DMPNet: Densely connected multi-scale pyramid networks for crowd counting

Pengfei Li [Corresp., 1] , **Min Zhang** [Corresp., 1] , **Jian Wan** [1] , **Ming Jiang** [1]

[1] Computer & Software School, Hangzhou Dianzi University, Hangzhou, Zhejiang, China

Corresponding Authors: Pengfei Li, Min Zhang
Email address: lipf@hdu.edu.cn, hz_andy@163.com

Crowd counting has been widely studied by de ep learning in recent years. However, due to scale variation caused by perspective distortion, crowd counting is still a challenging task. In this paper, we propose a Densely Connected Multi-scale Pyramid Network (DMPNet) for count estimation and the generation of high-quality density maps. The key component of our network is Multi-scale Pyramid Network (MPN), which can extract multi-scale features of the crowd effectively while keeping the resolution of the input feature map and the number of channels unchanged. To increase the information transfer between the network layer, we use dense connections to connect multiple MPNs. In addition, we also design a novel loss function, which can help our model achieve better convergence. To evaluate our method, we conduct extensive experiments on three challenging benchmark crowd counting datasets. Experimental results show that compared with the state-of-the-art algorithms, DMPNet performs well in both parameters and results. Code is available at: https://github.com/lpfworld/DMPNet .

1   DMPNet: Densely Connected Multi-scale Pyramid
2   Networks for Crowd Counting

3   Pengfei Li[1], Min Zhang[1], Jian Wan[1], Ming Jiang[1]

4   [1] Computer & Software School, Hangzhou Dianzi University, Hangzhou, Zhejiang Province,
5   China

6   Corresponding Author: Min Zhang

7   Baiyang Road #2, Hangzhou, Zhejiang Province, 310018, China

8   Email address: hz_andy@163.com

9   ABSTRACT

10   Crowd counting has been widely studied by deep learning in recent years. However, due to scale
11   variation caused by perspective distortion, crowd counting is still a challenging task. In this
12   paper, we propose a Densely Connected Multi-scale Pyramid Network (DMPNet) for count
13   estimation and the generation of high-quality density maps. The key component of our network
14   is Multi-scale Pyramid Network (MPN), which can extract multi-scale features of the crowd
15   effectively while keeping the resolution of the input feature map and the number of channels
16   unchanged. To increase the information transfer between the network layer, we use dense
17   connections to connect multiple MPNs. In addition, we also design a novel loss function, which
18   can help our model achieve better convergence. To evaluate our method, we conduct extensive
19   experiments on three challenging benchmark crowd counting datasets. Experimental results
20   show that compared with the state-of-the-art algorithms, DMPNet performs well in both
21   parameters and results. Code is available at: https://github.com/lpfworld/DMPNet.

22   **Keywords** Crowd counting, Density map, Multi-scale, Pyramid convolution, Group convolution

23   INTRODUCTION

24   With the increase of the world population, crowd counting has been widely applied in video
25   surveillance, crowd analysis, sporting events, and other public security services (Chan, Liang &
26   Vasconcelos, 2008; Boominathan, Kruthiventi & Babu 2016; Cao et al.,2018; Xiong et al.,2019).
27   In addition, it is extended to cell or bacterial counts in the medical field and vehicle counts in
28   transportation field (Xie, Noble & Zisserman, 2018; Hu et al., 2020). However, crowd counting
29   still is a challenging task due to scale variations, cluttered backgrounds, and heavy occlusion.
30   Among them, scale variation is the most important research issue, as shown in Figure 1.

31   CNN-based (Convolutional Neural Networks, CNN) methods have made remarkable progress in
32   crowd counting in recent years. To extract multi-scale features of crowds, researchers designed
33   multi-column or multi-branch networks (Sam, Surya & Babu, 2016; Zhang et al., 2016; Liu,
34   Salzmann & Fua, 2019; Jiang, Zhang & Xu, 2020). However, most networks are limited in their
35   ability to extract multi-scale features due to the similarity of different columns or branches
36   (Zhang et al., 2016; Sam et al., 2016, Zhang et al., 2018). In addition, multi-scale extraction
37   modules in these networks require a lot of computation because of the complexity of the network

38   structure (Li, Zhang & Chen, 2018; Guo et al., 2019). Our MPN also adopts a multi-branch
39   structure to ensure multi-scale feature extraction, in which pyramid convolution and group
40   convolution are used to effectively reduce parameters.

41   The higher resolution feature map contains finer details and the resulting density map is of
42   higher quality, which is helpful for count estimation (Cao et al., 2018; Jia, Antoni & Chan,
43   2019). To increase receptive fields of networks, pooling operations are adopted. However, the
44   resolution of feature maps generated by the network become smaller, resulting in the loss of
45   crowd image details. To keep the input and output resolutions unchanged, the encoder-decoder
46   structure is usually utilized (Jiang et al.,2019; Thanasutives et al., 2021). The network of
47   encoder-decoder structure uses encoder to extract input image features and combine them, and
48   then decodes the higher-level features required by these features through a specially designed
49   decoder. Take M-SFANet (Multi-Scale-Aware Fusion Network with Attention mechanism)
50   (Thanasutives et al., 2021) for example, the encoder of M-SFANet (Thanasutives et al., 2021) is
51   enhanced with ASSP (Atrous Spatial Pyramid Pooling, ASSP) (Chen et al.,2017), which can
52   extract multi-scale features of the target object and fuse large context information. In order to
53   further deal with the scale variation of the input image, they used the context module called CAN
54   (Context Aware Network, CAN) (Liu et al.,2019) as the decoder. Similar to these works, we
55   keep the input and output resolutions of MPN unchanged to ensure that the final density map
56   generated by DMPNet contains sufficient detailed crowd information.

57   Different layers of neural network contain different crowd information, but with the increase of
58   network depth, some details are gradually lost. DSNet (Dense Scale Network, DSNet) (Dai et al.,
59   2021) proposed that using dense connected networks in the field of crowd counting can
60   effectively extract long-distance context information and maximize the retention of network
61   layer information. We follow this operation and connect MPNs with dense connections.

62   Euclidean loss is the most common loss function in crowd counting SOTA methods, which is
63   based on pixel independence (Cao et al., 2018; Liu et al.,2020; Zhang et al.,2020). However,
64   texture features and pixel correlation of different regions in crowd images are different.
65   Euclidean loss ignores the local correlation of the crowd image and does not consider the global
66   counting error of the crowd image (Cao et al., 2018; Dai et al., 2021). Therefore, when designing
67   the loss function, we not only consider the local density consistency of the image, but also
68   consider the global counting loss of the image.

69   In this paper, we propose the Densely connected Multi-scale Pyramid Network (DMPNet) for
70   crowd counting, as shown in Figure 2. The important component Multi-scale Pyramid Network
71   (MPN) consists of Local Pyramid Network (LPN), Global Pyramid Network (GPN), and Multi-
72   scale Feature Fusion Network (MFFN). LPN is used to capture small heads and extract multi-
73   scale fine-grained features, while GPN is used to capture large heads and global features. They
74   are composed of multiple levels, and each level has filters of different sizes and depths, whose
75   output local and global features are combined by MFFN. To maximize the flow of information
76   between layers of the network, MPNs in the network are densely connected, with each MPN

77 receiving as input the results of MPNs before it. To optimize the loss function, we combine
78 Euclidean loss, density level consistency loss, and MAE loss to improve the performance of
79 DMPNet. Experiments on three datasets (ShanghaiTech Part A and Part B (Zhang et al., 2016),
80 UCF-QNRF (Idrees et al., 2018), UCF_CC_50 (Idrees et al., 2013) prove the effectiveness and
81 robustness of the proposed method.

82 ## RELATED WORK

83 Generally, the existing crowd counting methods can be mainly classified into two categories:
84 traditional methods and CNN-based methods (Sindagi & Patel, 2017; Gao et al.,2020). In this
85 section, we give a brief review of crowd counting methods and explain the differences between
86 our methods.

