A. Appendices

A.1. Bounding Box Generation

Given a vertex V_m in a cluster Ψ, we compute their coordinates in the input image $\omega_m = (i_m, j_m) \cdot \mu_{stride} + \mu_{offset}$, where $\mu_{stride} = 16$ and $\mu_{offset} = 8$. Then the bounding box size and orientation of each cluster are computed based on Principle Component Analysis (PCA). Given a set of coordinates $\omega = \{\omega_m | m = 1, 2, \cdots\}$ of a cluster , we compute the its eigenvectors θ_1 and θ_2 as well as the corresponding eigenvalues λ_1 and λ_2. The coordinates of the four corners of the bounding box is computed by:

$$\begin{aligned} c_1 &= A(\lambda_1 \cdot \theta_1 + \lambda_2 \cdot \theta_2) + \phi \\ c_2 &= A(\lambda_1 \cdot \theta_1 - \lambda_2 \cdot \theta_2) + \phi \\ c_3 &= A(-\lambda_1 \cdot \theta_1 - \lambda_2 \cdot \theta_2) + \phi \\ c_4 &= A(-\lambda_1 \cdot \theta_1 + \lambda_2 \cdot \theta_2) + \phi \end{aligned} \tag{14}$$

where ϕ is the center of the cluster and A denotes the scaling factor which is set to 1.75.

A.2. From Image to Stochastic Flow

Crucially, accurate object detection relies on correct flow prediction. In MCN, the flows f_0, f_1, f_2 and f_3 are the outputs of the Flow Mapping Layer (FML) with regional object probability P and correlation measurement S_1, S_2 and S_3 as inputs. P is generated by the Fore-/Background Network (FBN), while S_1, S_2 and S_3 are output by Local Correlation Network (LCN). Both FBN and LCN are starting at the *conv5_3* of VGG-16 pretrained network.

A.2.1 Fore-/Background Network

As shown in Figure 8 (a), the Fore-/Background Network is an FPN-based network [14] with spatial recurrent components and *softmax* output to predict the object score $P \in (0, 1)^{H_{1/16} \times W_{1/16}}$. The output of *conv5_3* is further processed by a Feature Pyramid Network (FPN) and a 2-dimensional Recurrent Neural Network (2D-RNN) successively. In FPN shown in Figure 8 (b), input with size of $H/16 \times W/16$ is processed by four convolutional blocks with 2×2 pooling layers to obtain additional feature maps with resolution of $1/32$, $1/64$, $1/128$ and $1/256$. These feature maps together with the input are fused to resolution of $1/16$ by deconvolution consisted of layer-wise addition, bilinear upsampling and convolution. By fusing features with different resolution in a pyramid manner, our method have larger capacity to detect multiscale objects with less parameters. Subsequently, the output of FPN is fed to an 2D Recurrent Neural Network (2D-RNN) before region-based classification. We consider a spatial feature map as a 2D sequence which can be directly analyzed by a 2D-RNN. The structure of the proposed 2D-RNN is shown in Figure 8 (c). A 2D-RNN is composed of two Bidirectional RNNs (RNN-H and RNN-V), which are applied to the rows and columns of the input feature map independently. As shown in Figure 8, the outputs of 2D-RNN is constructed by concatenating two feature maps produced by RNN-H and RNN-V with size of $H_{1/16} \times W_{1/16}$ along depth axis. Finally, a region-based classification is performed on the output feature map by a 2-layer convolutional network with *softmax* output, Figure 9 (d).

A.2.2 Local Correlation Subnetwork

To predict the semantic and spatial correlation between adjacent subregions, we build another subnetwork with additional four convolutional blocks and a *softmax* classifier starting at *conv5_3*, shown in Figure 9. The network outputs three correlation measurements S_1, S_2 and S_3 $\in (0, 1)^{H_{1/16} \times W_{1/16}}$ representing the semantic and spatial correlation between current anchor and its three neighbors (bottom, right and left) respectively. As the *conv5_3* features is corresponding to subregions of input image with stride of 16, the LCN is actually measuring the correlation among these overlapping subregions. Together with output of objetness network P, S_1, S_2 and S_3 are mapped to the Stochastic Flow f_0, f_1, f_2 and f_3 by Flow Mapping Layer (FML).

