Efficient Crowd Counting Model using Feature Pyramid Network and ResNeXt

Kalyani G (Kalyanichandrak@gmail.com)
Velagapudi Ramakrishna Siddhartha Engineering College https://orcid.org/0000-0001-8544-6930
Janakiramaiah B
Prasad V Potluri Siddhartha Institute Of Technology
Narasimha Prasad L.V
Institute of Aeronautical Engineering
Karuna A
JNTUK UCEV: Jawaharlal Nehru Technological University Kakinada University College of Engineering Vizianagaram
Mohan Babu A
Audhishankara college of Engineering & Technology

Research Article

Keywords: Crowd Counting, Density Map, ResNeXt, Feature Pyramid Network, Feature Map

Posted Date: April 23rd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-438067/v1

Version of Record: A version of this preprint was published at Soft Computing on July 5th, 2021. See the published version at https://doi.org/10.1007/s00500-021-05993-x.

Noname manuscript No.
(will be inserted by the editor)

Efficient Crowd Counting Model using Feature Pyramid Network and ResNeXt

G Kalyani[1] · **B Janakiramaiah**[2] · **L V Narasimha Prasad** [3] · **A Karuna**[4] · **A Mohan Babu** [5]

Received: date / Accepted: date

Abstract Crowd counting is one of the most challenging issues in computer vision community for safety and security through surveillance systems. It has extensive range of applications, such as disaster management, surveillance event detection, intelligence gathering and analysis, public safety control, traffic monitoring, design of public spaces, anomaly detection and military. Early approaches still encounter many issues, like non-uniform density distribution, partial occlusion and discrepancies in scale and point of view. To address the above problems, Feature Pyramid Networks are introduced in deep convolution networks for counting the individuals in the Crowd. The designed network has extracted the features at all resolutions and is constructed rapidly from

G Kalyani
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, Andhra Pradesh, india.
E-mail: kalyanichandrak@gmail.com

B. Janakiramaiah
Prasad V. Potluri Siddhartha Institute of Technology
Vijayawada, Andhra Pradesh, india.
E-mail: bjanakiramaiah@gmail.com

L V Narasimha Prasad
Institute of Aeronautical Engineering
Hyderabad, Telangana, India.
E-mail: lvnprasad@iare.ac.in

A Karuna
University College of Engineering Kakinada(A)
Jawaharlal Nehru Technological University Kakinada,
Andhra Pradesh, India.
E-mail: karunagouthana@gmail.com

A Mohan Babu
Audhishankara College of Engineering & Technology,
Gudur, Andhra Pradesh, India.
E-mail: mohanphy57@gmail.com

only one input image. This method achieves out performance results compared to the well-known networks on three standard crowd counting datasets.

Keywords Crowd Counting · Density Map · ResNeXt · Feature Pyramid Network · Feature Map.

1 Introduction

Human populace on the planet is expanding significantly. This development, subsequently from development and urbanization around the world, has by implication made enhancement in the crowd. Huge social gathering occasions of individuals can be seen at shrouded zones, for example, in building corridors, air terminals and arenas and also in open zones like walkways, parks, sport occasions, political meetings and public exhibitions. Figure 1 shows some pictures of the above locations. The motivation behind the social occasions has significant impact for the huge scope properties and practices of the group. Subsequently the investigation of group elements and practices is a subject of extraordinary premium in numerous logical explores in psychology, sociology, public administrations, security and computer vision.

Crowd disruption is a general cause for crowd catastrophes, which are due to pushing, mass-frenzy, group smashes, and leads to overall loss of control [1]. There exist numerous misfortunes to delineate this issue like Water Festival stampede 2010 in Colombia where 380 people died approximately [2], [3]. Another renowned group pulverize scenario that has been concentrated a lot occurred in 2010 Love Parade music festival in Germany, where 21 members died and more than 500 were effected [4], [5], [6]. Sample images of the above two scenarios is shown in Figure 2. Some other destructive crowd scenarios of are presented in Table 1. To forestall such destructive mishaps, early programmed recognition of basic and strange circumstances in huge scope crowd is required. It would surely help, accordingly, to settle on fitting choices for crisis and well being control.