87 ## Traditional methods

88 In early studies, detection-based methods used sliding windows to detect the target, and
89 manually extract features of the human body or specific body parts (Wu & Nevatia, 2007;
90 Enzweiler & Gavrila,2009; Felzenszwalb et al., 2010). However, even if only heads or smaller
91 body parts of pedestrians are detected, these methods often fail to make accurate counts of dense
92 crowd scenes due to occlusion and illumination. To improve the performance of crowd counting,
93 feature-based regression methods attempted to extract various features from local image blocks
94 and generate low-level information (Chan & Vasconcelos, 2009; Ryan et al.,2009; Ke et al.,
95 2012). Idrees et al. (2013) tried to fuse the features obtained by Fourier analysis and SIFT (Scale-
96 invariant feature transform, SIFT) interest points. However, they ignored the scale information.
97 To overcome the problem, density estimation-based method considers the relationship between
98 image features and data regression. Victor & Andrew (2010) adopted the method of extracting
99 features in local areas and establishing linear mapping between features and density maps. Pham
100 et al. (2015) tried to use random forest regression to get a nonlinear map.

101 ## CNN-based methods

102 The CNN-based methods can be classified into the multi-column CNN-based methods and the
103 single-column CNN-based methods. The multi-column CNN-based methods use multi-column
104 networks to extract the human head features of different scales and then fuse them to generate
105 density maps. Zhang et al. (2016) (Multi-Column Convolutional Neural Network, MCNN)
106 proposed to extract features using three-column networks with convolution kernels of different
107 sizes respectively, and fused them through 1x1 convolution. Sam et al. (2016) (Switching
108 Convolutional Neural Network, Switch-CNN) proposed to design an additional switch based on
109 MCNN, that is to use the switch to select the most appropriate CNN column for different input
110 images to improve the counting accuracy. Inspired by the image generation methods, Viresh et
111 al. (2018) (Iterative Crowd Counting CNN, ic-CNN) proposed a two-column networks to
112 gradually refine the obtained low-resolution density map to high-resolution density map. Sindagi
113 et al., (2017) (Contextual Pyramid CNN, CP-CNN) used global and local feature information to
114 generate density maps for crowd images. Zhu et al., (2020) (Relational Attention Network,

115 RANet) proposed to use the stacked hourglass structure in human pose, optimized outputs from
116 each hourglass module with local attention LSA and global attention GSA, and then fused the
117 two features with a relational module. Zhu et al. (2019) (Multi-Scale Fusion Network with
118 Attention mechanism, SFANet) proposed a dual path multi-scale fusion network architecture
119 with attention mechanism, which contains a VGG as the front-end feature map extractor and a
120 dual path multi-scale fusion networks as the back-end to generate density map. Jiang et al.
121 (2020) (Attention Scaling Network, ASNet) proposed to use different columns to generate
122 density maps and scale factors, then multiply them by the mask of the region of interest to output
123 multiple attention-based density maps, and add the density maps to obtain a high-quality density
124 map. These methods have a strong ability in extracting multi-scale features and improving the
125 performance of crowd counting. However, they also have some disadvantages: these networks
126 usually have a lot of parameters, and the similarity of networks with different columns results in
127 limited feature extraction ability. In addition, training multiple CNNs at the same time will lead
128 to slower training speed (Zhang et al., 2018; Cao et al., 2018; Jiang et al., 2020).

129 The single-column CNN methods try to use the multi-branch structure for optimization, which
130 can extract multi-scale information and effectively reduce parameters (Zhang et al., 2018; Cao et
131 al., 2018; Liu et al., 2019). Zhang et al. (2018) (Congested Scene Recognition Network,
132 CSRNet) proposed the network structure of the front and back end, in which the front-end
133 network adopts VGG16 (Simonyan & Zisserman, 2014), and the back-end network uses dilated
134 convolution to increase the receptive field and extract multi-scale features. Cao et al. (2018)
135 (Scale Aggregation Network, SANet) proposed to extract multi-scale features by using
136 convolution containing multiple levels, and the convolution kernel of each level is different in
137 size. At the back end of SANet, the resolution of the feature map is restored to the size of the
138 input image by deconvolution, and the final density map is obtained. Liu et al. (2019) (Context
139 Aware Network, CAN) proposed a pooling pyramid network to extract multi-scale features and
140 adaptively assign weights to crowd regions of different scales in images. Shi et al., (2019)
141 (Perspective-Aware CNN, PACNN) proposed a perspective-aware network, which can integrate
142 the perspective information into density regression to provide additional knowledge of scale
143 variations in images. Miao et al., (2020) (Shallow feature based Dense Attention Network,
144 SDANet) proposed to reduce the influence of background by introducing an attentional model
145 based on shallow features, and to capture multi-scale information through dense connections of
146 hierarchical features. Thanasutives et al., (2021) (M-SFANet) proposed to use ASPP (Chen et al.,
147 2017) containing parallel atrous convolutional layers with different sampling rates to enhance the
148 network, which can extract multi-scale features of the target object and incorporate larger
149 context. Jiang et al., (2019) (Trellis Encoder-Decoder Network, TEDNet) proposed to build
150 multiple decoding paths in different coding stages to aggregate features of different layers. Ma et
151 al., (2019) (Bayesian Loss, BL) regarded crowd counting as a probability problem, the predicted
152 density map is a probability map, each point represents the probability of existence at the point,
153 and each point of the density map is regarded as the sample observation value.

154   Our DMPNet is a single-column network with multi-branch, similar to some works (Cao et al.,
155   2018; Liu, Salzmann & Fua, 2019; Dai et al., 2021). We differ them from three aspects: (1) Each
156   branch of our convolution kernel is not only different in size, but also different in the number of
157   channels, which improves the ability of feature extraction of similar networks. (2) We use group
158   convolution to process convolution kernels of different sizes, effectively reducing network
159   parameters, and the calculation process is similar to Google MixNet (Mixed Depthwise
160   Convolutional Network, MixNet) (Tan & Le, 2019). (3) Our DMPNet is an end-to-end
161   architecture, without adding extra perspective maps or attention maps.

162   ## METHODS

163   The basic idea of our approach is to implement an end-to-end network that can capture multi-
164   scale features and generate a high-quality density map, to achieve accurate crowd estimation. In
165   this section, we first introduce our proposed DMPNet architecture, then present our loss function.

166   ## DMPNet architecture

167   Similar to CSRNet (Li, Zhang & Chen, 2018), our DMPNet architecture includes a front-end
168   network and a back-end network (see Figure 2). In the front-end network, the first ten layers with
169   three pooling layers of VGG16 are used to extract features from crowd images. Several works
170   have proved that VGG16 achieves a trade-off between accuracy and computation, and is suitable
171   for crowd counting (Cao et al., 2018; Tanjan, Le & Hoai,2018; Jia, Antoni & Chan, 2019). In the
172   back-end network, MPNs that can extract multi-scale features are connected in a dense way to
173   improve information flow between layers. The integration between the different layers in the
174   network can also be further retained multi-scale features (Huang et al., 2016; Miao et al., 2020;
175   Amaranageswarao, Deivalakshmi & Ko, 2020). In ablation experiments, we demonstrated the
176   effectiveness of dense connections.

177   ## Multi-scale Pyramid Network (MPN)

178   MPN consists of three parts: Local Pyramid Network (LPN), Global Pyramid Network (GPN),
179   and Multi-scale Feature Fusion Network (MFFN), illustrated in Figure 2. The design principle of
180   MPN is to keep the resolution and channel number of input and output features unchanged, and
181   effectively extract multi-scale features.

182   ## Pyramid Convolution and Group Convolution

183   Pyramid convolution has been applied to image segmentation, image classification and other
184   fields, and achieved remarkable results (Lin et al., 2017; Duta et al., 2020; Wang et al., 2020;
185   Richardson et al., 2020). Inspired by this, we propose to apply pyramid convolution to crowd
186   counting. Compared to standard convolution, pyramid convolution is composed of convolution
187   kernels of different sizes and depths in N level, without increasing computational cost and
188   complexity, illustrated in Figure 3.

