Learning Scales from Points: A Scale-aware Probabilistic model for Crowd Counting

AKSHAD SHYAM

AI22MTECH02006

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

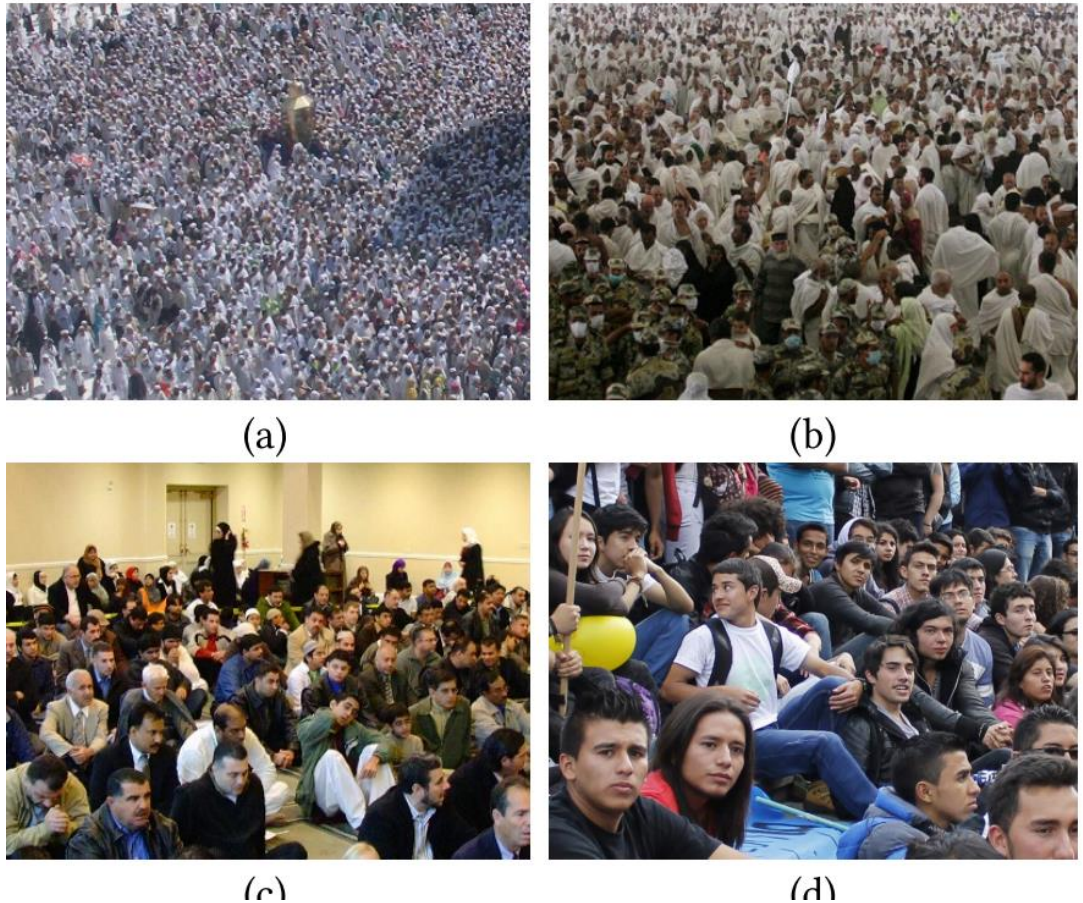
Abstract

- Paper Title: Learning Scales from Points: A Scale-aware Probabilistic Model for Crowd Counting by Zhiheng Ma et al 2020
- Counting people automatically through computer vision methods is challenging
- Due to large scale variations of instances caused by frame perspective.
- The paper presents a density pyramid network, where each level handles instances within a particular scale range.
- A proposed scale distribution estimator to learn scales of people from input data
- Also, an instance level probabilistic scale-aware model to guide the multi-scale training of the DPN.