Table 1 Some Example Destructive Crowd Scenarios.

Event Name	Location	Year	Number of deaths
Hindu festival	Datia	2013	115
Loveparademusicfestival	Duisburg	2010	21
WaterFestival	PhnomPhen	2010	more than 380
Pilgrimage	Mena	2006	363
ReligiousProcession,	Baghdad	2005	more than 640

Intelligent visual reconnaissance at region under perception is widely concentrated by researchers of computer vision domain [6], [7]. Intelligent visual reconnaissance includes exact information preparing, proficient data combination and requires many less human administrators. Actually, it has an extraordinary benefit contrasted with the customary CCTV innovations which

Fig. 1 Images of various crowded scenes.(a) Parade (b) Sports Stadium (c) Musical Concert (d) Political Rally.

Fig. 2 Pictures before crowd crush of Water Festival stampede and Love Parade music event 2010.

require an enormous count of individual administrators, more human asset cost, to continually screen reconnaissance cameras. Analyzing the crowd is quite is possibly the most difficult assignments in such smart visual observation frameworks. It may be utilized for automated discovery of basic group level, recognizing and checking individuals, and furthermore identifying of abnormalities and disturbing group blemishes. Moreover it very well may be utilized for tracing the people or a gathering of individuals in a group [8]. Counting crowd stream is a significant video-outline dissecting measure in crowd examination since density of the crowd is one of the essential portrayals of the crowd status. Automatic procedures for crowd density assessment and tallying had got lot of attention in security control and assume a fundamental part in crowd management. It is also very well can be utilized for estimating the solace level of the group and recognizing possible danger to forestall over crowd disasters. In visual checking frameworks, the group size is one of the significant essential pointers for identifying dangers like revolting, savage dissent, battling, mass frenzy [9], [10], [11].

Crowd estimation and analysis has assortment of important applications some of which are as follows:

- Safety observing: The utilization of video observation cameras for security and well being points in different places, for example, sports arenas, places of interest, shopping centers and air terminals has empowered simpler checking of group in such situations.
- Disaster Management: Many situations including group social affairs, for example, games, music shows, public showings and political meetings countenance the danger of group interrelated calamities, for example, rushes that can be hazardous. In the above said cases, crowd examination can be utilized as a compelling apparatus for early congestion recognition and proper administration of group, subsequently, possible repugnance of any catastrophe [12], [13].
- Devise of public spaces: Crowd examination on accessible public spots like air terminals, train stations, shopping centers and other public structures [14] can uncover significant plan inadequacies from group well being and accommodation perspective.
- Virtual conditions: Crowd investigation strategies can be utilized to comprehend the basic marvel accordingly empowering us to set up numerical models that can give precise reproductions. These numerical models can be additionally utilized for recreation of group wonders for different applications, for example, PC games, embedding special visualizations in film scenes and planning clearing plans [15],[16].
- Forensic investigate: Crowd investigation also be utilized to look for suspect and casualties in occasions like besieging, shooting or mishaps in huge social events. Conventional face discovery and acknowledgement calculations can be speeded up utilizing crowd examination strategies which are more skilled at dealing with such situations [17].

– Strategic planning in Defence: In military and security applications, people counting plays a crucial role to analyse the crowd and for taking the strategic decisions at the time of war according to the crowd observed.

In this work, we plan a deep network named as ResNeXtFP network with feature sharing to check individuals and gauge full-resolution density maps. Convolutional kernels of 3x3 at pyramid layers for extracting the features while looking after resolution. Skip associations are used to coordinate multi-scale semantics and to expand scale insight capacity with less boundaries than multi-section models. The remainder of this paper is arranged as: Section 2 quickly presents some new procedures on swarm tallying. Followed by the specific information about our model and planning courses of action in Section 3. In Section 4, three public testing swarm datasets are clarified alongside the presentation measures and result relationships among top tier methodologies and our own. Finally section 5 finishes up the paper.