189 Each level of pyramid convolution is computed with all input features. To use different depths of
190 the kernels at each level of pyramid convolution, we do this using group convolution. The input
191 features are divided into four groups, and the convolution kernels are applied separately for each
192 input group, illustrated in Figure 4.

193 We compare the parameters of standard convolution and group convolution. (1) Standard
194 convolution contains a single type of convolution kernel (with height K, width K), and the depth
195 is equal to the number of channels of input features C_1. C_2 such convolution kernels and input
196 features (with height H, width W) are calculated to get output features (with height $H^{'}$, width $W^{'}$
197). Therefore, the parameter number of standard convolution is $k^2 C_1 C_2$. (2) Group Convolution
198 divides the input feature map (with height H, width W) into g groups, the depth is equal to the
199 number of channels of input features C_1, and then performs convolution calculation within each
200 group. The convolution kernels (with height K, width K, and the number of channels C_2) are also
201 divided into corresponding g groups. Each group of convolution generates feature maps (with
202 height W', width H', and the number of channels C_2/g). Therefore, the parameter number of

203 group convolution is $k^2 * \left(\frac{C_1}{g}\right) * \left(\frac{C_2}{g}\right) * g = k^2 C_1 C_2 / g$. The width and height of the output depend

204 on the convolution step size, and these two values are not considered here. The above calculation
205 results prove that group convolution can generate feature maps with fewer parameters. The more
206 feature maps, the more information that can be encoded for the crowd counting network.

207 LPN, GPN, and MFFN

208 Based on the ability of pyramid convolution and group convolution, we design LPN and MPN to
209 extract local and global features of crowd images, and use MFFN to combine the two, as shown
210 in Figure 5.

211 (a) LPN is mainly used for fine-grained feature extraction. Detailed information is shown in
212 Figure 5(a). First, we use 1x1 convolution to reduce the channel of F_I to 512. Then, four-level
213 pyramid convolution with different convolution kernels sizes (9x9,7x7,5x5, and 3x3) is used to
214 extract multi-scale features. The corresponding channel number is 32,64,128,256, and the group
215 convolution size is 16,8,4,1. Finally, we use 1x1 convolution to increase the channel numbers of
216 the four-level features to 512, and the output feature F_L is obtained. All convolution operations
217 are followed by BN and ReLU.

218 (b) GPN is mainly used for coarse-grained feature extraction. Detailed information is shown in
219 Figure 5(b). The intermediate processing of GPN and LPN is the same, but the difference is that
220 the input feature F_I first goes through a layer of 9x9 adaptive average pool to ensure that
221 complete global information can be obtained. In addition, to restore the resolution of the output
222 feature map, we use bilinear interpolation for up-sampling to obtain the final output F_G.

223   (c) MFFN is mainly used for global and local feature fusion (fine-grain and coarse-grain
224   features). Detailed information is shown in Figure 5(c). First, the output of LPN and GPN is
225   combined into the features with 1024 channels as the input of MFFN. Then, through a layer of
226   3x3 convolution output the features of 256 channels. Finally, we use 1x1 convolution to restore
227   the channel numbers to 512 and obtain feature F_o.

Loss function

229   Euclidean loss is the most common loss function in crowd counting. It evaluates the difference
230   between the ground truth and the estimated density map based on pixel independence, without
231   considering the local density correlation of images. However, the local features of the crowd are
232   generally consistent. In addition, Euclidean loss does not consider the counting error of the
233   image (Cao et al., 2018; Dai et al., 2021). Therefore, we combine density-level consistency loss
234   and MAE loss with Euclidean loss in the loss function.

Euclidean loss

236   Euclidean loss can estimate the pixel-level error between the estimated density map and the
237   ground truth. It is the most common loss function in crowd counting. The Euclidean loss
238   function can be defined as follow:

239    $L_\mathrm{E} = \frac{1}{N}\Sigma_{i=1}^{N}||\mathrm{F}(X_i;\theta) - F_i||_2^2$

240   Where N is the size of training batch, θ is the variable parameters of DMPNet. X_i is the input
241   image, F_i represent the ground truth, and $\mathrm{F}(X_i;\theta)$ is the output of DMPNet.

Density level consistency loss

243   Due to the imbalance of crowd distribution, the density map has a local correlation, and the
244   density level of different sub-regions is not the same. Therefore, the density map generated by
245   the model should be consistent with the ground truth (Jia, Antoni & Chan, 2019; Jiang et al.,
246   2020). Referring to the setting of reference (Dai et al., 2021), we divide the density map into sub-
247   regions of different sizes and formed pool representations. Three outputs of different sizes are
248   used (1x1,2x2,4x4), with 1x1 representing the global density level of the density map and the
249   other two representing the density level of different local sizes in the density map. The density
250   level consistency loss can be defined as follow:

251
$$\mathrm{L_D} = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}\sum_{j=1}^{S}\frac{1}{k_j^2}||P_{ave}\big(\mathrm{F}(X_i;\theta),k_j\big) - P_{ave}(D_i^{GT},k_j)||_1$$

252   Where S represents the number of scale levels, P_{ave} is the average pooling operation, and k_j
253   represents the specified output size of average pooling.

MAE loss

255 Mean absolute error (MAE) loss can estimate the real count and the estimated count. The MAE
256 loss can be defined as follow:

257 $L_A = \frac{1}{N}\sum_{i=1}^{N}|C(I_i) - C^{GT}(I'_i)|$

258 where I_i and I'_i represent the density map generated by DMPNet and the real density map of X_i

259 separately. C represents the sum of all pixels. $C(I_i)$ and $C^{GT}(I'_i)$ represent the estimated count and

260 the real count of X_i separately.

261 *The final loss*

262 The final loss consists of L_S, L_c, and L_E. α and β are weighting factors of L_S and L_c. According to

263 our experiments, they are set as 10-4 and 10-3, separately.

264 $$L(\Theta) = L_E + aL_D + \beta L_A$$

265 **EXPERIMENTAL AND DISCUSSION**

266 **Training methods**

267 *Ground truth generation*

268 The ground truth density map can represent the image containing N people. Following the
269 methods in (Zhang et al., 2016; Liu, Salzmann & Fua, 2019; Jiang et al., 2020), We convolve $\delta($
270 $x - x_i$) with a Gaussian kernel $G_{\sigma_i}(x)$ (which is normalized to 1) with parameter σ_i to blur each

271 head annotation. The ground truth density map can be defined as follow:

272 $F(x) = \sum_{i=1}^{N}\delta(x - x_i) * G_{\sigma_i}(x),$ with $\sigma_i = \beta\overline{d^i}$

273 Where x_i represents the position of pixel, $\overline{d^i}$ is the average distance of k nearest neighbors, β is a
274 constant. We set k=3 and β=0.3. σ_i is standard deviation, the setups are shown in Table 1.

275 *Training details*

276 Our DMPNet is implemented based on the PyTorch framework. It consists of a front-end
277 network with the first 10 layers of VGG16 (Simonyan & Zisserman, 2014) and a back-end
278 network with three densely connected MPNs. The training batch size is 1, optimized by Adam
279 (Kingma & Ba, 2014), and the learning rate is 5e-6 and the weight decay of 5e-4. Random
280 Gaussian initialization with 0.01 standard deviation is used. Besides, we perform data
281 enhancement on the image, and the enhancement principle followed CSRNet (Li, Zhang & Chen,
282 2018). Considering the illumination changes, we carry out gamma transform and gray transform
283 on images, and the transformation principle follows DSNet (Dai et al., 2021).

284 **Evaluation metrics**

285   Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) can evaluate the

286   performance of crowd counting (Jia, Antoni & Chan, 2019; Jiang et al., 2020). MAE and RMSE

287   represent the accuracy and robustness of the network respectively, and they can be defined as

288   follows:

289     $MAE = \frac{1}{N}|D_i - D_i^{GT}|$

290     $RMSE = \sqrt{\frac{1}{N}\Sigma_{i=1}^{N}(D_i - D_i^{GT})^2}$

291   where N is the number of test images. D_i and D_i^{GT} represent the actual and estimated numbers of

292   people in the i-th image respectively.