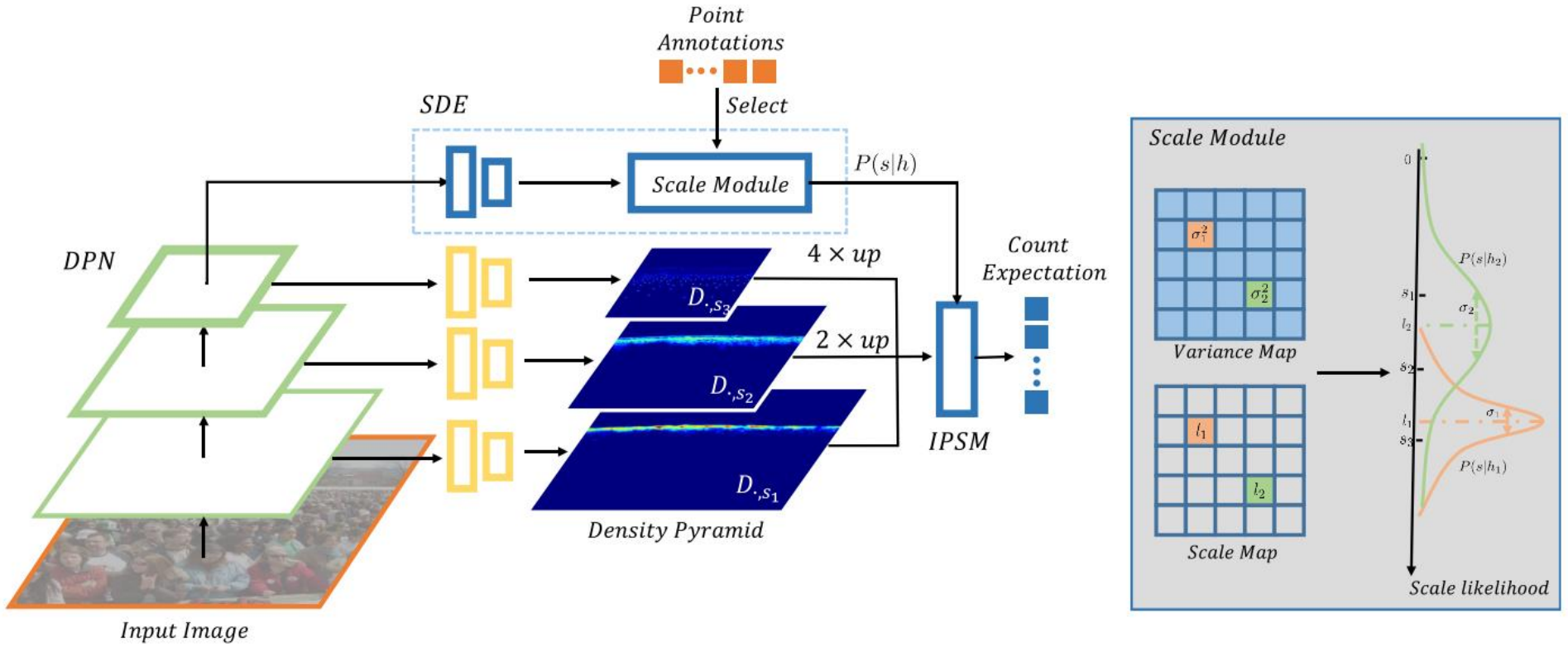
Background and Motivation

- Feature pyramid networks along with instance-level hard assignment is used to handle scale variations in object detection methods.
- Smaller objects are assigned to a finer-resolution pyramid level and vice versa
- But this only works if we have accurate scale supervision available
- Datasets which involve crowds of instances (crowd counting) usually annotate the instances with points, hence no scale supervision is present
- Hence a scale-aware probabilistic model along with the dense pyramid network is proposed

Targeted Data



Proposed Network



Scale Aware Probabilistic Model

- Target: To train the DPN in a scale-aware manner, ensuring each pyramid level focuses on a specific scale range.
- Three parts – Density Pyramid, SDE ,IPSM
- Scale of an instance *h* is modelled as a learnable gaussian distribution, whose mean and variance are predicted by SDE.
- The specific mean and variance are selected according to the spatial coordinates of the instance(given by the annotated point) to give the scale likelihood.

