the latent feature vectors at neighboring scales upon inter-layer attention variable $m_{s-1,s}^{ij}$, which is defined as:

$$\begin{aligned} \Psi(\mathbf{H}, \mathbf{M}) &= \sum_{s=2}^{S} \sum_{i,j} \psi(m_{s-1,s}^{ij}, \mathbf{h}_s^i, \mathbf{h}_{s-1}^j) \\ &= m_{s-1,s}^{i,j} \psi_h(\mathbf{h}_s^i, \mathbf{h}_{s-1}^j) \end{aligned} \tag{4}$$

As in previous work [2, 32], we consider a dot-product similarity to enforce the estimated latent features to be close to their corresponding observations. Following the non-local mean operation [2] and the generic non-local operation in deep neural networks [32], we use a normalized dot-product similarity to define $\psi_h(\mathbf{h}_s^i, \mathbf{h}_{s-1}^j)$ as:

$$\psi_h(\mathbf{h}_s^i, \mathbf{h}_{s-1}^j) = \mathbf{h}_{s-1}^j (\mathbf{h}_s^i)^\top. \tag{5}$$

The third term in (2) enocdes the pair-wise relationships between hidden feature vectors at the same scale upon the intra-layer attention variable a_s^{ij}, which is defined as:

$$\Xi(\mathbf{H}, \mathbf{A}) = \sum_{s=1}^{S} \sum_{i,j} \xi_h(a_s^{ij}, \mathbf{h}_s^i, \mathbf{h}_s^j) \tag{6}$$

Specifically, we define:

$$\xi_h(a_s^{ij}, \mathbf{h}_s^i, \mathbf{h}_s^j) = a_s^{ij} \xi_h(\mathbf{h}_s^i, \mathbf{h}_s^j) = a_s^{ij} \mathbf{h}_s^j (\mathbf{h}_s^i)^\top. \tag{7}$$

3.3. Inference

Following the previous works [43, 22], we employ the mean-field approximation in order to derive a tractable inference procedure. Under the mean field theory, the best approximation of these variables is the distribution Q that minimizes the Kullback-Leibler (KL) divergence between these variables and Q. The solution for Q is formed in [22]. By considering the potentials defined in (2), (4) and (6) and denoting $\mathbb{E}_q$ as the expectation with distribution q, we have:

$$\begin{aligned} q(\mathbf{h}_s^i) \propto \exp(&\phi(\mathbf{h}_s^i, \mathbf{f}_s^i) \\ &+ \sum_j \mathbb{E}_{q(a_s^{ij})}\{a_s^{ij}\} \mathbb{E}_{q(\mathbf{h}_s^j)}\{\xi_h(a_s^{ij}, \mathbf{h}_s^i, \mathbf{h}_s^j)\} \\ &+ \sum_j \mathbb{E}_{q(m_{s-1,s}^{ij})}\{m_{s-1,s}^{ij}\} \mathbb{E}_{q(\mathbf{h}_{s-1}^j)}\{\psi_h(\mathbf{h}_s^i, \mathbf{h}_{s-1}^j)\}) \end{aligned} \tag{8}$$

$$q(m_{s-1,s}^{ij}) \propto \exp(m_{s-1,s}^{ij} \mathbb{E}_{q(\mathbf{h}_s^i)}\{\mathbb{E}_{q(\mathbf{h}_{s-1}^j)}\{\psi_h(\mathbf{h}_s^i, \mathbf{h}_{s-1}^j)\}\}) \tag{9}$$

$$q(a_s^{ij}) \propto \exp(a_s^{ij} \mathbb{E}_{q(\mathbf{h}_s^i)}\{\mathbb{E}_{q(\mathbf{h}_s^j)}\{\xi_h(\mathbf{h}_s^i, \mathbf{h}_s^j)\}\}). \tag{10}$$

