Multimedia Tools and Applications (2024) 83:26687–26709
https://doi.org/10.1007/s11042-023-16442-2

Dual convolutional neural network for crowd counting

Huaping Guo[1,2] · **Rui Wang**[3] · **Li Zhang**[1,4] · **Yange Sun**[1,2]

Received: 4 December 2022 / Revised: 27 July 2023 / Accepted: 2 August 2023 /
Published online: 4 September 2023

Abstract
As a challenging issue in computer vision, crowd counting has been increasingly studied. A convolutional neural network (CNN) is an effective system for handling crowd counting, based on constructing a CNN to generate a high-quality density estimation map. However, conventional CNN-based methods only consider the mapping from the crowd image to the density map, neglecting reconstruction from the density map to the crowd image and the impact of this reconstruction on the CNN performance. Here, we present a novel model denoted a dual-CNN (DualCNN) to improve the conventional CNN performance on crowd counting. Our DualCNN comprises a primal network for generating the density maps from the crowd image and a secondary network for reconstructing the crowd image from the density map. The two networks are trained through an iterative and alternating learning process, and the performance of the final model is improved by considering the interactions of the two networks. In addition, we introduce the attention mechanism into the dual network to enhance the primal network robustness against the background influence of the crowd image. The experimental results indicate that the proposed method significantly improves the performance of CNNs in crowd counting.

Keywords Crowd counting · Dual network · Convolutional neural network

1 Introduction

Recently, rapid urbanization has resulted in rapid crowd gatherings, such as in squares, stations and cinemas. Unfortunately, such gatherings have led to many stampede incidents [16,

✉ Huaping Guo
hpguo@xynu.edu.cn

✉ Yange Sun
yangesun@xynu.edu.cn

1 School of Computer and Information Technology, Xinyang Normal University, Xinyang 464000, Henan, China

2 Research Center of Precision Sensing and Control, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

3 School of Big Data and Artificial Intelligence, Xinyang University, Xinyang 464000, Henan, China

4 School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou 450001, Henan, China

43]. For example, a crowd stampede occurred in Shanghai, China, on New Year Eve in 2015 and resulted in 36 deaths and 49 injuries. Accurately counting the number of people in crowd images or videos offers great significance for crowd control and public safety [15, 67]. Techniques for crowd counting extend to related fields, including cell microscopy [17], animal monitoring [12] and environment surveys [69]. In addition, two crowd distribution (density) maps with the same number of people may greatly differ from each other as shown in Fig. 1, where the crowd in the left image is mainly distributed at the top of the crowd scene, while the crowd in the right image is mainly distributed at the middle of the crowd scene. Furthermore, congestion may only occur in some patches of an image, and these crowded places are at greater risk for stampede incidents. As shown in Fig. 2, most people are congested on the side of the road; thus, alarm systems should pay more attention to the real-time situation on the roadside. Therefore, it is meaningful to determine the crowd distribution in a high-risk environment.

Many approaches exist for generating high-quality crowd density maps [28, 62], and convolutional neural network (CNN), which is widely used in human activity recognition [37], disease monitoring [26, 38] facial expression recognition [51, 57] and event Summarization [24, 25], is one of the most frequently used approaches [14, 15, 43, 54]. CNNs have been used as the mapping function to generate the density map from the crowd image and estimate the crowd count from the density map. For example, Li et al. [34] introduced dilated kernels to replace pooling operations to deliver larger reception fields and proposed a dilated convolution neural network for solving the crowd counting problem, and Jiang et al. [21] proposed a trellis encoder-decoder network (TEDnet) using multiple decoding paths to fuse

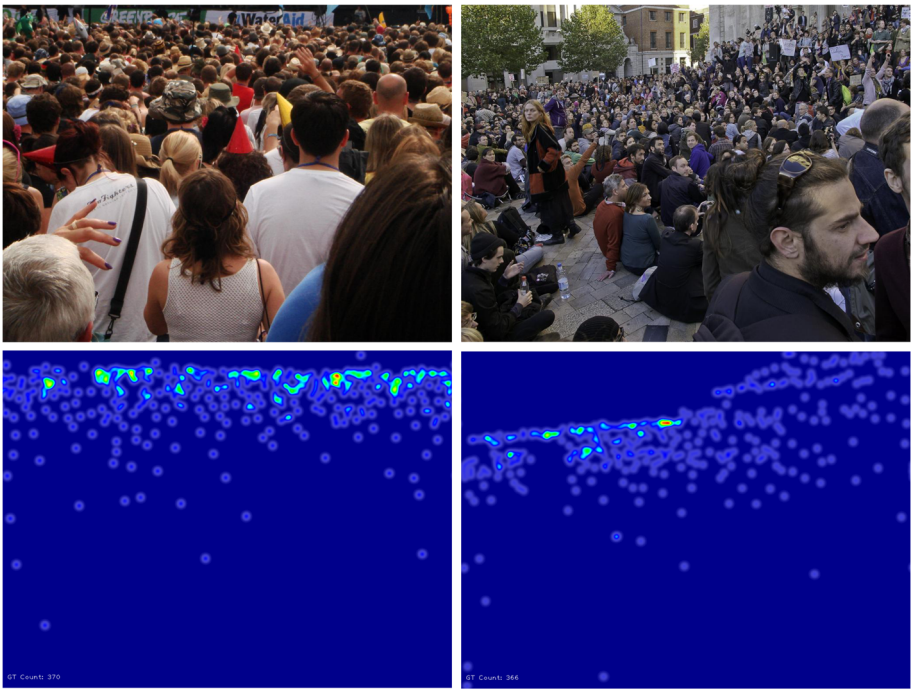

Fig. 1 The two images in the first row both contain 167 people in the ShanghaiTech Part_A dataset, while they have quite different spatial distributions. Images in the second row show their density maps

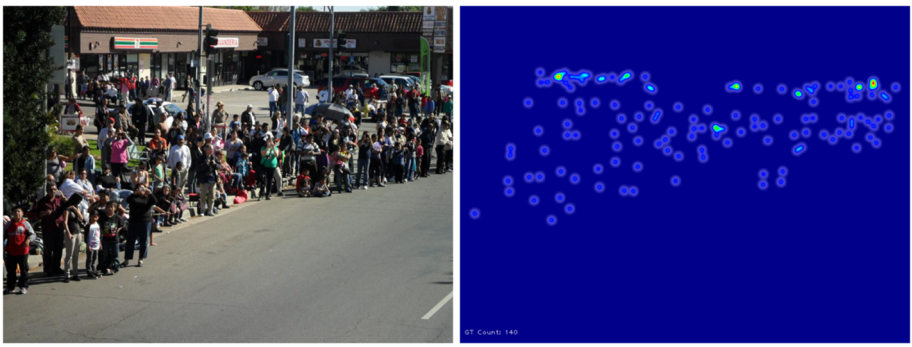

Fig. 2 People in this scene are mainly distributed along the road: (left) crowd image and (right) corresponding density map

multiscale features and exploiting dense skip connections to sufficiently facilitate multiscale feature fusion. Figure 3 provides the architecture of conventional convolutional neural networks for the crowd counting problem, where the ground truth density map is generated from the crowd image with people heads annotated. The limitation with crowd-counting oriented CNNs is that they typically only consider the mapping from the crowd image to the density map, neglecting the reconstruction from the density map to the crowd image and the impact of the reconstruction on the model performance.

To solve the above problem, we build a novel CNN-based framework that generates the crowd density map from a crowd image and reconstructs the crowd image from the crowd density map. The developed framework is intended to improve the performance of CNNs for the crowd counting problem by considering the impact of the reconstruction. According to the above consideration, a dual convolutional neural network (DualCNN) framework is proposed to enhance model performance on the crowd counting problem. This framework is inspired by the natural language processing (NLP) dual learning method [18], which uses one agent to represent the model for the primal task and the other agent to represent the model for the dual task, and then, it asks these agents to teach each other through a loop reinforcement learning process.

Similar to the NLP dual learning method, the proposed DualCNN comprises two agents (i.e., convolutional neural networks): the primal network executes the primal task, i.e., generating the density or distribution map from the crowd image, and the dual network executes the dual task, i.e., reconstructing the crowd image from the density map. DualCNN mainly differs from the NLP dual learning method as follows: the NLP dual learning model achieves mutual translation between languages, and the two languages are equally important, whereas DualCNN mainly aims to learn the crowd counting model with high precision and focuses more on the performance of the primal network.