2 Review of Literature

Based on the characteristics of the available methods for estimating the crowd density and counting the individuals, they are categorized into direct and indirect methods. The direct methodologies attempts to fragment and identify every person in the crowd scenes and after ward counting them by considering a classifier. In this strategy, counting the individuals can be done as long as individuals are effectively segmented yet the interaction can be more perplexing when an extreme group or impediments happened. In the indirect methodologies, individuals counting is conveyed typically utilizing the estimations of certain highlights with learning calculations or statistical examination of the entire group to accomplish count measure. Indirect strategies are further classified as pixel-based, texture-based and corner-point analysis techniques. In later years the achievement of Convolutional Neural Networks (CNNs) in a variety of computer vision tasks has propelled scientists to utilize their capabilities for getting non-linear functions from crowd pictures to their relating density maps or count of individuals. An assortment of CNN-based techniques has been introduced in the literature. CNN-based techniques are categorized by considering the property of the network and training process. In view of the property of the networks, the methodologies are categorized into the accompanying classes: fundamental CNNs, Scale-aware models, Context-aware models, multi-tasks systems. There exists another type of categorization, with respect to the inference strategy into two categories which are Patch-based inference, Whole picture based inference techniques.

To handle the issue of crowd density estimation, the use of CNNs has started by the authors Wang et al. [18] and Fu et al. [19]. Wang et al. proposed a regression model based on a deep CNN for counting individuals from pictures in very dence groups. They embraced AlexNet [20] in the framework where the last completely associated layer of 4096 neurons is supplanted with

a solitary neuron layer for foreseeing the check. Moreover, to lessen bogus reactions, backgrounds like buildings and trees in the pictures, training data is expanded with extra negative instances whose ground truth count is set zero. In an alternate methodology, Fu et al. projected to characterize the picture into one of the five classes: very high, high, medium, low and very low density as opposed to approximate the density maps. Multi-stage ConvNet by Sermanet et al. [21] was received for enhanced move, scale and mutilation invariance. Likewise, they utilized a course of two classifiers to accomplish boosting in which the first explicitly tests misclassified pictures though the subsequent one reclassifies the misclassified instances.

Zhang et al. [22] examined the active techniques to recognize that the performance diminishes radically when applied to another scene that is not quite the same as the training data. To conquer this issue, they proposed a mapping function from the given image to crowd count. To accomplish this, they train their network alternatively on two related objective functions which are crowd counting and estimating the density. By training the network alternatively to optimize these two objectives one can acquire better neighbourhood optima. To adjust this network to another new scene, the trained network is fine-tuned by considering the training samples that are similar to the new scene. The main point to be noted in their approach is that the network is adjusted to new picture with no additional labeled data.

Assessing count of the crowds stays a difficult task because of the issues of scale varieties, non-uniform circulation and typical backgrounds. The authors of the paper [23] proposed a multi-goal consideration convolutional neural organization (MRA-CNN) to address the task of crowd counting. Aside from the task of counting, the authors used an extra grouping task at density level during training and consolidate highlights learned for the two tasks, consequently shaping multi-scale, multi-logical highlights to adapt to the scale variety and non-uniform appropriation. In addition, they used a multi-resolution attention (MRA) model to create score maps, in which head areas are with superior scores to train the network to have more attention on head locales and stifle non-head regions paying little mind to the unpredictable backgrounds. In the generation of score maps, atrous convolution layers are utilized to extend the receptive field with less number of parameters, accordingly getting more elevated level highlights and giving the MRA model more exhaustive data. The designed network was examined on ShanghaiTech, WorldExpo'10 and UCF datasets to show the viability of proposed network.