A.2.3 Flow Mapping Layer

The Flow Mapping Layer (FML) is point-wise non-linear function with input of P, S_1, S_2 and S_3 and output of f_0, f_1, f_2 and f_3. The mapping is shown below:

$$f_0 = e^{-\alpha[1-\mu(1-P)] \cdot [S_1^2 + S_2^2 + S_3^2]} \tag{15}$$

$$f_1 = (1 - f_0) \cdot \frac{S_1}{S_1 + S_2 + S_3} \tag{16}$$

$$f_2 = (1 - f_0) \cdot \frac{S_2}{S_1 + S_2 + S_3} \tag{17}$$

$$f_3 = (1 - f_0) \cdot \frac{S_3}{S_1 + S_2 + S_3} \tag{18}$$

$$\mu(x) = \frac{1}{1 + e^{-\beta(x-\gamma)}}. \tag{19}$$

Here, f_0 is actually the transition probability of *self-loop*, which is controlled by the likehood of background $(1 - P)$ and the correlation measurement between current vertex and its neighbors (S_1, S_2 and S_3). It is designed to be weak for vertices within the same object region and to be strong for a vertex which corresponds to the background or is just the attractor of a cluster. This behavior is realized by firstly measuring the correlation intensity ($S_1^2 + S_2^2 + S_3^2$) modulated by an *on-off* function $\mu(x)$, and then projecting it to the exponential space. $\mu(x)$ is parameterized by trainable

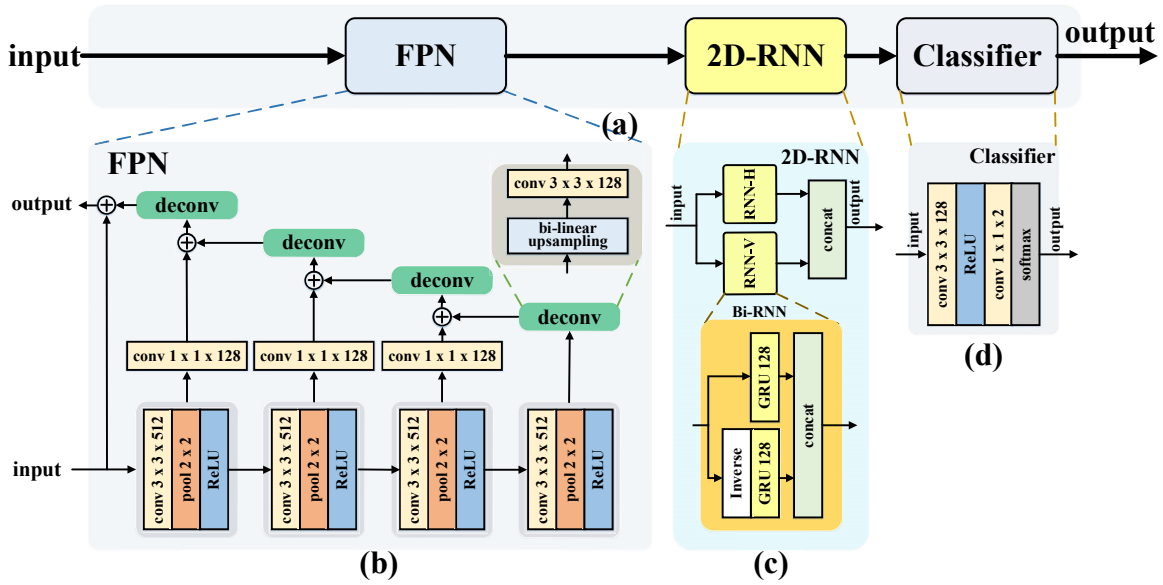

Figure 8. (a) Architecture of Fore-/Background Network (FBN); (b) Feature Pyramid Network (FPN) fusing feature maps with different resolutions; (c) 2-dimensional Recurrent Neural Network (2D-RNN) encoding contextual representations; (d) Regional objectness classifier predicting presence of an object with stride of 16×16 pixels.

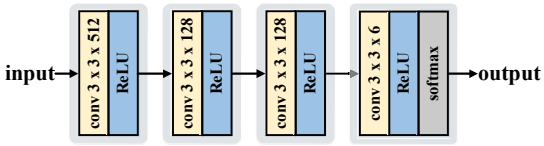

Figure 9. Local Correlation Network (LCN) with four convolution blocks and a softmax layer outputs correlation measurements S_1, S_2 and S_3 between current anchor and its three neighbors (bottom, right and left).

variables α, β and γ. It takes $1 - P$ as input and produces an on-off signal to control f_0. It will disables the effect of S_1, S_2 and S_3 and drive f_0 approaching to 1 when a vertex is in the background region. Accordingly, the values of f_1, f_2 and f_3 will be small, making all the background vertices to be isolated. In the object region, the correlation intensity S_1, S_2 and S_3 take control of f_0 since $1 - P$ is small. In this case, f_0 will be large if weak correlation is measured and the vertex will become the attractor of a cluster. Otherwise, the vectices belongs to the same object region will be connected through f_1, f_2 and f_3 and the flows of a cluster will end at the attractor.