By considering the potentials defined in (3), (4) and (6) and denoting

$$\bar{a}_s^{ij} = \mathbb{E}_{q(a_s^{ij})}\{a_s^{ij}\} \tag{11}$$

$$\bar{m}_{s-1,s}^{ij} = \mathbb{E}_{q(m_{s-1,s}^{ij})}\{m_{s,s-1}^{ij}\} \tag{12}$$

$$\bar{\mathbf{h}}_s^i = \mathbb{E}_{q(\mathbf{h}_s^i)}\{\mathbf{h}_s^i\}, \tag{13}$$

we can derive mean-field update for the latent feature representation:

$$\bar{\mathbf{h}}_s^i = \mathbf{f}_s^i + \sum_j \bar{m}_{s,s-1}^{ij} \bar{\mathbf{h}}_{s-1}^j + \sum_j \bar{a}_s^{ij} \bar{\mathbf{h}}_s^j. \tag{14}$$

Since $m_{s-1,s}^{ij}$ and $\bar{\mathbf{h}}_{s-1}^j (\bar{\mathbf{h}}_s^i)^\top$ are in the range of [0,1], their expectation can be derived considering (9), (10) as the approximate form:

$$\bar{m}_{s,s-1}^{ij} = \frac{\exp(\bar{\mathbf{h}}_{s-1}^j (\bar{\mathbf{h}}_s^i)^\top) - 1}{\bar{\mathbf{h}}_{s-1}^j (\bar{\mathbf{h}}_s^i)^\top}. \tag{15}$$

We use the $\mathrm{softmax}$ computation to normalize $\bar{m}_{s-1,s}^{ij}$ following previous work [32]. This can be seen from the fact that, for a given i, $\bar{m}_{s-1,s}^{ij}$ indicates the correlation between the feature in position i and position j across the scale s and $s-1$. And $\bar{a}_s^{ij}$ can be computed as:

$$\bar{a}_s^{ij} = \frac{\exp(\bar{\mathbf{h}}_s^j (\bar{\mathbf{h}}_s^i)^\top) - 1}{\bar{\mathbf{h}}_s^j (\bar{\mathbf{h}}_s^i)^\top}. \tag{16}$$

We can simultaneously learn the parameters of the CRFs and those of the encoder-decoder network. To infer the latent multi-scale representations $\mathbf{H}$, the inter-layer attention variables $\mathbf{M}$ and the intra-layer attention variables $\mathbf{A}$, we implement the mean-field updates with a neural network by multiple iterations.

3.4. Message Passing

We perform the mean-field updates jointly for both the attention variables and the latent feature maps, according to the derivation described in Sec. 3.3. By message passing from all positions in same layer or different layers to attention variables, the long-range dependencies can be captured and passed to the final representations.

To perform mean-field updates for the inter-layer attention $\mathbf{M}$, we use (14) to update each of inter-layer attention variable $m_{s,s-1}^{ij}$ over several steps, as follows: (i) We perform the message passing from the associated feature maps $\bar{\mathbf{h}}_s$, where $\bar{\mathbf{h}}_s$ is initialized with corresponding feature observations $\mathbf{f}_s$. (ii) Message passing from the associated feature $\bar{\mathbf{h}}_s$ and neighboring feature $\bar{\mathbf{h}}_{s-1}$ to inter-attention feature $\bar{\mathbf{m}}_{s,s-1}$ is performed via (15) as $\bar{\mathbf{m}}_{s,s-1} \leftarrow \frac{\exp(\bar{\mathbf{h}}_{s-1} \bar{\mathbf{h}}_s^\top) - 1}{\bar{\mathbf{h}}_{s-1} \bar{\mathbf{h}}_s^\top}$, where $\bar{\mathbf{h}}_{s-1} \bar{\mathbf{h}}_s^\top$ is calculated via a matrix multiplication operation and the normlization. Similarly, we perform the mean-field update of intra-attention variables $\bar{\mathbf{a}}_s$ using (16).