$$P(s|h) = \mathcal{N}(s|l_h, \sigma_h^2)$$
$$= \frac{1}{\sqrt{2\pi V_{y_h}}} \exp\left(-\frac{(s - Ly_h)^2}{2V_{y_h}}\right).$$

- The DPN levels are classified on the basis of scale set S = {$\log_2 8$, $\log_2 16$,$\log_2 32$}. Thus the scale likelihood P($s|h$) can be interpreted as assigning *h* to the pyramid level of the scale *s,s* from the subset S.
- The count expectation c_h of the person *h* can be calculated as

$$c_h = \sum_{x \in \mathcal{X}} \sum_{s \in \mathcal{S}} P(h|x, s) D_{x,s}.$$

- Here P(h|x,s) is the posterior probability where x represents the spatial coordinates. The ground truth value is 1.

- The posterior probability is derived from scale likelihood P(s|h) and spatial likelihood P(x|h), where P(x|h) is defined as a gaussian distribution centered on the annotated point with constant variance.
- Assuming P(h) to be a uniform distribution and P(s|h),P(x|h) to be conditionally independent we get the below equation.
- The count loss is to minimize the discrepancy between estimated and ground-truth count expectation.

$$P(h|x,s) = \frac{P(x,s|h)P(h)}{\sum_{h\in\mathcal{H}} P(x,s|h)P(h)} = \frac{P(x|h)P(s|h)}{\sum_{h\in\mathcal{H}} P(x|h)P(s|h)}.$$

$$\mathcal{J}^{count} = \sum_{h\in\mathcal{H}} \|c_h - \hat{c}_h\|_p.$$

- The count loss is further refined by taking the background into consideration. The background is treated as a negative class b and its ground-truth expectation should be 0.
- An intuitive assumption made here is that the pixel far from any annotated point should belong to the background, so the background spatial distance of a pixel is inversely related to the spatial distance between the pixel and its closest head point.
- The spatial likelihood of the background is given as

$$\bar{x} = \frac{\alpha}{\min_{h \in \mathcal{H}} \|x - y_h\|_2^2 + \epsilon},$$

$$P(x|b) = \frac{1}{\sqrt{2\pi}\beta} \exp(-\frac{\bar{x}^2}{2\beta^2}),$$

- The derivation of the posterior probability including the background is given below.

$$P(s|b) = \mathcal{N}(s|0, \sigma_b^2).$$

$$P(b|x,s) = \frac{P(x|b)P(s|b)}{P(x|b)P(s|b) + \sum_{h\in\mathcal{H}} P(x|h)P(s|h)}, \qquad P(h|x,s) = \frac{P(x,s|h)P(h)}{P(x,s|b)P(b) + \sum_{h\in\mathcal{H}} P(x,s|h)P(h)}$$