Datasets

294   We evaluate DMPNet on three benchmark crowd counting datasets: ShanghaiTech (Zhang et al.,

295   2016), UCF-QNRF (Idrees et al., 2018), UCF CC 50 (Idrees et al., 2013). (1) ShanghaiTech: It

296   includes Part A and Part B, with a total 1,198 images and 330,165 annotations. Part A contains

297   300 training images and 182 testing images for congested crowd scenes, counting from 33 to

298   3139. Part B contains 400 training images and 316 testing images, for sparse crowd scenes,

299   counting from 9 to 578. (2) UCF-QNRF: It is the largest and most recently released dataset on

300   crowd counting with 1,535 dense crowd images from various websites, counting from 49 to

301   12865. (3) UCF CC 50: It contains 50 images with 63974 annotations, counting from 94 to 4543.

302   The average number of people in the image is 1280.

Comparison with State-of-the-Art

304   We evaluate and compare our DMPNet and SOTA methods on three challenging crowd counting

305   datasets. The experimental results are shown in Table 2. As you can see, our DMPNet is in the

306   top two in multiple comparisons. (1) On ShanghaiTech part A (Zhang et al., 2016), MAE of our

307   model is 98.3, which is the second best result. RMSE is 63.7, 7.2 % higher than that of the

308   optimal model RANet (Zhu et al., 2019). On ShanghaiTech Part B (Zhang et al., 2016), MAE

309   and RMSE are 13.4% and 15.6% higher than DSNet (Dai et al., 2021) and SDANet (Miao et al.,

310   2020), respectively. The images of Part A are from the Internet with highly congested scenes.

311   The images of Part B come from streets captured by fixed cameras with relatively sparse crowd

312   scenes. It indicates that our DMPNet can perform well both congested and sparse crowd scenes.

313   (2) On UCF_QNRF (Idrees et al., 2018), although we do not reach the best, we still have a good

314   performance. MAE and RMSE are 98.7 and 179.8, respectively, 15.3% and 18.9% higher than

315   M-SFANet (Thanasutives et al., 2021). UCF_QNRF has lots of different scenes, in which the

316   viewpoint and lighting variations are more diverse. In addition, due to the great change of crowd

317   density, the perspective distortion of the head is more serious. Our model can handle this data

318   well, which proves that our model has a certain adaptability to multiple scenes. In the face of

319   crowd images close to real high-density scenes in UCF_QNRF, DMPNet can produce more

320   accurate counting. (3) On UCF_CC_50 (Idrees et al., 2013), 5-fold cross-validation is used to

321   evaluate our DMPNet and achieve the second-best results of MAE and RMSE, 24.7% and 25.3%
322   higher than M-SFANet (Thanasutives et al., 2021) and DSNet (Dai et al., 2021) respectively.
323   UCF_CC_50 is a challenging dataset with few samples and low image resolution. The results of
324   this data demonstrate that we can also achieve high results on small datasets.

325   The visualization results of our DMPNet are shown in Figure 6, and the quality of density maps
326   generated by DMPNet and SOTA methods is compared on ShanghaiTech Part A and Part B
327   datasets (Zhang et al., 2016) are shown in Figure 7. The comparison of visualization results and
328   counting results shows that DMPNet can extract different types of crowd image information, and
329   the density map is closer to the ground truth and higher in counting accuracy than MCNN
330   (Zhang et al., 2016) and CSRNet (Li, Zhang & Chen, 2018). Our DMPNet has well solved the
331   problems of crowd occlusion, perspective distortion, and scale variations.

Ablation experiments

333   In this subsection, we perform several ablation experiments including Multi-scale Pyramid
334   Network (i.e., LPN, GPN, and LPN+GPN), connected network (i.e., dense connection and
335   without dense connection), and loss function. Following the previous works (Li, Zhang & Chen,
336   2018; Jiang et al., 2020; Zhang et al., 2020), ablation experiments are conducted on
337   ShanghaiTech Part A (Zhang et al., 2016).

Effect of LPN and GPN

339   To verify the effects of LPN and GPN, we adjust the network structure with three different
340   combinations. The results of LPN and GPN are summarized in Table 3. In comparison, LPN
341   achieves better results than GPN, with MAE and MSE lower 4.4% and 7.1%, respectively. When
342   the two networks are used together, the results are further reduced by 5.4% and 6.7% relative to
343   LPN. The results show that the proposed multi-scale extraction module is effective in capturing
344   coarse-grained and fine-grained scales.

Effect of Dense connection

346   To verify the effects of dense connections, we compare two structures, one with dense
347   connections and the other without dense connections, and the results are shown in Table 4.
348   Results are significantly better when dense connections are used, with MAE and MSE decreasing
349   by 6.8% and 9.5%, respectively. This indicates that dense connection effectively prevents feature
350   loss, increases information flow between different network layers, further enlarges scale
351   diversity, and makes the feature more effective.

Effect of loss function

353   To verify the effect of different loss function combinations, we design four different
354   combinations, and the results are shown in Table 5. MSE Loss, as the most common loss
355   function in crowd counting, still plays a major role. However, after density level consistency loss
356   and MAE loss are added, the effect is improved to a certain extent. When both are used, MAE

357 and MSE decrease by 9.0% and 9.3%, respectively, indicating that the combination of density
358 level consistency loss and MAE loss can help the model to better converge and improve the
359 counting performance.

360 *Effect of the number of MPN*

361 In order to verify the influence of the number of MPNs on the results, the number of MPNs is
362 gradually increased and dense connections are used in different structures. The results are shown
363 in Table 6. When the number of N is not greater than 3, the result of crowd counting is better as
364 the number of MPN increases. When N=3, MAE and RMSE are 63.7 and 98.3, respectively.
365 When N=4, the results were 64.4 and 97.7, with no significant improvement. In DMPNet, we use
366 dense connection, so there is no need to set too many MPN numbers, which will cause the
367 increase of parameters and the redundancy of calculation.

368 CONCLUSION

369 In this paper, we propose a novel end-to-end model called DMPNet for accurate crowd counting
370 and high-quality density map generation. The front-end network of DMPNet is VGG16, and the
371 back-end network is stacked by three densely connected MPNs. As an important component
372 module of DMPNet, MPN can effectively extract multi-scale features while keeping the input
373 and output resolution unchanged. The ability of the network is further enhanced by densely
374 connecting multiple MPNs. In addition, we combine Euclidean loss with density level
375 consistency loss and MAE loss to further improve the effect of the model. Experimental results
376 on three challenging datasets validate the adaptability and robustness of our method in different
377 crowd scenes. Although we deal with scale variation well, we do not eliminate background noise
378 in the crowd density map, which will affect the counting accuracy to some extent. In future
379 work, we will introduce attention mechanism to deal with background noise.

380 REFERENCES

381 **Boominathan L, Kruthiventi S, Babu R V. 2016.** CrowdNet: A deep convolutional network
382   for dense crowd counting. *In Proceedings of the 24th ACM international conference on*
383   *Multimedia.* 640-644. DOI 10.1145/2964284.2967300.

384 **Cao X, Wang Z, Zhao Y, et al. 2018.** Scale aggregation network for accurate and efficient
385   crowd counting. *In Proceedings of the European Conference on Computer Vision*
386   *(ECCV).* 734-750. DOI 10.1007/978-3-030-01228-1_45.

387 **Chan A B, Liang Z S, Vasconcelos N. 2008.** Privacy preserving crowd monitoring: Counting
388   people without people models or tracking. *2008 IEEE Conference on Computer Vision*
389   *and Pattern Recognition (CVPR).* 1-7. DOI 10.1109/CVPR.2008.4587569.

390 **Xiong H, Lu H, Liu C, et al. 2019.** From open set to closed set: Counting objects by spatial
391   divide-and-conquer. *2019 IEEE/CVF International Conference on Computer Vision*
392   *(ICCV). 8362–8371.* DOI 10.1109/ICCV.2019.00845.