In addition, we observe that the background of the crowd image is an important factor for testing the robustness of the model since it is difficult to completely eliminate the interference of the background on the performance of the model. Many convolutional neural networks with attention mechanisms have been proposed for the crowd counting problem [31, 36, 59, 70, 78], where attention mechanism-based structures are designed for capturing/extracting the features of the region of interest. In this paper, we also introduce the attention mechanism into the proposed dual convolutional neural network framework to make the framework focus more on the foreground of the crowd image, thereby reducing the impact of the background

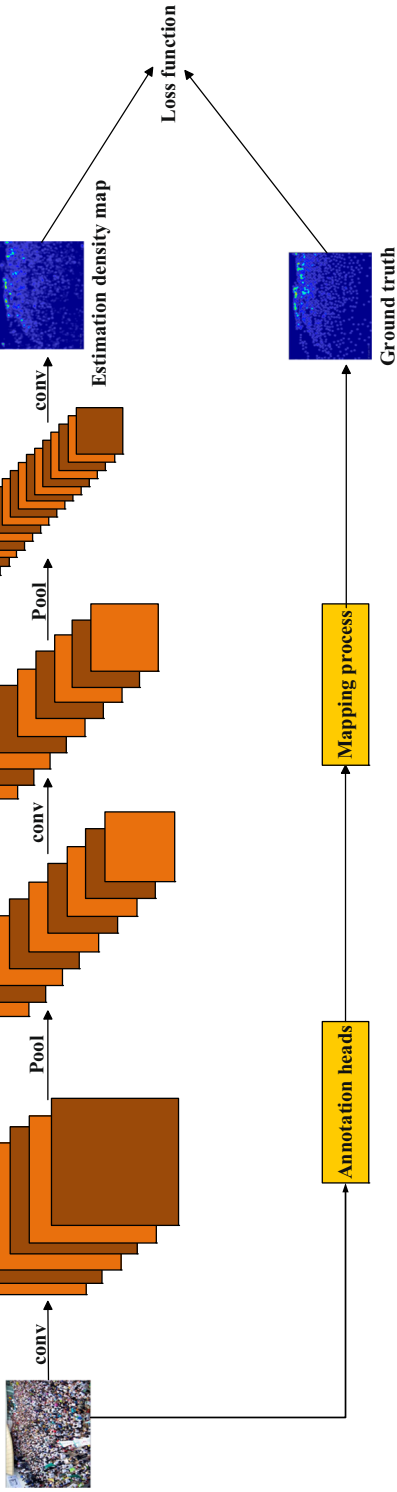

Fig. 3 Architecture of convolutional neural network for the crowd counting problem, where conv and pool in the figure represent the operations of convolution and pooling respectively. The true density map are generated based on the annotated heads in the crowd image

of the crowd image on the proposed dual framework. Therefore, the contributions of this paper are as follows:

During model construction, we observed that the background of the crowd image is an important factor for testing the robustness of the model since it is difficult to completely eliminate the interference of the background on the performance of the model. Many CNNs with attention mechanisms have been proposed for the crowd counting problem [31, 36, 59, 70, 78], where attention mechanism-based structures are designed for capturing/extracting the features of the region of interest. Here, we also introduce the attention mechanism into the proposed dual convolutional neural network framework to make the framework focus more on the foreground of the crowd image, thereby reducing the impact of the background of the crowd image on the proposed dual framework. To summarize, the contributions of this paper are as follows:

- We propose a novel dual learning framework comprising two convolutional neural networks: the primal network for generating the density maps from the crowd image and the dual network for reconstructing the crowd image from the density map. The framework iteratively learns the parameters of the primary network and the dual network and improves the performance of traditional CNNs by considering the mapping from the crowd image to the density map and the impact of the mapping on the primal network.
- We introduce the attention mechanism into the proposed dual learning framework to make the framework focus more on the foreground of the crowd image, thereby reducing the impact of the background of the crowd image on the framework; this further improves the performance of the traditional CNN (i.e., the primal network) on the crowd counting problem.
- We perform an ablation experiment to validate the effectiveness of the proposed framework and the attention mechanism. A comparison with other advanced methods shows that the proposed DualCNN achieves the lowest mean absolute error (MAE) and mean square error (MSE). Three traditional crowd counting oriented convolutional neural networks were selected as the primal networks, and the corresponding experimental results validate the robustness of DualCNN in improving the model performance on the crowd counting problem.

The remainder of this paper is organized as follows: after presenting related work in Section 2, Section 3 describes the proposed method; Section 4 presents the experimental results; and, finally, conclusions are provided in Section 5.

2 Related work

2.1 Crowd counting methods

The crowd counting problem has become increasingly important in computer vision technology. Many approaches [1, 55] have been proposed to handle this problem, and these efforts can be categorized into three levels, i.e., detection-based level, regression-based level and density estimation-based level.

Detection-based level methods commonly assume that every entity can be detected with a moving-window-like detector. Reported studies of detection-based approaches can be grouped into two types: detection based on the whole body [29] and detection based on the partial body [11]. For the former, the detector extracts low-level features from the whole human body, such as Haar wavelets [13] and HOGs (histogram oriented gradients) [47].

Then, the method identifies people using classifiers, such as random forest [29] and support vector machine [56]. Therefore, approaches based on whole-body detection are mainly suitable for sparse crowd scenes and performs poorly on highly congested scenes due to crowd occlusion. To address this problem, partial body-based detection methods are proposed to count the number of people by detecting particular body parts, including the head and arms.

A limitation existing in detection-based level methods is that they are incapable to handle crowd occlusions. To resolve this, regression-based level approaches [5, 33, 46, 48] learn the regression of crowd image features to the crowd number using the following three steps: 1) foreground extraction; 2) feature learning; and 3) counting regression. Novel feature extraction methods, including hybrid dynamic texture [2] and wavelet analysis [6], have been employed to extract crowd foreground features for further consideration. For the feature learning step, the detector identified effective features of the foreground features extracted in the first step, such as crowd edge information, texture and perimeter. With respect to the last step, the detector used regression methods, such as linear regression, ridge regression [65] and Gaussian regression [35], to learn the mapping from the identified effective features to the total number of people. Regression-based level methods alleviated the occlusion problem of detection-based level methods; however, these methods forfeited localization capability and thus could not perceive crowd distributions [21].

To solve the problem that regression-based approaches forfeit the ability to perceive crowd distributions, density estimation-based methods [45, 50, 53] are proposed to generate a density map that reflects the crowd distribution and then estimate the number of people from the density map. Many density estimation-based methods [4, 30, 49, 76, 77] have been proposed, and those based on convolutional neural networks (CNNs) have received the most widespread attention. For example, Zhang et al. [75] proposed a multicolumn convolutional neural network (MCNN) model that provides flexible receptive fields by utilizing receptive filters with different sizes across the columns. Li et al. [34] observed that a deeper single-column network may outperform the MCNN and thus proposed a novel deep network called CSRNet for the crowd counting problem. CSRNet is composed of two components: a conventional CNN and a dilated CNN, where the dilated CNN uses dilated kernels to deliver larger reception fields and to replace pooling operations. Jiang et al. [21] proposed a trellis encoder-decoder network (TEDnet) that uses multiple decoding paths to fuse multiscale features of different encoding stages and employs dense skip connections interleaved across paths to sufficiently facilitate multiscale feature fusion. Liu et al. [32] proposed an end-to-end deep network that combines features obtained by utilizing filters with receptive fields of different sizes and learns the importance of each such feature at each image location. Vishwanath and Vishal [59] applied attention mechanisms to enhance the features of the convolutional neural network and proposed a hierarchical attention-based network (HA-CCA) to accurately estimate crowd density.

To resolve regression-based approaches forfeiting the ability to perceive crowd distributions, density estimation-based methods [45, 50, 53] were proposed to generate a density map that reflects the crowd distribution and then estimate the number of people from the density map. Many density estimation-based methods [4, 30, 49, 76, 77] have been proposed, and those based on CNNs have received the most widespread attention. For example, Zhang et al. [75] proposed a multicolumn CNN (MCNN) model that provided flexible receptive fields by utilizing receptive filters with different sizes across the columns. Li et al. [34] observed that a deeper single-column network may outperform the MCNN and thus proposed a novel deep network called CSRNet for the crowd counting problem. CSRNet comprised two components: a conventional CNN and a dilated CNN, where the dilated CNN uses dilated kernels to deliver larger reception fields and to replace pooling operations. Jiang et al. [21] proposed

a trellis encoder-decoder network (TEDnet) that uses multiple decoding paths to fuse multi-scale features of different encoding stages and employed dense skip connections interleaved across paths to sufficiently facilitate multiscale feature fusion. Liu et al. [32] proposed an end-to-end deep network that combines features obtained by utilizing filters with receptive fields of different sizes and learns the importance of each such feature at each image location. Vishwanath and Vishal [59] applied attention mechanisms to enhance the features of the convolutional neural network and proposed a hierarchical attention-based network (HA-CCA) to accurately estimate crowd density.