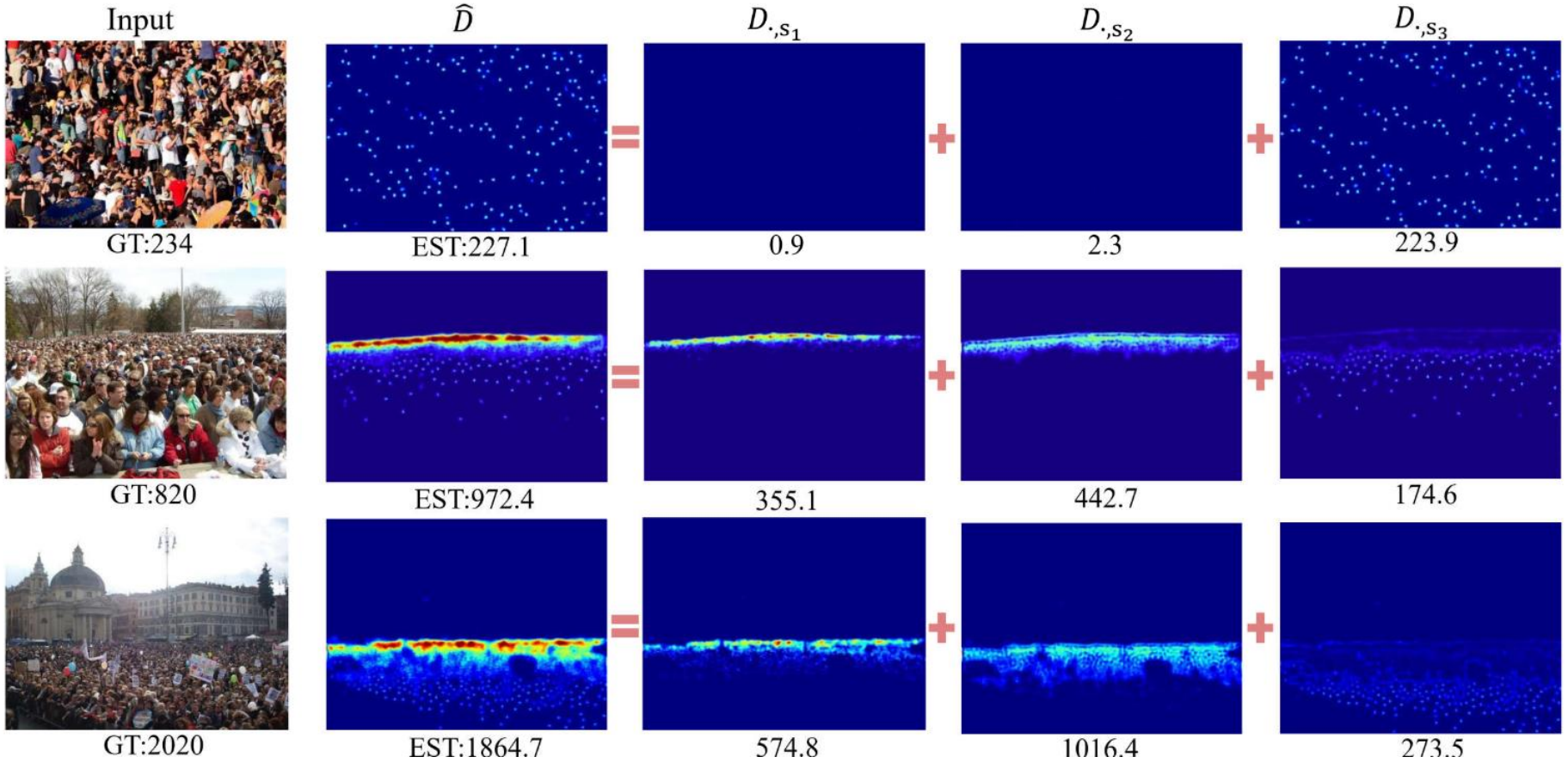
$$= \frac{P(x|h)P(s|h)}{P(x|b)P(s|b) + \sum_{h\in\mathcal{H}} P(x|h)P(s|h)},$$

$$c_b = \sum_{x\in\mathcal{X}} \sum_{s\in\mathcal{S}} P(b|x,s) D_{x,s}.$$

- The complete count loss is to regress count expectations of the foreground instances and the background to their ground-truth counts.
- The scale loss is to estimate the scale likelihood under weak supervision of annotated points.
- The final target is to estimate the total count of the people in a image. Let C denote the total count. It can be calculated by summing up the count in DP layers.

$$
\begin{aligned}
C &= \sum_{h \in \mathcal{H}} c_h = \sum_{h \in \mathcal{H}} c_h + c_b \\
&= \Big(\sum_{h \in \mathcal{H}} P(h|x,s) + P(b|x,s) \Big) \sum_{x \in \mathcal{X}} \sum_{s \in \mathcal{S}} D_{x,s} \\
&= \sum_{x \in \mathcal{X}} \sum_{s \in \mathcal{S}} D_{x,s} .
\end{aligned}
$$

Output Images



Visualization of the various levels of the dense pyramid, from low to high resolution.