393 **Hu Y, Jiang X, Liu X, et al. 2020.** NAS-Count: Counting-by-Density with Neural Architecture
394 Search. *In Proceedings of the European Conference on Computer Vision (ECCV)*. 747-
395 766. DOI 10.1007/978-3-030-58542-6_45.

396 **Xie W, Noble J A, Zisserman A. 2018.** Microscopy cell counting and detection with fully
397 convolutional regression networks. *Computer methods in biomechanics and biomedical*
398 *engineering: Imaging & Visualization.* 6(3):283-292. DOI
399 10.1080/21681163.2016.1149104.

400 **Sam D B, Surya S, Babu R V.2016.** Switching convolutional neural network for crowd
401 counting. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition*
402 *(CVPR)*. 4031-4039. DOI 10.1109/CVPR.2017.429.

403 **Zhang Y, Zhou D, Chen S, et al. 2016.** Single-image crowd counting via multi-column
404 convolutional neural network. *In Proceedings of the IEEE conference on computer vision*
405 *and pattern recognition (CVPR)*. 589-597. DOI 10.1109/CVPR.2016.70.

406 **Jiang, Xiaolong, et al. 2019.** Crowd counting and density estimation by trellis encoder-decoder
407 networks. *In Proceedings of the IEEE conference on computer vision and pattern*
408 *recognition (CVPR)*. 2019. 6133-6142. DOI: 10.1109/CVPR.2019.00629.

409 **Thanasutives, Pongpisit, et al. 2021**. Encoder-Decoder Based Convolutional Neural Networks
410 with Multi-Scale-Aware Modules for Crowd Counting. *25th International Conference on*
411 *Pattern Recognition (ICPR)*. 2382-2389. DOI:10.1109/ICPR48806.2021.9413286.

412 **Chen L C, Papandreou G, Schroff F, et al. 2017.** Rethinking atrous convolution for semantic
413 image segmentation. arXiv preprint arXiv:1706.05587, 2017. 2, 3, 7.

414 **Liu W, Salzmann M, Fua P. 2019.** Context-aware crowd counting. *In Proceedings of the IEEE*
415 *Conference on Computer Vision and Pattern Recognition (CVPR)*. 5099-5108.
416 DOI 10.1109/CVPR.2019.00524.

417 **Zhu L, Zhao Z, Lu C, et al. 2019.** Dual path multi-scale fusion networks with attention for
418 crowd counting. arXiv preprint arXiv:1902.01115.

419 **X Jiang, Zhang L, Xu M, et al. 2020.** Attention Scaling for Crowd Counting. *2020 IEEE/CVF*
420 *Conference on Computer Vision and Pattern Recognition (CVPR)*. 4705-4714. DOI
421 10.1109/CVPR42600.2020.00476.

422 **Li Y, Zhang X, Chen D. 2018.** CSRNet: Dilated convolutional neural networks for
423 understanding the highly congested scenes. *In Proceedings of the IEEE conference on*
424 *computer vision and pattern recognition (CVPR)*.1091-1100. DOI 10.1109 /CVPR.
425 2018.00120.

426 **Shi M, Yang Z, Xu C, et al. 2018.** Perspective-Aware CNN For Crowd Counting.
427 CoRR abs/1807.01989 (2018) .

428 **Jiang X, Xiao Z, Zhang B, et al.** Crowd Counting and Density Estimation by Trellis Encoder-
429   Decoder Networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern*
430   *Recognition (CVPR) IEEE, 2020.* DOI: 10.1109/CVPR.2019.00629.

431 **Guo D, Li K, Zha Z J, et al. 2019.** DADNet: Dilated-Attention-Deformable ConvNet for
432   Crowd Counting. *In Proceedings of the 27th ACM International Conference on*
433   *Multimedia.* 1823-1832. DOI 10.1145/3343031.3350881.

434 **Jia Wan, Antoni B. Chan. 2019.** Adaptive Density Map Generation for Crowd Counting. *2019*
435   *IEEE/CVF International Conference on Computer Vision (ICCV)*.1130-1139. DOI
436   10.1109/ICCV.2019.00122.

437 **Liu L, Qiu Z, Li G, et al. 2020.** Crowd counting with deep structured scale integration network.
438   *2019 IEEE/CVF International Confere nce on Computer Vision (ICCV).* 1774-1783.
439   DOI 10.1109/ICCV.2019.00186.

440 **Zhang A, Shen J, Xiao Z, et al. 2020.** Relational Attention Network for Crowd Counting. *2019*
441   *IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2020.* DOI
442   10.1109/ICCV.2019.00689.

443 **Simonyan K, Zisserman A. 2014.** Very Deep Convolutional Networks for Large-Scale Image
444   Recognition. *Computer Science.* arXiv preprint arXiv:1409.1556.

445 **Idrees H, Tayyab M, Athrey K, et al. 2018.** Composition loss for counting, density map
446   estimation and localization in dense crowds. *In Proceedings of the European Conference*
447   *on Computer Vision (ECCV).* 532–546. DOI 10.1007/978-3-030-01216-8_33.

448 **Idrees H, Saleemi I, Seibert C, et al. 2013.** Multi-source multi-scale counting in extremely
449   dense crowd images. *In Proceedings of the IEEE Conference on Computer Vision and*
450   *Pattern Recognition (CVPR).* 2547–2554. DOI 10.1109/CVPR.2013.329.

451 **Sindagi V A, Patel V M. 2017.** A Survey of Recent Advances in CNN-based Single Image
452   Crowd Counting and Density Estimation. *Pattern Recognition Letters.* 107:3-16. DOI
453   10.1016/j.patrec.2017.07.007.

454 **Gao G, Gao J, Liu Q, et al. 2020.** CNN-based Density Estimation and Crowd Counting: A
455   Survey. *CoRR abs/2003.12783.*

456 **Enzweiler M, Gavrila D M. Gavrila. 2009.** Monocular pedestrian detection: Survey and
457   experiments. *IEEE transactions on pattern analysis and machine intelligence.*
458   31(12): 2179-2195. DOI 10.1109/TPAMI.2008.260.

459 **Felzenszwalb, Pedro, F, et al. 2010.** Object detection with discriminatively trained part-based
460   models. *IEEE Transactions on Pattern Analysis & Machine Intelligence.* 32(9):1627-
461   1645. DOI 10.1109/TPAMI.2009.167.

462 **Wu B, Nevatia R. 2007.** Detection and Tracking of Multiple, Partially Occluded Humans by
463   Bayesian Combination of Edgelet based Part Detectors. *International Journal of*
464   *Computer Vision.* 75(2):247-266. DOI 10.1007/s11263-006-0027-7.

465 **Chan A B, Vasconcelos N. 2009.** Bayesian poisson regression for crowd counting. *In 2009*
466   *IEEE 12th international conference on computer vision.* 545-551.
467   DOI 10.1109/ICCV.2009.54 59191.

468 **Ryan D, Denman S, Fookes C B, et al. 2009.** Crowd counting using multiple local features. *In*
469   *2009 Digital Image Computing: Techniques and Applications.* IEEE, 81-88. DOI
470   10.1109/DICTA.2009.22.

471 **Ke C, Chen C L, Gong S, et al. 2012.** Feature mining for localised crowd counting. *British*
472   *Machine Vision Conference(BMVC).* 1-1. DOI 10.5244/C.26.21.

473 **Victor S. Lempitsky, Andrew Zisserman. 2010.** Learning to count objects in images. *24th*
474   *Annual Conference on Neural Information Processing Systems.* 1324–1332. DOI
475   10.1.1.231.2318.

476 **Pham V Q, Kozakaya T, Yamaguchi O, et al. 2015.** Count forest: Co-voting uncertain number
477   of targets using random forest for crowd density estimation. *2015 IEEE International*
478   *Conference on Computer Vision (ICCV).* 3253–3261. DOI 10.1109/ICCV.2015.372.

479 **Viresh R J, Le H, Hoai M. 2018.** Iterative Crowd Counting. *15th European Conference on*
480   *Computer Vision.* 278-293. DOI 10.1007/978-3-030-01234-2_17.