Here, we propose a novel dual framework based on a CNN that learns the network for generating the density map from the crowd image and the dual network for generating the crowd image from the density map. The model's performance on the crowd counting problem is intended to be improved due to favourable use of the interactions between the two networks.

2.2 Attention mechanism

The attention mechanism of machine vision forces the model to more heavily consider areas of interest and to ignore areas that contribute minimally to the final result. The attention mechanism has been extensively applied to image subtitles [9, 60], visual question and answering systems [3, 8, 64, 66], pose estimation [10], stock price prediction [27], classification [19, 40, 44, 63] and detection [39, 41, 42, 73]. For example, Xu et al. [64] proposed a novel spatial attention architecture that aligns words with image patches in the first hop and obtained improved results by adding a second attention hop that considers the whole question to choose visual evidence based on the results of the first hop. Chu et al. [10] incorporated CNNs with a multi-context attention mechanism into an end-to-end framework for human pose estimation. Wang et al. [61] proposed a residual attention network (RAN) using an attention mechanism that can be incorporated with a state-of-art feed forwards network architecture in an end-to-end training fashion to improve the image classification performance. Hu et al. [19] proposed compact squeeze-and-excitation (SE) block modelling interdependencies between channels to boost the representational power of a network.

In crowd counting research, efforts have further improved the model performances by introducing an attention mechanism. Vishwanath and Vishal [59] used the attention mechanism to enhance the characteristics of convolutional neural networks and proposed a hierarchical attention-based network (HA-CCA) to accurately estimate crowd density. Zhang et al. [74] proposed an attention neural field (ANF) that combined conditional random fields and nonlocal attention mechanisms to capture multiscale features and large-scale dependencies and enhance the network's ability to handle large-scale changes. Furthermore, Zhang et al. [72] proposed a relational attention network (RANet) that used both local and global self-attention mechanisms to capture the interdependence information between pixels and obtain more informative feature representations. Jiang et al. [22] proposed a scale attention network that combined the density attention network (DANet) and the attention scale network (ASNet). DANet provided ASNet with attention masks related to areas with different density levels. ASNet first generated the density map and scale factors and then multiplied them by the attention mask to output a separate attention-based density map. These density maps are added to obtain the final density map.

The attention mechanism adopted here is used in the process of density map reconstruction. Since the mapping from the density map to the crowd map is not unique, the dual network cannot restore a given density map to a different crowd map. Therefore, the atten-

tion mechanism is introduced in the mapping process of the density map to map the relevant density map to the corresponding crowd attention map.

2.3 Dual model

Because nearly monolingual data are available via the web, the dual-learning mechanism was developed by He et al. [18] to leverage these data and boost the performance of neural machine translation (NMT) systems. This mechanism is inspired by any machine translation task having a corresponding dual task, e.g., English-to-French translation (primal) versus French-to-English translation (dual). In the dual-learning mechanism, one agent represents the model for the primal task, and the other agent represents the model for the dual task. Then, they are trained to teach each other through a reinforcement learning process. Inspired by the dual-learning mechanism, Yi et al. [68] proposed a novel dual model with generative adversarial networks (DualGAN) that allows images from either domain to be translated and then reconstructed by image translators.

The main differences between the dual convolutional neural network model proposed in the paper and the above dual network are as follows: 1) The goals of machine translation and DualGAN are to achieve inter-language translation and picture-to-picture conversion, and the goal of DualCNN is to learn a high-precision crowd skill counting model, that is, to pay more attention to the performance of the main model, thereby focusing on the impact of the main model on reconstruction loss rather than the impact of the dual model; and 2) The learning method used by machine translation and DualGAN is unsupervised learning, and the DualCNN model is mainly used to realize the mapping of the crowd map (density map) to the density map (crowd map). The crowd map and the density map have a one-to-one correspondence, which belongs to supervised learning.

The main differences between the dual convolutional neural network model proposed here and the above dual network are as follows: 1) The goals of machine translation and DualGAN are to achieve interlanguage translation and picture-to-picture conversion, and the goal of DualCNN is to learn a high-precision crowd skill counting model, that is, to better consider the performance of the main model, thereby focusing on the impact of the main model on reconstruction loss rather than the impact of the dual model; and 2) The learning method used by machine translation and DualGAN is unsupervised learning, and the DualCNN model mainly realizes the mapping of the crowd map (density map) to the density map (crowd map). The crowd map and the density map exhibit a one-to-one correspondence, which represents a supervised learning system.

3 Proposed solution

Conventional crowd-counting oriented CNNs often focus on generating a high-quality density map and then estimating the number of people from the map. The proposed dual-learning convolutional neural network (DualCNN) framework comprises one network for the primal task to generate the density map from the crowd image (called the primal network) and another network for the dual task to generate the crowd image from the density map (called the dual network). Figure 4 shows the overall architecture and data flow of the DualCNN.

Let $\boldsymbol{x} \in \boldsymbol{X}$ represent a crowd image and $\boldsymbol{y} \in \boldsymbol{Y}$ represent the ground truth density map corresponding to $\boldsymbol{x}$. The main task of the DualCNN is to learn the primal convolutional neural network model $C_A : \boldsymbol{X} \mapsto \boldsymbol{Y}$, mapping the crowd image $\boldsymbol{x} \in \boldsymbol{X}$ to the corresponding ground

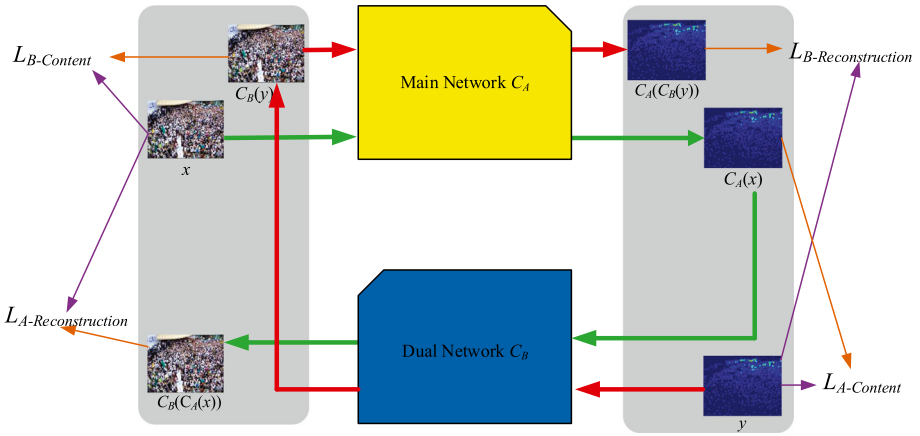

Fig. 4 Architecture and data flow of the dual-learning convolutional neural network

truth density map $\boldsymbol{y} \in \boldsymbol{Y}$, while the dual task is to train an inverse model $C_B : \boldsymbol{Y} \mapsto \boldsymbol{X}$, mapping $\boldsymbol{y} \in \boldsymbol{Y}$ to $\boldsymbol{x} \in \boldsymbol{X}$. In the training process, DualCNN considers the mutual effect between C_A and C_B to improve the performance of convolutional neural networks on crowd counting.

We observe that the density map generated from the crowd image is unique while multiple crowd images can be generated from a density map, i.e., there is a nonunique mapping from the density map to the crowd map. Therefore, it is difficult to recover the crowd image from a given density map, especially considering the restoration of the image background. In this section, we introduce the attention mechanism presented in Section 3.3 to reduce the influence of the crowd image background.

We observe that the density map generated from the crowd image is unique, and multiple crowd images can be generated from a given density map, i.e., there is a nonunique mapping from the density map to the crowd map. Therefore, it is difficult to recover the crowd image from a given density map, especially considering the restoration of the image background. In this section, we introduce the attention mechanism presented in Section 3.3 to reduce the influence of the crowd image background.

3.1 Primal network

In the dual-learning framework proposed here, a primary goal is to improve the mapping performance of the primal network (C_A model) from the crowd image to the density map. This mapping can be described as follows:

$$\boldsymbol{y} = C_A(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{y}$ is the density map predicted by network C_A and $\boldsymbol{x}$ represents a true crowd image or the crowd image obtained by dual network C_B (refer to Section 3.2). Therefore, the primal network is regarded as a dual model of the dual network, which maps the crowd image generated by the dual network C_B to a density map.

The proposed dual-learning framework primarily seeks to improve the performance of the primal network (C_A), which maps the crowd image to the density map. Our framework requires training both the primal and dual networks, which results in slow training speed.