Scale variety due to viewpoint contortion is as yet a difficult task for examining the crowds. To tackle this issue, an atrous convolution spatial pyramid organization (ACSPNet) is introduced by the authors of [24], to perform swarm checks and density maps for sparse and crowded situations. Dilated convolutions sequenced with expanding dilation rates are used to misrepresent the responsive field and to keep up the goal of extricated highlights. Atrous Spa-

tial Pyramid Pooling (ASPP) is utilized to resample data at various scales and consists of global setting. The proposed ACSPNet is evaluated by authors on five benchmark datasets for crowd counting and they claimed that our strategy accomplishes minimum absolute error and squared error.

Applying the crowd counting on aerial images with the help of an embedded system is a difficult assignment, because of high quality pictures, low computing resources, and restricted memory. To handle this issue, the authors of the paper [25] proposed an effective deep learning model named Flounder-Net. In the Flounder-Net, a novel interleaved group of convolutions are used to dispense with the duplication of the network, and a fast shrink of feature maps is utilized to handle the issue of high-resolution. The authors of the work in [26], proposed an effective encoder-decoder framework, named MobileCount, which is explicitly intended for high accuracy continuous group counting on versatile or installed gadgets with restricted computation assets. For the encoder part, MobileNetV2 is custom fitted to altogether diminish FLOPs at somewhat cost of performance drop, which has four bottleneck blocks before a maximum pooling layer with stride 2. The plan of decoder is roused by Light-weight RefineNet, which added performance of counting by 10% expansion of FLOPs. The proposed framework accomplishes similar counting performance with 1=10 FLOPs with various benchmarks when compared to existing methods. Finally, the authors proposed a distillation method with multi-layer network to additionally lift the performance of the MobileCount without expanding its FLOPs.

Crowd counting has attracted far attention in the domain of computer vision. However it is amazingly difficult on account of the changing scales and densities. Various existing strategies centred on improving the multi-scale portrayal by using multi-section or multi-branch models with various kernel sizes. Nonetheless, those networks can't retrieve the feature maps with enormous receptive fields because of limit of profundity. Also, the significance of using the staggered highlight data in a deep network is disregarded. To handle this task the authors of the paper [27] proposed a multi-scale feature aggregation network (MFANet) for precise and effective group counting, and it tends to be trained in an end-to-end manner. A fundamental part of the network is the scale level aggregation module (SLAM), which can separate multi-scale highlights and utilize multi-level feature data for more precise assessment. The best performance has observed when six SLAMs are stacked together and applied in the network. Exploratory outcomes show that the proposed MFANet accomplishes cutting edge execution in group counting and localization of the group.

Numerous CNN-based counting methods achieve great performance. But, these strategies just focusing on the neighbourhood appearance highlights of group scenes however overlook the huge reach pixel-wise relevant and group consideration data. To handle the above issues, the authors of the paper [28],

presented the Spatial-/Channel-wise Attention Models into the conventional regression CNN to assess the density map, called as "SCAR". It comprises of two modules, specifically Spatial-wise and Channel-wise Attention Models. At last, two kinds of attention data and customary CNN's feature maps are incorporated by a concatenation operation. The authors claimed that the outcomes show that the proposed technique accomplishes cutting edge results.

In any case, existing networks which are based on CNN basically center with respect to improving precision yet once in a while think about the simplicity of network. In particular, they have the accompanying limits: 1) taking high computational intricacy [29],[30], 2) having such a large number of parameters [31],[32], 3) biased with fixed size of picture as input [33],[34]. These limitations cut-off the applicability of the techniques as embedded systems with restricted memory and computational force and limit the versatility of the network to an assortment of imaging equipment.

3 Proposed methodology

This section describes the proposed network by considering the concept of feature pyramid network based on ResNeXt network for counting the individuals in the work.