Once the inter-attention and intra-attention maps are updated, we use them as guidance to update the latent feature maps $\mathbf{h}_s$. The mean-field updates of $\bar{\mathbf{h}}_s$ can be carried out using (14) as follows: (i) Message passing from the inter-attention feature $\bar{\mathbf{m}}_{s,s-1}$ and the feature at $s-1$ scale to the feature s scale is performed by the matrix multiplication operation in $\hat{\mathbf{h}}_s \leftarrow \bar{\mathbf{m}}_{s,s-1} \bar{\mathbf{h}}_{s-1}$. (ii) Message passing within the same scale is applied with intra-attention feature $\tilde{\mathbf{h}}_s \leftarrow \bar{\mathbf{a}}_s \bar{\mathbf{h}}_s$ by matrix multiplication operation. (iii) The message is passed to the final $\mathbf{h}_s$ by adding the unary term $\bar{\mathbf{h}}_s \leftarrow \hat{\mathbf{h}}_s \oplus \tilde{\mathbf{h}}_s \oplus \mathbf{f}_s$, where $\oplus$ denotes the element-wise sum operation.

3.5. Optimization

Our attentional neural fields integrate CRFs and non-local attentions in the convolutional encoder-decoder network, which can be learned by jointly optimizing the parameter Θ_c of the network and the parameters Θ_f of the attentional fields.

Given the training set $\mathbf{T} = \{(\mathbf{X}_i, \mathbf{D}_i)\}_{i=1}^N$, we minimize the difference between predicted density map and that of the

ground truth by ℓ_2-norm in the objective function, which takes the following form:

$$\mathcal{L}_F(\mathbf{X}, \mathbf{D}; \Theta_c, \Theta_f) = \sum_{i=1}^{N} ||F(\mathbf{X}_i; \Theta_c, \Theta_f) - \mathbf{D}_i||_2^2. \quad (17)$$

The optimization is conducted in a end-to-end manner by the back propagation algorithm using mini-batch based stochastic gradient descent. The inference of latent variables is performed along with each iteration of parameter updates of the network.

4. Experiments

We conduct extensive experiments on four benchmark datasets, including ShanghaiTech [41], WorldExpo 10 [40], UCF-CC-50 [11] and UCF-QNRF [12]. The results demonstrate that the proposed attentional neural field (ANF) consistently delivers high performance on all datasets, and exceeds most previous methods. The ablation studies further verify the great effectiveness of the ANF.

4.1. Implementation Details

We provide our implementation details, including data augmentation, ground truth generation, evaluation metrics and architecture design, to facilitate comparison with other methods.

Data augmentation. In this work, we use a patch-based training and image-based testing scheme. To fully make use of the dataset with limited number of training samples, we train our network by random scaling and cropping of images. Firstly, we select a random value to change the original image to different scales, which increases the network's robustness against size variations in human objects. Then, we randomly crop patches from the images at different locations. During testing, we feed the whole images into the network instead of the cropped patches.

Ground-Truth Generation. Since annotations for crowd images are labeled at the center of the pedestrian head, we use the Gaussian kernel to convert these points to generate crowd density maps. The normalized Gaussian kernel is defined as :

$$D(x) = \sum_{x_i \in S} \delta(x - x_i) * G_\sigma, \quad (18)$$

where D denotes the crowd density map and S is the set of all annotated points. A point at pixel x_i can be represented with a delta function $\delta(x - x_i)$. The density map can be obtained by convolving $\delta(x - x_i)$ with a Gaussian kernel and parameter G_σ. We fix the Gaussian kernel size to be 15×15.

Evaluation Metrics. Counting error is commonly measured by two metrics, i.e., Mean Absolute Error (MAE) and Mean Squared Error (MSE):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |y_i - y_i^{'}| \quad (19)$$

$$\text{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} |y_i - y_i^{'}|^2}, \quad (20)$$

where N is number of test samples, y_i is the ground truth count and $y_i^{'}$ is the estimated count corresponding to the i^{th} sample. MAE indicates the accuracy of the predicted result and MSE measures the robustness.