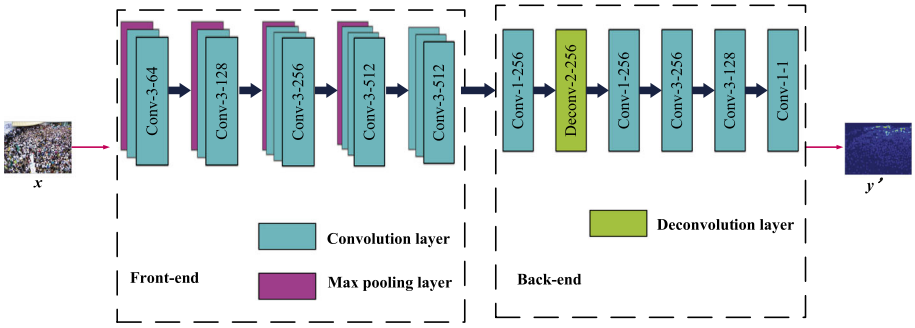

Fig. 5 Architecture and data flow of the primal convolutional neural network

Therefore, we design a simple and easy-to-train convolutional neural network. Figure 5 shows this architecture and data flow of the primal CNN, including two parts: a front and back end. Similar to CSRNet [34], VGG-16 [52] (the last downsampling layer and the fully connected layer are removed) is selected as the front end because of its strong transfer learning capability and its flexible architecture for easily concatenating the back end for density map generation. The corresponding parameters are initialized using VGG-16 parameters trained on ImageNet. The back end includes 6 convolutional layers, where the first and third layers are compression layers with a kernel size of 1×1, a stride equal to 1, and a channel number equal to 256. The second layer is a deconvolution layer with a deconvolution kernel size of 2×2, a stride equal to 2 and a channel number equal to 256 to expand the size of the density map. The fourth layer is similar to the fifth layer, with a convolution kernel size of 3×3 and a stride equal to 1. The difference between the fourth and fifth layers is that their channel numbers are equal to 256 and 128, respectively. The last layer with a kernel size of 1×1 and channel number equal to 1 outputs the density map.

3.2 Dual network

The dual network (C_B) of the proposed dual-learning framework reconstructs the crowd map from the density map, which is described as:

$$\boldsymbol{x} = C_B(\boldsymbol{y}) \tag{2}$$

where $\boldsymbol{x}$ is the true density map or the result predicted by C_A and $\boldsymbol{y}$ represents a crowd density map. The fundamental concept of the dual network (C_B) is that the primal network proposed in Section 3.1 can map the crowd image $\boldsymbol{x}$ to the density map $\boldsymbol{y}$ with high precision using (1) if $\boldsymbol{x}$ can be reconstructed from the obtained density map $\boldsymbol{y}$ with high precision using (2).

As mentioned in Section 3.1, the proposed dual-learning framework needs to train both the primal network and the dual network, which leads to a slow learning process. In addition, the main purpose of crowd counting is to improve the performance of the primal network that maps the crowd image to the density map, and therefore, the generalization performance of the dual network (C_B) does not need to be too high. For the above consideration, a relatively simple convolutional neural network is designed as the dual network C_B, as shown Fig. 6. The dual network contains four convolutional modules, where each of the first two modules contains 3 convolutional layers, and each of the latter two modules contains 2 convolutional layers. The size of the kernel of all convolution layers is 3×3, and the stride equals 1. The

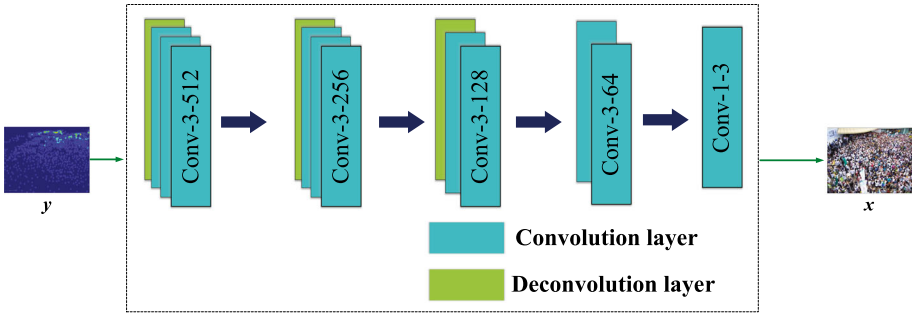

Fig. 6 Architecture and data flow of the dual convolutional neural network

number of convolution kernels of the four convolution modules is 512, 256, 128 and 64. Each convolutional module is followed by a deconvolutional layer with a kernel size of 2×2 and stride equal to 2 to enlarge the size of the feature map. Finally, a convolutional layer with kernel size of 1×1, 3-channel outputs and a stride equal to 1 is used to output the crowd image.

As mentioned in Section 3.1, this dual-learning framework needs to train both the primal network and the dual network, which results in slow learning. In addition, the main purpose of crowd counting research is to improve the performance of the primal network that maps the crowd image to the density map. Therefore, the generalization performance of the dual network (C_B) does not need to be excellent. Based on the above considerations, a relatively simple CNN is designed as the dual network C_B, as shown Fig. 6. The dual network contains four convolutional modules, where each of the first two modules contains 3 convolutional layers, and each of the latter two modules contains 2 convolutional layers. The size of the kernel of all convolution layers is 3×3, and the stride equals 1. The numbers of convolution kernels of the four convolution modules are 512, 256, 128 and 64. Each convolutional module is followed by a deconvolutional layer with a kernel size of 2×2 and stride equal to 2 to enlarge the size of the feature map. Finally, a convolutional layer with a kernel size of 1×1, 3-channel outputs and a stride equal to 1 outputs the crowd image.

3.3 Attention mechanism

Limits in the proposed dual-learning framework (Section 3.2) are that although there is a unique mapping from the crowd map to the density map, the mapping from the density map to the crowd map is not unique, especially considering crowd images with different backgrounds. For example, crowd images with the same population distribution but different backgrounds show that existing density map generation methods based on head annotation may generate exactly the same density map. However, the dual network proposed in Section 3.2 cannot restore a density map to different crowd images. Therefore, it is crucial to remove the impact of the image background on the performance of the dual network.

The space-based attention mechanism can effectively address this problem. We introduce a space attention mechanism [59] into the dual network to reduce the impact of the background on the model performance. The dual network with the attention mechanism is described as:

$$\boldsymbol{x} = \boldsymbol{M} \odot C_B(\boldsymbol{y}) \tag{3}$$

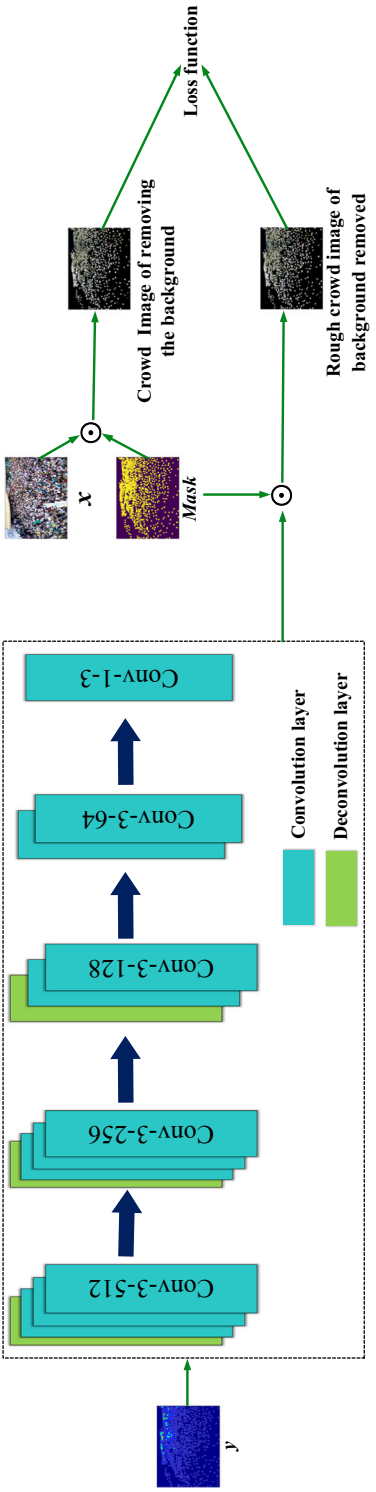

Fig. 7 Architecture and data flow of the dual network with attention mechanism

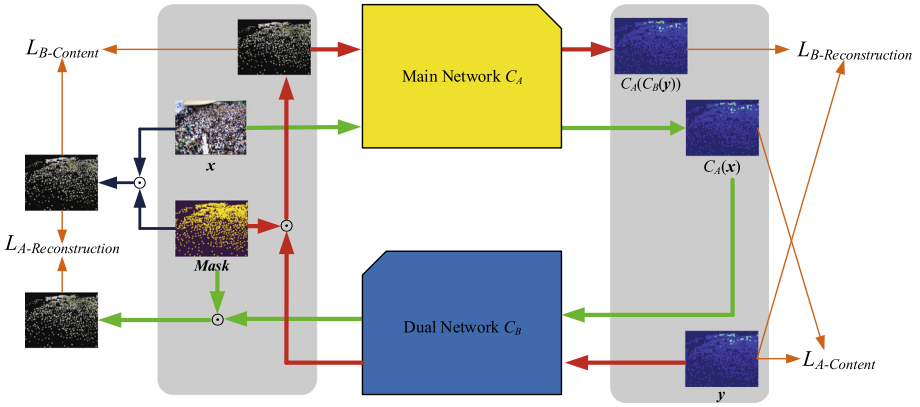

Fig. 8 Architecture and data flow of the dual model with an attention mechanism

where $\odot$ is the multiplication of elements, and $\boldsymbol{M}$ is the mask matrix. The element value of $\boldsymbol{M}$ equals 1 or 0, indicating the existence or absence of a head at the corresponding position. Then, the recovery of the crowd density map from the result of the dual network introduced in Section 3.2 is described as:

$$\boldsymbol{y}' = C_A\left(\boldsymbol{M} \odot C_B\left(\boldsymbol{y}\right)\right) \tag{4}$$