481 **Dai F, Liu H, Ma Y, et al. 2021.** Dense Scale Network for Crowd Counting. *International*
482   *Conference on Multimedia Retrieval.* 64-72. DOI 10.1145/3460426.3463628.

483 **Tan M, Le Q V. 2019.** MixConv: Mixed Depthwise Convolutional Kernels. *30th British*
484   *Machine Vision Conference. 74.* arXiv preprint arXiv:1907.09595

485 **Huang G, Liu Z, Laurens V, et al. 2016.** Densely Connected Convolutional Networks. *2017*
486   *IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*
487   DOI: 10.1109/CVPR.2017.243.

488 **Shi M, Yang Z, Xu C, et al. 2019.** Revisiting perspective information for efficient crowd
489   counting. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*
490   *Recognition.* DOI: 10.1109/CVPR.2019.00745.

491 **Miao Y, Lin Z, Ding G, et al. 2020.** Shallow Feature Based Dense Attention Network for
492   Crowd Counting. *Proceedings of the AAAI Conference on Artificial Intelligence.*
493   34(7):11765-11772. DOI 10.1609/aaai.v34i07.6848

494   **Amaranageswarao G, Deivalakshmi S, Ko S B. 2020.** Deep Dilated and Densely Connected
495   Parallel Convolutional Groups for Compression Artifacts Reduction. *Digital Signal*
496   *Processing.* 106:102804. DOI 10.1016/j.dsp.2020.102804

497   **Lin T Y, Dollar P, Girshick R, et al. 2017.** Feature Pyramid Networks for Object Detection.
498   *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE
499   Computer Society. DOI 10.1109/CVPR.2017.106.

500   **Duta I C, Liu L, Zhu F, et al. 2020.** Pyramidal Convolution: Rethinking Convolutional Neural
501   Networks for Visual Recognition. *CoRR abs/2006.11538.*

502   **Wang X, Zhang S, Yu Z, et al. 2020.** Scale-Equalizing Pyramid Convolution for Object
503   Detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*
504   *(CVPR).* 13356-13365. DOI 10.1109/CVPR42600.2020.01337.

505   **Richardson E, Y Azar, Avioz O, et al. 2020.** It's All About The Scale-Efficient Text Detection
506   Using Adaptive Scaling. *2020 IEEE Winter Conference on Applications of Computer*
507   *Vision (WACV).* 1833-1842. DOI 10.1109/WACV45572.2020.9093534.

508   **Kingma D, Ba J. 2014.** Adam: A Method for Stochastic Optimization. Computer Science, 2014.
509   *arXiv preprint arXiv 1412.6980.*

510   **Ma Z, Wei X, Hong X, et al. 2019.** Bayesian loss for crowd count estimation with point
511   supervision. *In Proceedings of the IEEE International Conference on Computer Vision*
512   *(ICCV).* 6142–6151. DOI 10.1109/ICCV.2019.00624.

513   **Sindagi V A, Patel V M. 2017.** Generating High-Quality Crowd Density Maps Using
514   Contextual Pyramid CNNs. *2017 IEEE International Conference on Computer Vision*
515   *(ICCV).* 1879-1888. DOI 10.1109/ICCV.2017.206.

Table 1. The setups for different datasets.

Parameter settings for density maps generated from different datasets.

Datasets	Parameter Settings
ShanghaiTech Part_A (Zhang et al., 2016)	$\sigma_i=4$
ShanghaiTech Part_B (Zhang et al., 2016)	$\sigma_i=15$
UCF_QRNF (Idrees et al., 2018)	Geometry-adaptive kernels
UCF_CC_50 (Idrees et al., 2013)	Geometry-adaptive kernels

Table 2. Comparisons of our DMPNet with SOTA methods.

The empirical comparison of three mainstream datasets shows that our method is more effective on MAE and MSE. We have bolded the best two results from each dataset.

Methods	ShanghaiTech Part A		ShanghaiTech Part B		UCF_QNRF		UCF_CC_50	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MCNN	110.2	173.2	26.4	41.3	277.0	426.0	377.6	509.1
Switch-CNN	90.4	135.0	21.6	33.4	228.0	445.0	318.1	439.2
CP-CNN	73.6	106.4	20.1	30.1	-	-	295.8	320.9
ic-CNN	68.5	116.2	10.7	16.0	-	-	260.9	365.5
CSRNet	68.2	115.0	10.6	16.0	-	-	266.1	397.5
SANet	67.0	104.5	8.4	13.6	-	-	258.4	334.9
BL	62.8	101.8	7.7	12.7	**88.7**	**154.8**	229.3	308.2
RANet	**59.4**	102.0	7.9	12.9	111	190	239.8	319.4
SDANet	63.6	101.8	7.8	**10.2**	-	-	227.6	316.4
SFANet	**59.8**	99.3	6.9	10.9	100.8	174.5	219.6	316.2
PACNN	66.3	106.4	8.9	13.5	-	-	241.7	320.7
TEDNet	64.2	109.1	8.2	12.8	113	188	249.4	354.5
DSNet	61.7	102.6	**6.7**	**10.5**	91.4	160.4	**183.3**	**240.6**
M-SFANet	59.69	**95.66**	**6.76**	11.89	**85.60**	**151.23**	**162.33**	**276.76**
DMPNet	63.7	**98.3**	7.6	11.8	98.7	179.8	202.4	301.5

Table 3. The estimation errors of LPN and GPN are compared on ShanghaiTech Part A (Zhang et al., 2016).

In the following training, MFFN is used.

Methods	MAE	RMSE
w/ LPN, w/o GPN	67.3	105.4
w/ GPN, w/o LPN	70.4	113.5
w/ (LPN+GPN)	63.7	98.3

The estimation errors of dense connections are compared on ShanghaiTech Part A (Zhang et al., 2016).

In the following training, we used three MPNs.

Method	MAE	RMSE
w/o Dense connection	68.4	108.7
w/ Dense connection	63.7	98.3

The estimation errors of different loss function combinations are compared on ShanghaiTech Part A (Zhang et al., 2016).

Method	MAE	RMSE
L_E	70.0	108.4
$L_E + L_D$	67.3	105.8
$L_E + L_A$	69.6	107.6
$L_E + L_D + L_A$	63.7	98.3

The estimation errors of different MPN numbers are compared on ShanghaiTech Part A (Zhang et al., 2016).

MPN(n) represents that the network contains n MPNs.

Method	MAE	RMSE
MPN($\mathbf{1}$)	71.0	111.3
$\boldsymbol{MPN}(\mathbf{2})$	66.2	103.4
$\boldsymbol{MPN}(\mathbf{3})$	63.7	98.3
$\boldsymbol{MPN}(\mathbf{4})$	64.4	97.7

Figure 1

Figure 1. Different scales of heads exist in crowd counting datasets.

The first row shows samples of crowd images, The second row shows corresponding ground truth density maps.

Figure 2

Figure 2. The architecture of DMPNet for crowd counting and high-quality density map.