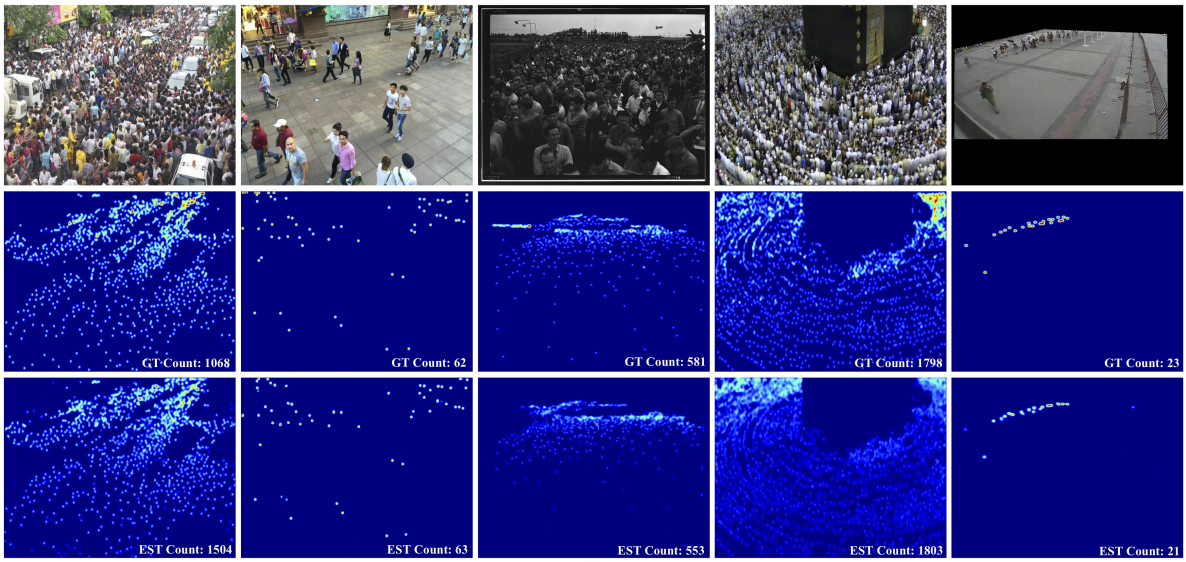
Network Architecture. We employ a simple architecture, which is made of one convolutional block for reducing inputs resolution and 6 residual convolutional blocks. Each residual convolutional block downsamples the feature map by a factor of 2 and outputs the feature map with the same number of channels. The decoder has 6 residual convolutional layers without a pooling operation. Our ANF uses both inter-attention and intra-attention with CRF, which experimentally produces the best overall performance.

4.2. Performance and Comparison

We show the performance on the four benchmark datasets and compare the proposed ANF with previous methods. Overall, our ANF produces the new state-of-the-art performance on all datasets and outperforms most of the compared methods.

ShanghaiTech. The ShanghaiTech dataset [41] contains still images with arbitrary camera perspectives and crowd densities. The ShanghaiTech dataset is composed of 1198 annotated images, including both internet and street view images. Images in Part A are randomly crawled from the Internet, and most of them have a large number of people. Images in Part B are taken from busy streets of metropolitan areas in Shanghai. There are tremendous occlusions for most people in each image, and the scale of the people is varies. Compared with other datasets, most images from ShanghaiTech have low resolution, and thus we maintain their resolution for training and testing.

We compare our ANF with recent methods evaluated on this dataset and the results are reported in Table 1. We achieve the best performance in terms of MAE and MSE on Part A, and gain competitive results on Part B. We show the samples from the ShanghaiTech dataset in Fig. 3. This demonstrates that the population density is unevenly distributed and heads in the crowds are highly diverse in scale. Our model can accurately determine each person's location.



(a) (b) (c) (d) (e)

Figure 3. Predictions in all datasets: (a) ShanghaiTech A, (b) ShanghaiTech B, (c) UCF-CC-50, (d) UCF-QNRF and (e) WorldExpo 10.

Table 1. Estimation errors on the ShanghaiTech, UCF-CC-50 and UCF-QNRF datasets.