Figure 7 shows the dual network and data flow with the spatial attention mechanism, where $\boldsymbol{y}$ represents the truth density map or the map resulting from the primal network. Compared with the dual network introduced in Section 3.2, the dual network with the spatial attention mechanism differs from the network without the spatial attention mechanism in that the former uses the mask matrix $\boldsymbol{x}$ to remove the impact of the mask matrix. Figure 8 presents the dual-learning framework with the spatial attention mechanism.

3.4 Loss function

This section uses the notations in Section 3.1. Given the crowd image, density map and mask matrix, $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M})$, the proposed dual-learning framework learns C_A and C_B alternately, learns C_A by fixing C_B and then learns C_B by fixing C_A. The learning process repeats until convergence. The model C_A is affected by the content loss and reconstruction loss, as shown in Fig. 4; thus, the loss function used for C_A is defined as

$$L_A\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M}\right) = \alpha L_{A-content}\left(\boldsymbol{x}, \boldsymbol{y}\right) + \left(1-\alpha\right) L_{A-reconstruction}\left(\boldsymbol{x}, \boldsymbol{M}\right) \tag{5}$$

where $\alpha \in [0, 1]$ is a coefficient to adjust the relative importance of the content loss versus reconstruction loss and $L_{A-content}$ and $L_{A-reconstruction}$ are the content loss and reconstruction loss of $\boldsymbol{x}$, respectively. These parameters are defined as:

$$L_{A-content}\left(\boldsymbol{x}, \boldsymbol{y}\right) = \left\|\boldsymbol{y} - C_A\left(\boldsymbol{x}\right)\right\|_2^2 \tag{6}$$

$$L_{A-reconstruction}\left(\boldsymbol{x}, \boldsymbol{M}\right) = \left\|\left(\boldsymbol{x} - C_B\left(C_A\left(\boldsymbol{x}\right)\right)\right) \odot \boldsymbol{M}\right\|_2^2 \tag{7}$$

Let $\alpha = 1$, then

$$L_A\left(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M}\right) = L_{A-content}\left(\boldsymbol{x}, \boldsymbol{y}\right) = \left\|\boldsymbol{y} - C_A\left(\boldsymbol{x}\right)\right\|_2^2 \tag{8}$$

From (8), the proposed loss defined as (5) degenerates into the loss of traditional convolutional neural networks for crowd counting; therefore, it can be used for pretraining C_A.

Similar to the loss function used for C_A, the loss function for C_B is defined as

$$L_B(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M}) = \alpha L_{B-content}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M}) + (1-\alpha) L_{B-reconstruction}(\boldsymbol{y}, \boldsymbol{M}) \tag{9}$$

where $\alpha \in [0, 1]$ is a coefficient to adjust the relative importance of the content loss versus reconstruction loss and $L_{B-content}$ and $L_{B-reconstruction}$ are the content loss and reconstruction loss of $\boldsymbol{y}$, respectively. These parameters are defined as:

$$L_{B-content}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M}) = \|(\boldsymbol{x} - C_B(\boldsymbol{y})) \odot M\|_2^2 \tag{10}$$

$$L_{B-reconstruction}(\boldsymbol{y}, \boldsymbol{M}) = \|\boldsymbol{y} - C_A(C_B(\boldsymbol{y}) \odot \boldsymbol{M})\|_2^2 \tag{11}$$

Similar to (5), let $\alpha = 1$, then

$$L_B(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M}) = L_{B-content}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M}) = \|(\boldsymbol{x} - C_B(\boldsymbol{y})) \odot M\|_2^2 \tag{12}$$

Therefore, (12) can be used for pretraining C_B.

Algorithm 1 DualCNN.

Require: crowd images set $\boldsymbol{X}$, ground truth set $\boldsymbol{Y}$, and mask matrix $\boldsymbol{M}$
1: initialize the parameters of C_A, i.e., θ_A, using the parameters of VGG16 pretrained on ImageNet
2: initialize the parameters of C_B, i.e., θ_B, using Gaussian with 0.01 standard deviation
3: **repeat**
4: **for** $\boldsymbol{x} \in \boldsymbol{X}$ and the corresponding $\boldsymbol{y} \in \boldsymbol{Y}$ **do**
5: $d_{\theta_A} = \frac{\partial L_A(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M})}{\partial \theta_A}$ //refer to (5)
6: fixing θ_B, update θ_A using Adam optimizer [23] with gradient of d_{θ_A}
7: $d_{\theta_B} = \frac{\partial L_B(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M})}{\partial \theta_B}$ //refer to (9)
8: fixing θ_A, update θ_B using Adam optimizer [23] with gradient of d_{θ_B}
9: **end for**
10: shuffle the crowd images set $\boldsymbol{X}$
11: **until** convergence
12: **return** C_A

3.5 Training procedure

Our DualCNN iteratively learns the parameters of C_A and C_B, which are the subnetworks for generating the density estimation map and the crowd image, respectively, as defined in Section 3.1. For each iteration, DualCNN learns C_A and C_B alternately by fixing C_B and learning C_A and then fixing C_A and learning C_B. Algorithm 1 shows the details of the learning process of DualCNN.

The inputs of DualCNN are the crowd image set $\boldsymbol{X}$, the corresponding ground truth (density map) set $\boldsymbol{Y}$ and the mask matrix $\boldsymbol{x}$. DualCNN first initializes the parameters of C_A and C_B (Lines 1–2) and then iteratively learns model parameters until convergence (Lines 3-11). For each iteration, DualCNN uses all training data, i.e., $\boldsymbol{X}$ and $\boldsymbol{Y}$, to update the parameters of C_A and C_B (Lines 4–9). Given a pair of crowd images and density maps, $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{M})$, DualCNN learns C_A and C_B alternately: fixing C_B and updating C_A by minimizing (5) (Lines 5–6) and then fixing C_A and updating C_B by minimizing (9) (Lines 7–8).

Fig. 9 Representative crowd images of the ShanghaiTech dataset

4 Experiments

4.1 Datasets and experiment setup

We verify the effectiveness of the proposed method using two challenging datasets, ShanghaiTech [71] and UCF_CC_50 [20]. ShanghaiTech is a large-scale dataset released in 2016, and it includes Part_A and Part_B, each containing its own training and test sets. Part A contains 482 images with highly congested scenes randomly downloaded from the internet, while Part B includes 716 images with relatively sparse crowd scenes taken from streets in Shanghai. Figure 9 shows representative pictures of the ShanghaiTech dataset. The UCF_CC_50 dataset contains 50 images with different perspectives and resolutions, and the number of annotated persons per image ranges from 94 to 4543 (average of 1280). A 5-fold cross-validation is performed following [20]. Figure 10 shows representative images of the UCF_CC_50 dataset. The images of the two challenging datasets are directly used in this paper.

The front-end part of the primal network presented in Section 3.1 is initialized with the parameters of VGG16 pretrained on ImageNet, and the back-end part and dual network in Section 3.2 are initialized using Gaussian initialization with 0.01 standard deviation. In addition, Adam [23] is used as the optimizer with the initial learning rate at $1e$-5 and a 0.008 reduction every two iterations.

Fig. 10 Representative crowd images of the UCF_CC_50 dataset

4.2 Evaluation metrics

Evaluation metrics are essential for evaluating the effectiveness of algorithms. For the crowd counting problem, the mean absolute error (MAE) and mean square error (MSE) are two commonly used evaluation metrics, which are defined as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \|y_i^{gt} - y_i^{pred}\| \tag{13}$$

$$MSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(y_i^{gt} - y_i^{pred}\right)^2} \tag{14}$$

where N is the number of test crowd images and y_i^{gt} and y_i^{pred} represent the ground-truth result and predicted result for the i^{th} test crowd image, respectively. C_i represents the estimated count, which is defined as:

$$C_i = \sum_{l=1}^{L}\sum_{w=1}^{W} c_{l,w} \tag{15}$$

where L and W are the length and width of the density map, respectively, and $c_{l,w}$ is the pixel at (l, w) of the generated density map. Here, we use MAE and MSE to assess the performance of the proposed DualCNN.