It contains VGG16 (Simonyan & Zisserman, 2014) as the front-end network and three MPNs stacked by dense connections as the back-end network. MPN is composed of LPN, GPN and MFFN. It is used to extract human head features at different scales, and the resolution and channel number of input feature maps remain unchanged.
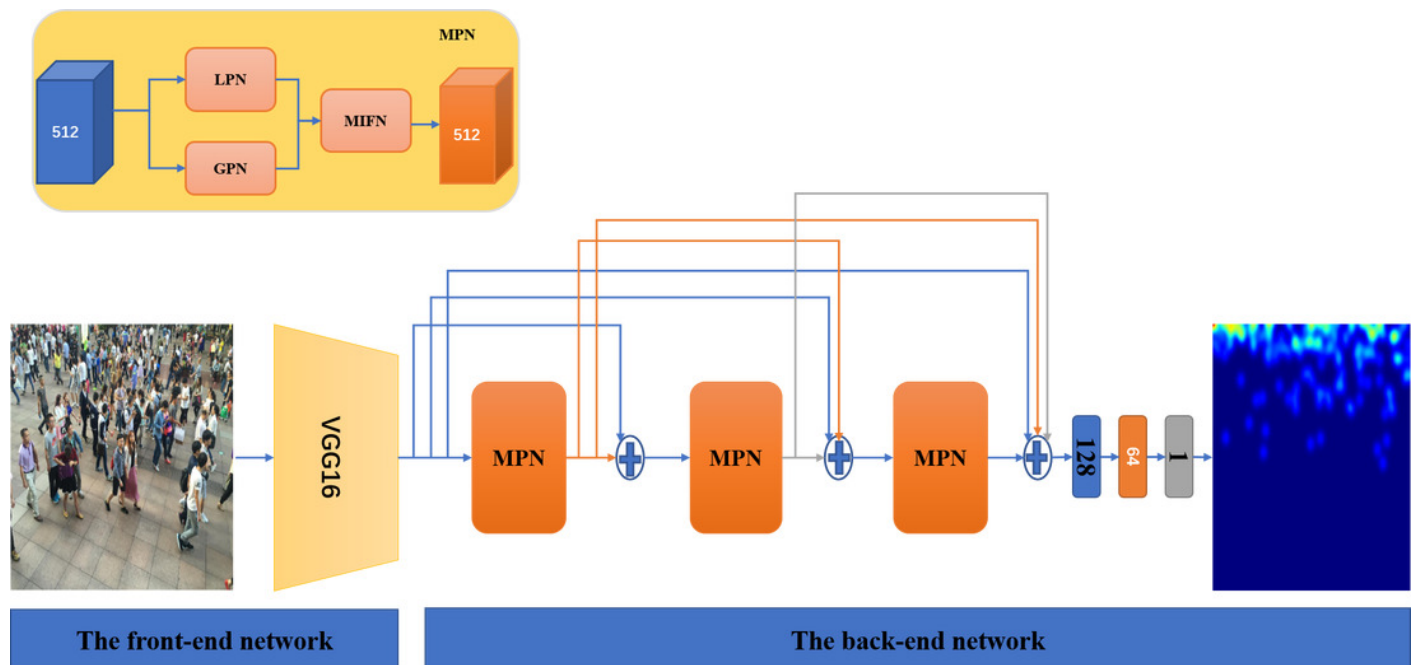
Figure 3

Figure 3. Compare the calculation process of Standard Convolution and Pyramid Convolution.

In pyramid convolution, the input feature map is calculated with convolution kernels of different sizes, and then the obtained feature map is connected by channel as the output feature map. The size of convolution kernel is decreasing, and the depth of convolution kernel is increasing.

(a) Standard Convolution   (b) Pyramidal Convolution
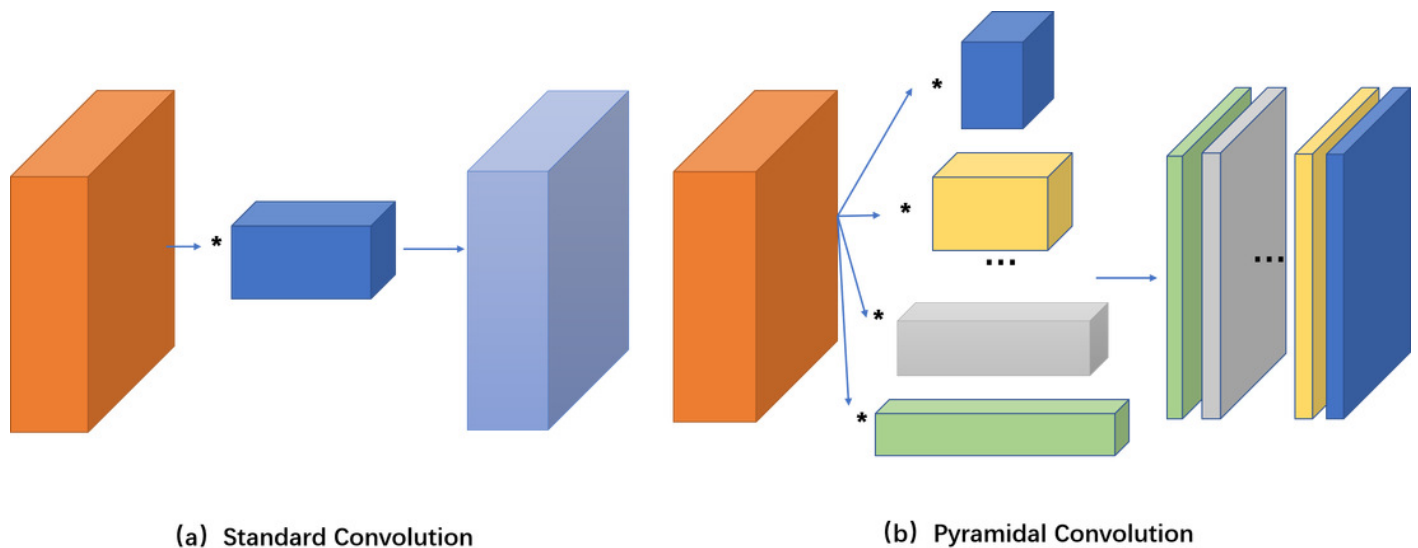
Figure 4

Figure 4. Compare the calculation process of Standard Convolution and Group Convolution.

In grouping convolution, the input feature map is divided into N groups, and the convolution kernel is also divided into N groups accordingly. The calculation is carried out in the corresponding group. Each group will generate a feature map, and a total of N feature maps are generated.
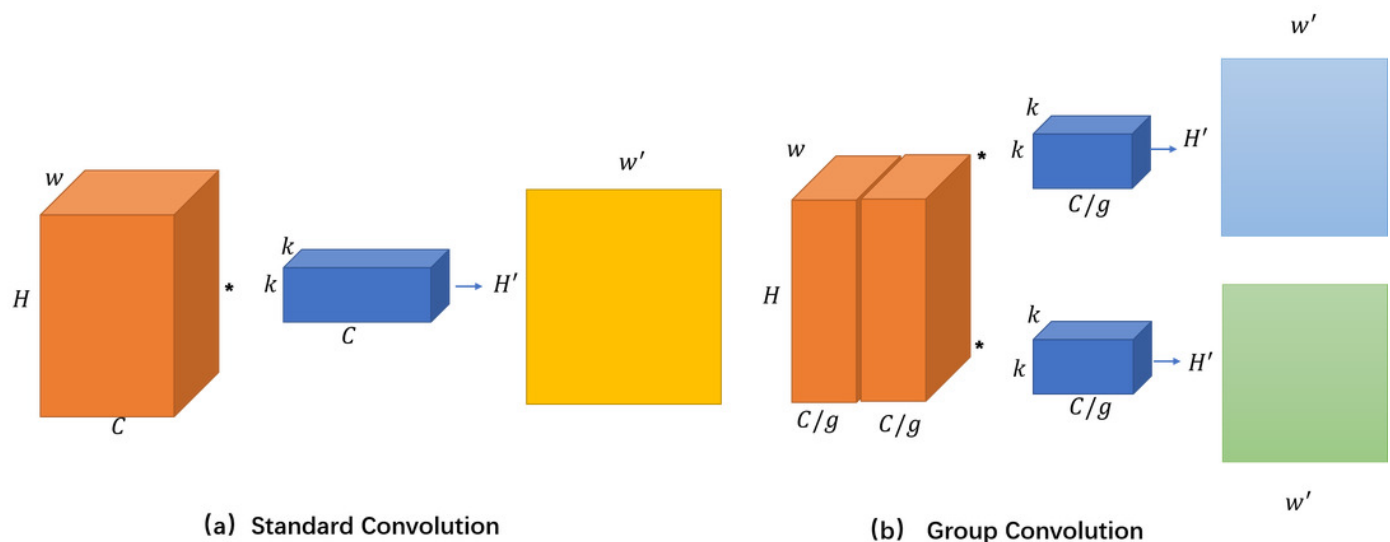
Figure 5

Figure 5. Three main components of Multi-scale Pyramid Network.