	ShanghaiTech Part A		ShanghaiTech Part B		UCF-CC-50		UCF-QNRF	
Method	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE
Zhang et al. [40]	181.8	277.7	32.0	49.8	467.0	498.5	-	-
MCNN [41]	110.2	173.2	26.4	41.3	3716.6	509.1	277	426
Cascaded-MTL [27]	101.3	152.4	20.0	31.1	322.8	397.9	252	514
Switching-CNN [26]	90.4	135.0	21.6	33.4	318.1	439.2	228	445
CP-CNN [28]	73.6	106.4	20.1	30.1	295.8	**320.9**	-	-
CSRNet [19]	68.2	115.0	10.6	16.0	266.1	397.5	-	-
SANet [3]	67.0	104.5	8.4	13.6	258.4	334.9	-	-
Idrees et al. [12]	-	-	-	-	-	-	132	191
ANF (Ours)	**63.9**	**99.4**	**8.3**	**13.2**	**250.2**	340.0	**110**	**174**

Table 2. Performance comparison on WorldExpo 10.

Method	S1	S2	S3	S4	S5	Avg
Zhang et al. [40]	9.8	14.1	14.3	22.2	3.7	12.9
MCNN [41]	3.4	20.6	12.9	13.0	8.1	11.6
CP-CNN [28]	2.9	14.7	10.5	10.4	5.8	8.9
CSRNet [19]	2.9	11.5	**8.6**	16.6	3.4	8.6
SANet [3]	2.6	13.2	9.0	13.3	**3.0**	8.2
ANF (Ours)	**2.1**	**10.6**	15.1	**9.6**	3.1	**8.1**

UCF-CC-50. The UCF-CC-50 dataset is introduced in [11]. It is a very small dataset with only 50 annotated crowd images. There is a large variation in crowd counts with the number of people in an image ranging from 96 to 4633. The limited number of images makes it a challenging dataset for deep learning methods. We follow the same settings as [41] and use five fold cross-validation for performance evaluation.

The comparison of the proposed ANF with other existing methods is summarized in Table 1. The proposed ANF produces the best performance in terms of MAE, and highly competitive performance in terms of MSE.

UCF-QNRF. The UCF-QNRF dataset [12] is a large-scale dataset that contains a wide variety of observation viewpoints, densities and lighting conditions. The number of people in an image ranges from 49 to 12865, which makes the crowd dense and challenging to count. We follow the settings in [12], and split the training and test set into 1201 and 334 images, respectively.

The comparison with previous methods is summarized in Table 1 on UCF-QNRF. Our ANF delivers the best result of all in terms of MAE, surpassing the second best approach by an improvement of 16.6% in MAE. It is also much better than previous methods in terms of MSE. The results on this dataset indicate the great ability of our ANF to handle extremely dense crowds.

WorldExpo 10. The WorldExpo 10 dataset [40] is composed of 1132 annotated video sequences captured by 108 surveillance cameras from Shanghai 2010 WorldExpo. The dataset can be classified into 5 different scenes, each containing 120 frames. This dataset provides perspective maps,

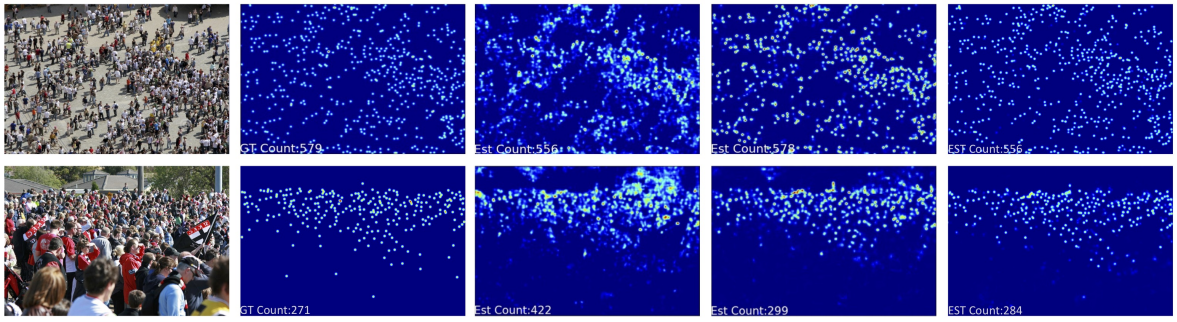


Figure 4. From left to right: RGB images, ground truth, MCNN [41] prediction, SANet [3] prediction, ANF prediction.