4.3 Ablation experiment

Here, we perform an ablation study to analyse the four configurations of the DualCNN on the ShanghaiTech dataset [71]. Table 1 shows the results of the ablation experiment, where the primal network is presented in Section 3.1, NDualCNN represents the proposed dual convolutional neural network without an attention mechanism, and DualCNN represents the dual network with an attention mechanism, as shown in Fig. 8. From Table 1, NDualCNN reduces the MAE (MSE) of the Primal Network on Part_A and Part_B to 62.4 (97.4) and 7.9 (12.8), respectively. DualCNN further reduces the MAE (MSE) of NDualCNN on Part A and Part B to 59.7 (96.5) and 7.1 (11.1), respectively.

4.4 Experimental results

This section presents the experimental results of the DualCNN compared to Switch-CNN [50], CP-CNN [53], IG-CNN [49], CSRNet [34], MCNN [75], SCAR [9], SANet [4], MLCNN [22], CAT-CNN [7], PCC-Net [14], IA-DCCN [54], ic-CNN [58], SFCN [46] and

Table 1 Experimental results of the ablation experiment on ShanghaiTech

Mathod	Part_A		Part_B	
	MAE	MSE	MAE	MSE
Primal Network	65.8	103.9	9.1	15.3
NDualCNN	62.4	97.4	7.9	12.8
DualCNN	59.7	96.5	7.1	11.1

Table 2 Experimental results of the proposed method DualCNN

Method	ShanghaiTech				UCF_CC_50	
	Part_A		Part_B		MAE	MSE
	MAE	MSE	MAE	MSE		
Switch-CNN [50]	90.4	135.0	21.6	33.4	318.1	439.2
CP-CNN [53]	73.6	106.4	20.1	30.1	295.8	320.9
IG-CNN [49]	72.5	118.2	13.6	21.1	291.4	349.4
CSRNet [34]	68.2	115.0	10.6	16.0	266.1	397.5
MCNN [75]	110.2	173.2	26.4	41.3	377.6	509.1
SCAR [9]	66.3	114.1	9.5	15.2	259.0	374.0
ic-CNN [58]	68.5	116.2	10.7	16.0	260.9	365.5
IA-DCCN [54]	66.9	108.4	10.2	16.0	264.2	394.4
SANet [4]	67.0	104.5	8.4	13.6	258.4	334.9
MLCNN [22]	71.2	112.5	12.1	19.3	242.4	317.8
CAT-CNN [7]	66.7	101.7	11.2	20.0	235.5	324.8
PCC-Net [14]	73.5	124.0	11.0	19.0	240.0	315.5
SFCN [46]	64.8	107.5	7.6	13.0	214.2	318.0
COMAL [76]	59.6	97.1	7.8	12.4	231.9	333.7
DualCNN	59.7	96.5	7.1	11.1	213.8	313.9

COMAL [76] on the ShanghaiTech dataset and UCF_CC_50 dataset. Table 2 presents the comparable results from these 13 methods.

Table 2 shows that DualCNN achieves the lowest MAE and MSE on ShanghaiTech Part_B and UCF_CC_50 and achieves the lowest MSE on Part_A. Specifically, DualCNN outperforms the other methods on Part A in terms of the MAE and is outperformed by HA-CNN in terms of the MSE. The MAE and MSE of DualCNN on Part B (UCF_CC_50) are 7.2 and 11.1, respectively, which are slightly better than those of the suboptimal model HA-CNN. The MAE and MSE of DualCNN on UCF_CC_50 are 213.8 and 313.9, respectively, which are lower than those of the suboptimal model CAT-CNN with 214.2 and 318.0 in terms of MAE and MSE, respectively.

4.5 Robustness of dualCNN

To validate the robustness of DualCNN, MCNN [75], CSRNet [34] and HA-CCN [59] were selected as the primal networks of the proposed dual framework. The corresponding dual convolutional neural networks are named DualCNN-MCNN, DualCNN-CSRNet and DualCNN-HA-CCN.

Because the MCNN includes two down-sampling layers with a kernel size of 2×2 and stride equal to 2, the width and height of the mapped density map are reduced to 1/4 those of the crowd image. Therefore, the dual network requires two deconvolution (upsampling) layers with a stride of 2 to recover the size of the crowd image. In addition, MCNN uses one input channel of the crowd image; therefore, the last layer of the dual network of DualCNN-MCNN requires a 1×1 convolution kernel and one channel to output the rough crowd image. Figure 11 shows the architecture and data flow of DualCNN-MCNN.

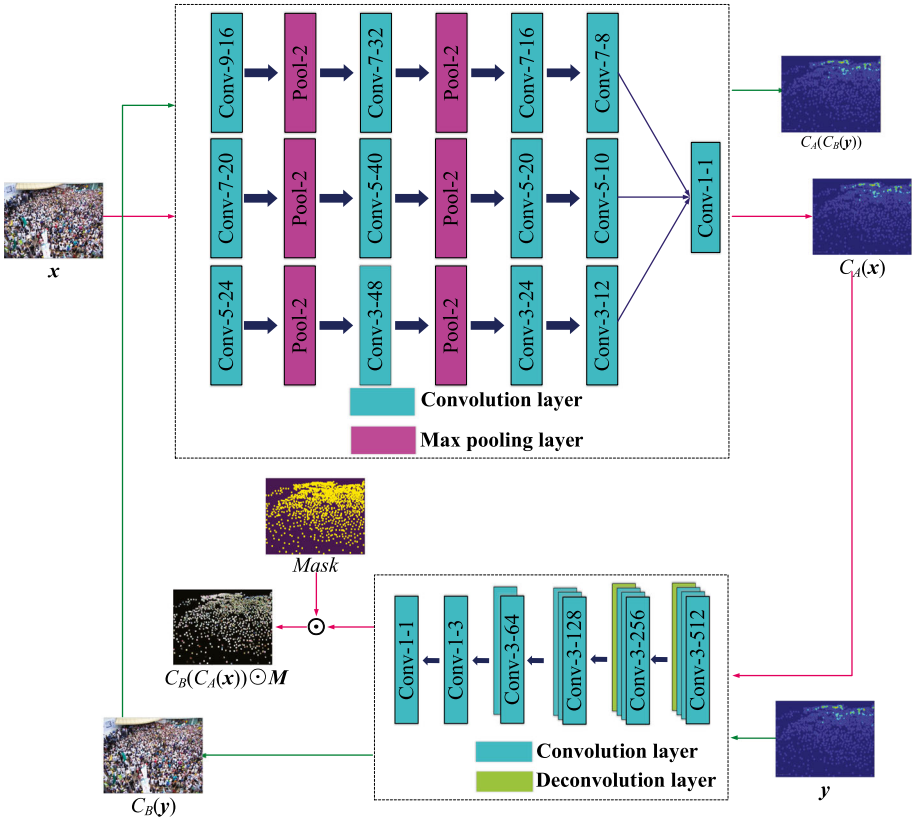

Fig. 11 Architecture and data flow of DualCNN-MCNN

Table 3 shows the MAE and MSE of the models on the ShanghaiTech and UCF_CC_50 datasets. We observe from Table 3: 1) DualCNN slightly improves the performance of MCNN, CSRNet and HA-CCN on dataset ShanghaiTech including Part_A and Part_B in terms of measures of MAE and MSE; 2) DualCNN significantly improves the performance of MCNN, CSRNet and HA-CCN on UCF_CC_50, specifically DualCNN-MCNN (DualCNN-CSRNet, DualCNN-HA-CCN) reduces the MAE and MSE of MCNN, CSRNet and HA-CCN by 60.5 (42.5, 25.0) and 96.7 (17.4, 33.6) respectively. These results indicate that the proposed dual-learning framework can effectively improve conventional convolutional neural networks on the crowd counting problem.

Table 3 shows the MAE and MSE of the models on the ShanghaiTech and UCF_CC_50 datasets. Table 3 shows that: 1) DualCNN slightly improves the performance of MCNN, CSRNet and HA-CCN on the ShanghaiTech dataset, including Part_A and Part_B, in terms of measures of MAE and MSE; 2) DualCNN significantly improves the performance of MCNN, CSRNet and HA-CCN on UCF_CC_50, specifically DualCNN-MCNN (DualCNN-CSRNet, DualCNN-HA-CCN) reduces the MAE and MSE of MCNN, CSRNet and HA-CCN by 60.5 (42.5, 25.0) and 96.7 (17.4, 33.6), respectively. These results indicate that the proposed dual-learning framework effectively improves the performance of CNNs for crowd counting.