F_I is the input features of LPN and GPN. F_L and F_G are output features of LPN and GPN, respectively. F_O is output features of MFFN.
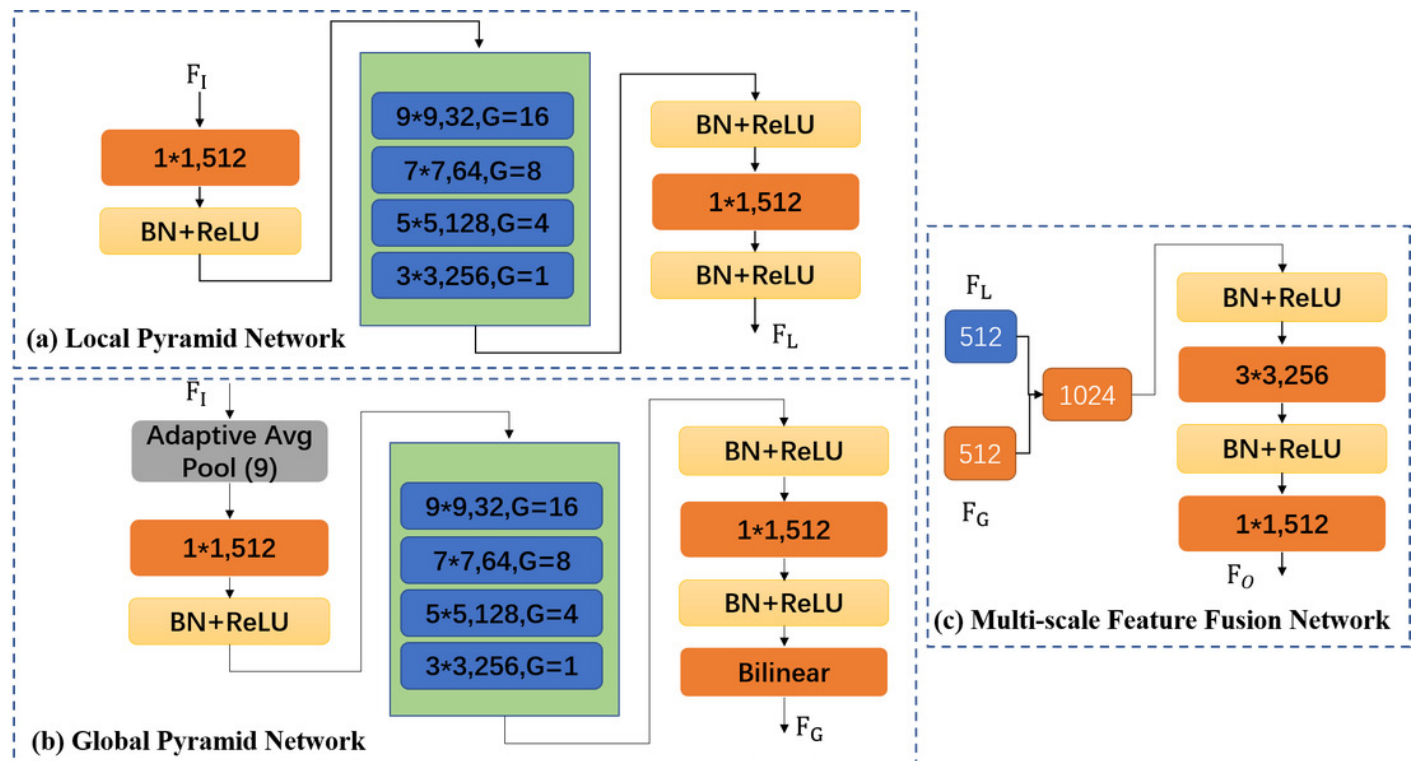
Figure 6

Figure 6. The visualization results and the corresponding counting results of our DMPNet.

The first row illustrates different test images from left to right: ShanghaiTech Part A (Zhang et al., 2016), ShanghaiTech Part B (Zhang et al., 2016), UCF-QNRF (Idrees et al., 2018), and UCF_CC_50 (Idrees et al., 2013). The second and third lines are the ground truth map and the estimated density map generated by DMPNet, respectively.
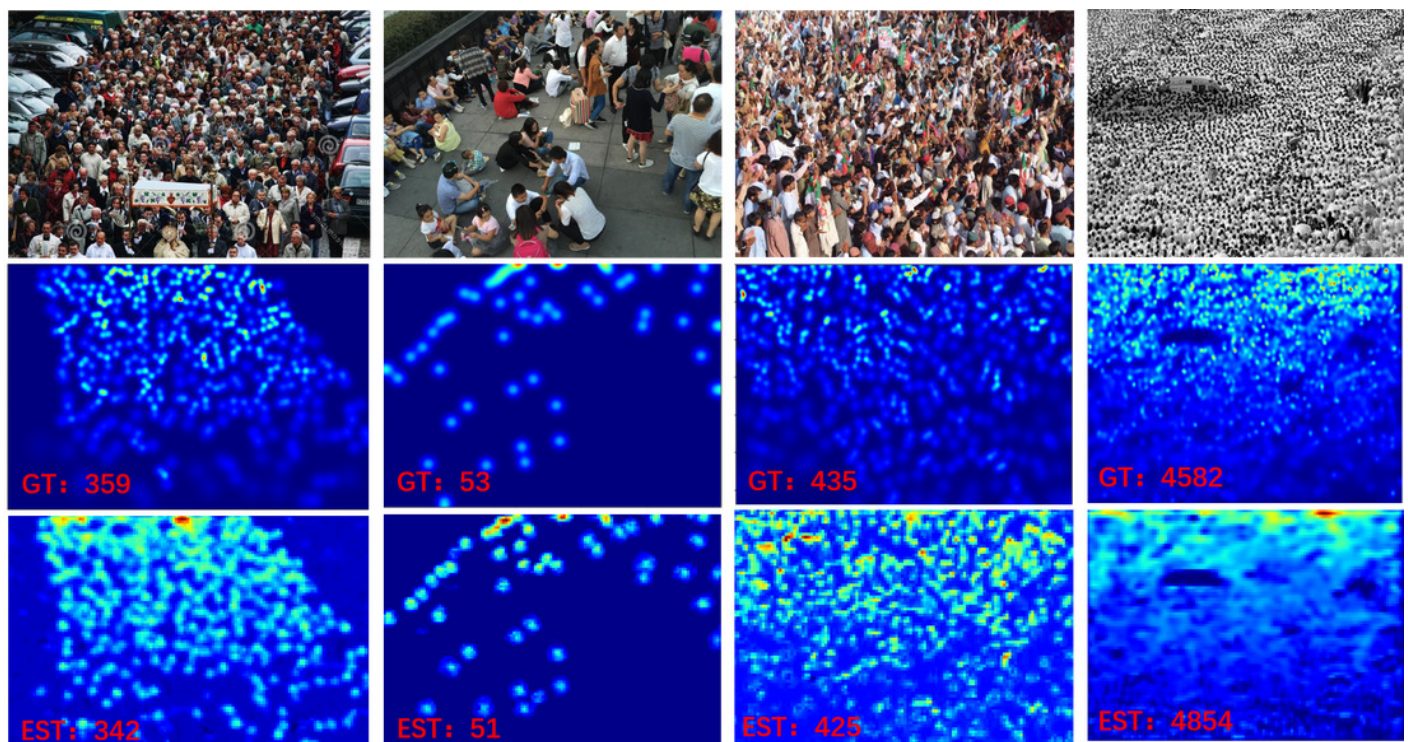
Figure 7

Figure 7. Comparison of density maps generated by different SOTA methods on ShanghaiTech Part A and Part B dataset (Zhang et al., 2016).

The six rows show that: (1) The test images, (2) the ground truth, (3) Density maps produced by MCNN (Zhang et al., 2016), (4) Density maps produced by CSRNet (Li, Zhang & Chen, 2018), (5) Density maps produced by DSNet (Dai et al., 2020), (6) Density maps produced by our DMPNet.