Table 3. Effectiveness of the proposed inter-/intra-layer attentions.

Method	MAE	MSE
baseline	66.2	110.8
intra & inter attentions	**63.9**	**99.4**

the value of which represents the number of pixels in the image covering one square meter of a real location.

Compared to UCF-QNRF, the crowds in this dataset are relatively sparse. Our ANF still yields the best performance on Scenes 1, 2 and 4, with highly competitive results on Scenes 3 and 5. The results demonstrate the great generality and effectiveness of the proposed ANF to handle both sparse and dense crowds.

4.3. Ablation study

To gain insight into the proposed ANF, we perform an ablation study to demonstrate the contribution of each of its components. We follow the previous work [19, 3, 21], using ShanghaiTech Part A as the benchmark for the ablation study. We compare the performance of our design choices with our baseline, and compare the density maps produced by our ANF with those produced by several state-of-the-art approaches. The ablation study verifies the great effectiveness of the proposed ANF for crowd counting.

Counting Accuracy. The ablation study results are shown in Table 3. The table is partitioned row-wisely into two groups, with three configurations. Each group contains the indexed configurations corresponding to one main contribution of ANF. These include the ANF with both inter-layer attention and intra-layer attention. In different columns, we report the counting accuracy of each configuration, using the MAE and MSE metrics. Qualitatively, we visualize the density maps generated by representive method (MCNN [41]), the current best-performing method (SANet [3]), and our ANF on the ShanghaiTech Part A dataset in Figure 4. ANF generates density maps closer to the ground truth, and produces more accurate crowd counts. It is worth noting that although SANet [3] achieves higher counting accuracy, the results by our ANF closer to the ground truth in terms of density estimation.

Table 4. Comparison of model sizes and Performance

Method	# Parameter	PSNR	SSIM	MAE
MCNN [41]	0.13M	21.4	0.52	110.2
CP-CNN [28]	68.4M	21.72	0.72	73.6
CSRNet [19]	16.26M	23.79	0.76	68.2
ANF (Ours)	**7.9M**	**24.1**	**0.78**	**63.9**

Density Map Quality. To demonstrate that our method produces high quality density maps, we use the measurements of SSIM and PSNR to compare with representative methods, in Table 4. We compare with CP-CNN, CSRNet and MCNN which have released codes publicly available. Morever, CP-CNN and CSRNet also emphasize that they can generate high-quality density maps, as mentioned in Section 2, and MCNN [41] is one of the most representative methods in density estimation based crowd counting. Our method achieves the best performance in terms of both SSIM and PSNR. In addition, we have also compared by model sizes with parameter numbers in Table 4, which shows that our ANF has relatively low computational cost while performing the best of all.

5. Conclusion

In this work, we have presented the attentional neural fields (ANF) for crowd counting. The ANF integrates the conditional random fields and an attention mechanism into the convolutional encoder-decoder framework, which enhances their abilities to fuse multi-scale features and capture long-range dependencies. With both quantitative and qualitative results, we have demonstrated that the ANF can introduce consistent performance improvements on four popular datasets, including ShanghaiTech, WorldEXPO 10, UCF-CC-50 and UCF-QNRF, showing its great effectiveness for crowd counting.

Acknowledgment This paper was supported by National Key Research and Development Program of China under Grant 2016YFB1200100, National Key Scientific Instrument and Equipment Development Project under Grant 61827901, and Natural Science Foundation of China under Grant 91538204, 91738301, 61871016, 61571147.

References

[1] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. Crowdnet: A deep convolutional network for dense crowd counting. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 640–644. ACM, 2016.