Table 3 Experimental results of ablation experiment on ShanghaiTech

Method	ShanghaiTech				UCF_CC_50	
	Part_A		Part_B		MAE	MSE
	MAE	MSE	MAE	MSE		
MCNN [75]	110.2	173.2	26.4	41.3	377.6	509.1
DualCNN-MCNN	105.7	164.2	20.5	34.0	317.1	412.4
CSRNet [34]	68.2	115.0	10.6	16.0	291.4	349.4
DualCNN-CSRNet	66.3	113.4	9.1	15.2	248.9	332.0
HA-CCN [59]	62.9	94.9	8.1	13.4	256.2	348.4
DualCNN-HA-CCN	61.3	92.4	7.5	10.2	231.2	314.8

Combining the experimental results, we conclude:

- The proposed DualCNN considers the reconstruction from the density map to the crowd image and the impact of this reconstruction on the CNN performance. Therefore, DualCNN shows high performance on the crowd counting problem and effectively improves the performance of conventional CNN-based methods for the problem;
- we introduce the attention mechanism into DualCNN to enhance the primal network robustness against the background influence of the crowd image, which furthermore improves the performance of CNN-based methods on the crowd counting problem.

5 Conclusion and future work

In this paper, we presented a novel dual-learning network framework called DualCNN to improve the performance of convolutional neural networks (CNNs) for crowd counting. DualCNN comprises two sub-CNNs: one used for the primal task (i.e., generating high-quality density maps from crowd images) and the other used for the dual task (i.e., generating crowd images from density maps). DualCNN iteratively and alternatively learns the two sub-CNNs, and it thus enhances the performance of conventional convolution neural networks on the crowd counting problem by utilizing these network interactions. In addition, an attention mechanism is introduced to further improve the performance of DualCNN by reducing the impact of the background of the crowd image on the proposed dual framework. Experimental results show that DualCNN significantly outperforms state-of-the-art methods on the crowd counting problem.

We discussed how crowd maps with different backgrounds may result in the same density map, but a density map can only map to a single crowd map. Therefore, a key to the success of the dual network is to reduce the impact of the image background on the network for the dual task, i.e., generating crowd images from density maps. The space-based attention mechanism effectively addresses this problem, as discussed in Section 3.3. Therefore, one future work of ours will be to design novel space-based attention modules that further reduce the design impact of the image background on the DualCNN. In addition, only three deep neural networks were applied here as candidates for validating the robustness of DualCNN on improving the performance of other crowd-count oriented models. Consequently, more experiments for validating the robustness of DualCNN are another possible future direction of this work.

Acknowledgements We wish to thank Xu Mingliang's team of Zhengzhou University and Wensheng Zhang' team of Chinese Academy of Sciences for their constructive comments and recommendations, which have significantly improved the presentation of this paper.

Funding This work is supported in part by the Natural Science Foundation of Henan Province(No. 222300420274 and No. 222300420275), and in part by Science and Technology Research key Project of the Education Department of Henan Province (No. 22A520008).

Data Availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflicts of Interest The authors declare no conflict of interest.

References

1. Abdou M, Erradi A (2020) Crowd Counting: A Survey of Machine Learning Approaches. IEEE International Conference on Informatics, IoT, and Enabling Technologies, Doha, Qatar, pp 48–54
2. Ali S, Bouguila N (2019) Dynamic Texture Recognition using a Hybrid Generative-Discriminative Approach with Hidden Markov Models and Support Vector Machines. IEEE Global Conference on Signal and Information Processing, pp 1–5
3. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp 6077–6086
4. Cao X, Wang Z, Zhao Y, Su F (2018) Scale Aggregation Network for Accurate and Efficient Crowd Counting. Proceedings of 15th European conference on computer vision, Part V, Munich, Germany, pp 757–773
5. Chan AB, Liang ZSJ, Vasconcelos N (2008) Privacy preserving crowd monitoring: Counting people without people models or tracking. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Anchorage, Alaska, USA
6. Chen Y, Li D, Zhang JQ (2019) Complementary Color Wavelet: A Novel Tool for the Color Image/Video Analysis and Processing. IEEE Trans Circuits Syst Video Technol 29(1):12–27
7. Chen J, Su W, Wang Z (2020) Crowd Counting with Crowd Attention Convolutional Neural Network. Neurocomput 382:210–220
8. Chen K, Wang J, Chen L-C, Gao H, Xu W, Nevatia R (2015) ABC-CNN: An Attention Based Convolutional Neural Network for Visual Question Answering. CoRR abs/1511.05960
9. Chen L, Zhang H, Xiao J, Nie L, Shao J, Liu W, Chua T-S (2017) SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp 6298–6306
10. Chu X, Yang W, Ouyang W, Ma C, Yuille AL, Wang X (2017) Multi-context Attention for Human Pose Estimation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp 1831–1840
11. Dollár P, Wojek C, Schiele B, Perona P (2012) Pedestrian Detection: An Evaluation of the State of the Art. IEEE Trans Pattern Anal Mach Intell 34(4):743–761
12. Dwibedi D, Aytar Y, Tompson J, Sermanet P, Zisserman A (2020) Counting Out Time: Class Agnostic Video Repetition Counting in the Wild. IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Seattle, WA, USA. pp 10384–10393
13. Foadian S, Pourgholi R, Tabasi SH, Damirchi J (2019) The inverse solution of the coupled nonlinear reaction-diffusion equations by the Haar wavelets. Int J Comput Math 96(1):105–125
14. Gao J, Wang Q, Li X (2020) PCC Net: Perspective Crowd Counting via Spatial Convolutional Network. IEEE Trans Circuits Syst Video Technol 30(10):3486–3498
15. Gao G, Gao J, Liu Q, Wang Q, Wang Y (2020) CNN-based Density Estimation and Crowd Counting: A Survey. CoRR abs/2003.12783
16. Hassen KBA, Machado JJM, Tavares JMRS (2022) Convolutional Neural Networks and Heuristic Methods for Crowd Counting: A Systematic Review. Sensors 22(14):5286

17. He S, Minn KT, Solnica-Krezel L, Anastasio MA, Li H (2021) Deeply-supervised density regression for automatic cell counting in microscopy images. Med Image Anal 68:101892
18. He D, Xia Y, Qin T, Wang L, Yu N, Liu TY, Ma WY (2016) Dual Learning for Machine Translation. Annual Conference on Neural Information Processing Systems, Barcelona, Spain, pp 820–828
19. Hu J, Shen L, Sun G (2018) Squeeze-and-Excitation Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp 7132–7141
20. Idrees H, Saleemi I, Seibert C, Shah M (2013) Multi-source multi-scale counting in extremely dense crowd images. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, pp 2547–2554
21. Jiang X, Xiao Z, Zhang B, Zhen X, Cao X, Doermann DS, Shao L (2019) Crowd Counting and Density Estimation by Trellis Encoder-Decoder Networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp 6133–6142
22. Jiang X, Zhang L, Xu M, Zhang T, Lv P, Zhou B, Yang X, Pang Y (2019) Attention Scaling for Crowd Counting. IEEE International Conference on Computer Vision (ICCV), Seattle, WA, USA, pp 4705–4714
23. Kingma DP, Ba J (2015) Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA
24. Kumar K, Deepti D (2018) Shrimankar: Deep Event Learning boosT-up Approach: DELTA. Multimed Tools Appl 77(20):26635–26655
25. Kumar K, Shrimankar DD (2018) F-DES: Fast and Deep Event Summarization. IEEE Trans Multimed 20(2):323–334
26. Kumari S, Singh M, Kumar K (2019) Prediction of liver disease using grouping of machine learning classifiers. Conference Proceedings of International Conference on Deep Learning, Artificial Intelligence and Robotics (ICDLAIR2019), pp 339–349
27. Kumar A, Purohit K, Kumar K (2021) Stock Price Prediction Using Recurrent Neural Network and Long Short-Term Memory. Conference Proceedings of International Conference on Deep Learning, Artificial Intelligence and Robotics (ICDLAIR), Salerno, Italy. Lecture Notes in Networks and Systems, vol 175, pp 153–160
28. Li T, Chang H, Wang M, Ni B, Hong R, Yan S (2015) Crowded scene analysis: A survey. IEEE Trans Circuits Syst Video Technol 25(3):367–386
29. Lin Z, Davis LS (2010) Shape-Based Human Detection and Segmentation via Hierarchical Part-Template Matching. IEEE Trans Pattern Anal Mach Intell 32(4):604–618
30. Liu Y, Wen Q, Chen H, Liu W, Qin J, Han G, He S (2020) Crowd Counting Via Cross-Stage Refinement Networks. IEEE Trans Image Process 29:6800–6812
31. Liu YB, Jia R, Liu QM, Zhang XL, Sun HM (2021) Crowd counting method based on the self-attention residual network. Appl Intell 51(1):427–440
32. Liu W, Salzmann M, Fua P (2019) Context-Aware Crowd Counting. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp 5099–5108
33. Liu B, Vasconcelos N (2015) Bayesian Model Adaptation for Crowd Counts. IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp 4175–4183
34. Li Y, Zhang X, Chen D (2018) Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp 1091–1100
35. Mallasto A, Feragen A (2018) Wrapped Gaussian Process Regression on Riemannian Manifolds. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp 5580–5588
36. Miao Y, Lin Z, Ding G, Han J (2020) Shallow Feature Based Dense Attention Network for Crowd Counting. The Thirty-Fourth AAAI Conference on Artificial Intelligence, New York, NY, USA, pp 11765–11772
37. Negi A , Kumar K, Chaudhari NS, Singh N, Chauhan P (2021) Predictive Analytics for Recognizing Human Activities Using Residual Network and Fine-Tuning. Proceedings of the 9th International Conference on Big Data Analytics,Virtual Event, pp 296–310
38. Negi A, Kumar K (2021) Classification and Detection of Citrus Diseases Using Deep Learning. Data Science and Its Applications, In book, pp 63–85
39. Negi A, Kumar K (2021) Face Mask Detection in Real-Time Video Stream Using Deep Learning. Computational Intelligence and Healthcare Informatics, In book, pp 255–268
40. Negi A, Kumar K, Chauhan P (2021) Deep Neural Network-Based Multi-Class Image Classification for Plant Diseases. Agricultural Informatics, In book, pp 117–129
41. Negi A, Kumar K, Chauhan P (2021) Deep Learning-Based Image Classifier for Malaria Cell Detection. Machine Learning for Healthcare Applications, In book, pp 187–197

42. Negi A, Chauhan P, Kumar K, Rajput RS (2020) Face Mask Detection Classifier and Model Pruning with Keras-Surgeon. 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, India, pp 1–6
43. Nguyen V, Ngo TD (2020) Single-image crowd counting: a comparative survey on deep learning-based approaches. Int J Multimed Inf Retriev 9(2):63–80
44. Park J, Woo S, Lee J-Y and Kweon IS (2018) BAM: Bottleneck Attention Module. CoRR abs/1807.06514
45. Pham V-Q, Kozakaya T, Yamaguchi O, Okada R (2015) COUNT Forest: CO-Voting Uncertain Number of Targets Using Random Forest for Crowd Density Estimation. IEEE International Conference on Computer Vision (CVPR), Santiago, Chile, pp 3253–3261
46. Qi W, Gao J, Lin W, Yuan Y (2021) Pixel-Wise Crowd Understanding via Synthetic Data. Int J Comput Vision 129(1):225–245
47. Rehman YAU, Po L, Liu M, Zou Z, Ou W (2019) Perturbing Convolutional Feature Maps with Histogram of Oriented Gradients for Face Liveness Detection. International Joint Conference: 12th International Conference on Computational Intelligence in Security for Information Systems and 10th International Conference on EUropean Transnational Education, Seville, Spain, pp 3–13
48. Ryan D, Denman S, Fookes C, Sridharan S (2009) Crowd Counting Using Multiple Local Features. Techniques and Applications, Melbourne, Australia, Digital Image Computing, pp 81–88
49. Sam DB, Sajjan NN, Babu RV, Srinivasan M (2018) Divide and Grow: Capturing Huge Diversity in Crowd Images With Incrementally Growing CNN. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp 3618–3626
50. Sam DB, Surya S, Babu RV (2017) Switching Convolutional Neural Network for Crowd Counting. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp 4031–4039
51. Sharma S, Kumar K, Singh N, (2017) D-FES: Deep facial expression recognition system. (2017) Conference on Information and Communication Technology (CICT). Gwalior, India, pp 1–6
52. Simonyan K, Zisserman A (2015) Very Deep Convolutional Networks for Large-Scale Image Recognition. Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA
53. Sindagi VA, Patel VM (2017) Generating High-Quality Crowd Density Maps Using Contextual Pyramid CNNs. IEEE International Conference on Computer Vision (ICCV), Honolulu, HI, USA, pp 1879–1888
54. Sindagi VA, Patel VM (2019) Inverse Attention Guided Deep Crowd Counting Network. IEEE International Conference on Advanced Video and Signal Based Surveillance, Taipei, Taiwan, pp 1–8
55. Sindagi VA, Patel VM (2017) A Survey of Recent Advances in CNN-based Single Image Crowd Counting and Density Estimation. Pattern Recognit Lett 107:3–16
56. Tian Y, Mirzabagheri M, Bamakan SMH, Wang H, Qu Q (2018) Ramp loss one-class support vector machine; A robust and effective approach to anomaly detection problems. Neurocomput 310:223–235
57. Vijayvergia A, Kumar K, (2018) STAR: rating of reviewS by exploiting variation in emoTions using trAnsfer leaRning framework. (2018) Conference on Information and Communication Technology (CICT). Busan, South Korea, pp 1–6
58. Viresh R, Le HM, Hoai M (2018) Iterative Crowd Counting. Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, pp 278–293
59. Vishwanath S, Vishal MP (2020) HA-CCN: Hierarchical Attention-Based Crowd Counting Network. IEEE Trans Image Process 29:323–335
60. Wang J, Jiang W, Ma L, Liu W, Xu Y (2018) Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp 7190–7198
61. Wang F, Jiang M, Qian C, Yang S, Li C, Zhang H, Wang X, Tang X (2017) Residual Attention Network for Image Classification. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, US, pp 3156–3164
62. Wang C, Zhang H, Yang L, Liu S, Cao X (2015) Deep people counting in extremely dense crowds. Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, pp 1299–1302l
63. Woo S, Park J, Lee JY, Kweon IS (2018) CBAM: Convolutional Block Attention Module. Proceedings of the 15th European Conference on Computer Vision, Munich, Germany, pp 3–19
64. Xu H, Saenko K (2016) Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering. Proceedings of the 14th European Conference on Computer Vision, Part VII, Amsterdam, The Netherlands, pp 451–466
65. Yan C, Li Y, Liu W, Li M, Chen J, Wang L (2020) An artificial bee colony-based kernel ridge regression for automobile insurance fraud identification. Neurocomput 393:115–125
66. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked Attention Networks for Image Question Answering. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp 21–29

67. Yang Y, Li G, Wu Z, Su L, Huang Q, Sebe N (2020) Reverse Perspective Network for Perspective-Aware Object Counting. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA. pp 4373–4382
68. Yi Z, Zhang H, Tan P, Gong M (2017) DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp 2868–2876
69. Yu JT, Jia RS, Li YC, Sun HM (2022) Automatic fish counting via a multi-scale dense residual network. Multimed Tools Appl 81(12):17223–17243
70. Zhang B, Wang N, Zhao Z, Abraham A, Liu H (2021) Crowd Counting Based on Attention-Guided Multi-Scale Fusion Networks. Neurocomput 451:12–24
71. Zhang C, Li H, Wang X, Yang X (2015) Cross-scene crowd counting via deep convolutional neural networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp 833–841
72. Zhang A, Shen J, Xiao Z, Zhu F, Zhen X, Cao X, Shao L (2019) Relational Attention Network for Crowd Counting. IEEE International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp 6787–6796
73. Zhang X, Wang T, Qi J, Lu H, Wang G (2018) Progressive Attention Guided Recurrent Network for Salient Object Detection. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp 714–722
74. Zhang A, Yue L, Shen J, Zhu F, Zhen X, Cao X, Shao L (2019) Attentional Neural Fields for Crowd Counting. IEEE International Conference on Computer Vision (ICCV), Seoul, Korea (South), pp 5713–5722
75. Zhang Y, Zhou D, Chen S, Gao S, Ma Y (2016) Single-Image Crowd Counting via Multi-Column Convolutional Neural Network. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp 589–597
76. Zhou F, Zhao H, Zhang Y, Zhang Q, Liang L, Li Y, Duan Z (2022) COMAL: compositional multi-scale feature enhanced learning for crowd counting. Multimed Tools Appl 81(15):20541–20560
77. Zhu M, Wang X, Tang J, Wang N, Qu L (2020) Attentive Multi-stage Convolutional Neural Network for Crowd Counting. Pattern Recognit Lett 135:279–285
78. Zhu A, Zheng Z, Huang Y, Wang T, Jin J, Hu F, Hua G, Snoussi H (2022) CACrowdGAN: Cascaded Attentional Generative Adversarial Network for Crowd Counting. IEEE Trans Intell Transp Syst 23(7):8090–8102