[2] Antoni Buades, Bartomeu Coll, and J-M Morel. A non-local algorithm for image denoising. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 60–65. IEEE, 2005.

[3] 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.

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

[5] Ke Chen, Chen Change Loy, Shaogang Gong, and Tony Xiang. Feature mining for localised crowd counting. In *British Machine Vision Confernce (BMVC)*, volume 1, page 3, 2012.

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[7] Xiao Chu, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Crf-cnn: Modeling structured information in human pose estimation, 2016.

[8] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[9] Adam W Harley, Konstantinos G Derpanis, and Iasonas Kokkinos. Segmentation-aware convolutional networks using local attention masks. In *IEEE International Conference on Computer Vision (ICCV)*, volume 2, page 7, 2017.

[10] Yuanjun Huang, Xianbin Cao, Xiantong Zhen, and Jungong Han. Attentive temporal pyramid network for dynamic scene classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8497–8504, 2019.

[11] 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.

[12] 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. *arXiv preprint arXiv:1808.01050*, 2018.

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

[14] Dan Kong, Douglas Gray, and Hai Tao. A viewpoint invariant approach for crowd counting. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 1187–1190. IEEE, 2006.

[15] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[16] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[17] Min Li, Zhaoxiang Zhang, Kaiqi Huang, and Tieniu Tan. Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[18] Peizhao Li, Anran Zhang, Lei Yue, Xiantong Zhen, and Xianbin Cao. Multi-scale aggregation network for direct face alignment. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2156–2165. IEEE, 2019.

[19] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes, 2018.

[20] Xin Miao, Xiantong Zhen, Xianglong Liu, Cheng Deng, Vassilis Athitsos, and Heng Huang. Direct shape regression networks for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5040–5049, 2018.

[21] Viresh Ranjan, Hieu Le, and Minh Hoai. Iterative crowd counting. *arXiv preprint arXiv:1807.09959*, 2018.

[22] Kosta Ristovski, Vladan Radosavljevic, Slobodan Vucetic, and Zoran Obradovic. Continuous conditional random fields for efficient regression in large fully connected graphs. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention ®C MICCAI 2015*, page 234®C241, 2015.

[24] David Ryan, Simon Denman, Sridha Sridharan, and Clinton Fookes. An evaluation of crowd counting methods, features and regression models. *Computer Vision and Image Understanding*, 130:1–17, 2015.

[25] Sami Abdulla Mohsen Saleh, Shahrel Azmin Suandi, and Haidi Ibrahim. Recent survey on crowd density estimation and counting for visual surveillance. *Engineering Applications of Artificial Intelligence*, 41:103–114, 2015.

[26] Deepak Babu Sam, Shiv Surya, and R Venkatesh Babu. Switching convolutional neural network for crowd counting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 6, 2017.

[27] Vishwanath A Sindagi and Vishal 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 14th IEEE International Conference on*, pages 1–6. IEEE, 2017.

[28] Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns.

In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1879–1888. IEEE, 2017.

[29] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[30] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[31] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. *arXiv preprint arXiv:1704.06904*, 2017.

[32] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. *arXiv preprint arXiv:1711.07971*, 10, 2017.

[33] Ze Wang, Zehao Xiao, Kai Xie, Qiang Qiu, Xiantong Zhen, and Xianbin Cao. In defense of single-column networks for crowd counting. In *British Machine Vision Conference (BMVC)*, 2018.

[34] Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of CVPR*, volume 1, 2017.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.

[36] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[37] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659, 2016.

[38] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2412, 2018.

[39] Lei Yue, Xin Miao, Pengbo Wang, Baochang Zhang, Xiantong Zhen, and Xianbin Cao. Attentional alignment networks. In *British Machine Vision Conference (BMVC)*, 2018.

[40] 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.

[41] 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.

[42] Jiewan Zheng, Xianbin Cao, Baochang Zhang, Xiantong Zhen, and Xiangbo Su. Deep ensemble machine for video classification. *IEEE transactions on neural networks and learning systems*, 30(2):553–565, 2018.

[43] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.