3.1 Feature Pyramid Network

Recognizing objects in various scales is generally a typical issue specifically in the case small objects. We can utilize a pyramid of same picture at various scales to recognize objects. Notwithstanding, handling various scale pictures is tedious and the memory requirement is too high for training all at the same time. On the other hand, we can make a pyramid of feature and utilize them for object discovery. Nonetheless, feature maps nearer to the input layer made out of low-level designs that are not powerful for exact object identification. Feature Pyramid Network (FPN) [35] is intended to extract the features for such pyramid concept in view of improving the accuracy and speed. It creates different feature map layers with preferred quality data over the ordinary feature pyramid for object recognition. FPN consists of bottom-up and top-down strategies. The bottom-up strategy is the standard convolution network for extracting the features. As we go up, the resolution diminishes. With all the more significant level designs distinguished, the semantic incentive for each layer increments. The bottom- up strategy is the feed-forward calculation of the convolution network which used in the background. It makes out of numerous convolution modules each has numerous convolution layers. As we climb, the spatial measurement is decreased. It is considered that single pyramid level is for every stage. The outcome of the final layer is used as the reference set of feature maps for advancing the top-down pathway by parallel association.

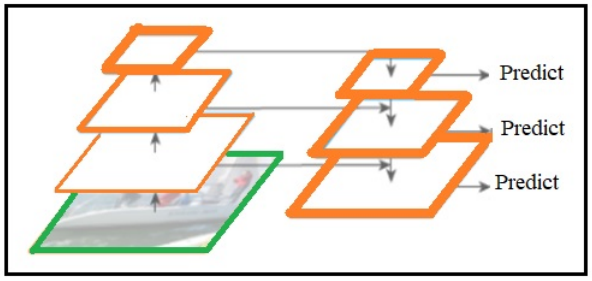

Fig. 3 Overview of FPN

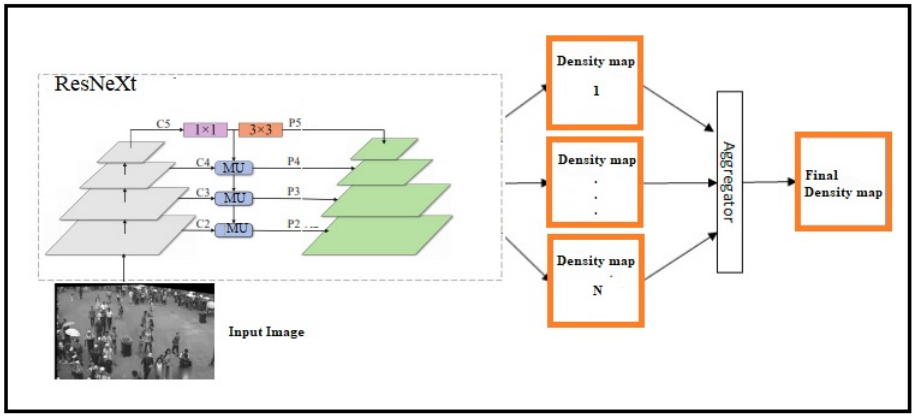

Fig. 4 Proposed ResNextFPNet Architecture

In Top-down strategy, the higher resolution feature is up-sampled spatially coarser, however semantically more grounded, include maps from upper levels of pyramid. All the more explicitly, the spatial goal is up-sampled by a factor of 2 utilizing the closest neighbour. Every lateral connection consolidates feature maps of a similar spatial size from the bottom-up strategy and the top-down strategy. The outline of FPN is shown in Figure 3.

3.2 Proposed ResNeXtFP Network Architecture

The structural design of the network which is designed in this paper for crowd counting is called ResNeXt based Feature Pyramid Network (ResNextFPNet) is shown in Figure 4. In this work, ResNeXt is used as a backbone network for FPN. We planned the model to have the option to use more than one output with various scale from FPN for counting the individuals in the crowd. To oblige this capability, the FPN is designed in such a way that each feature pyramid from FPN to create an intermediate density map. Thereafter, all density maps are accumulated with an aggregator module to create a final

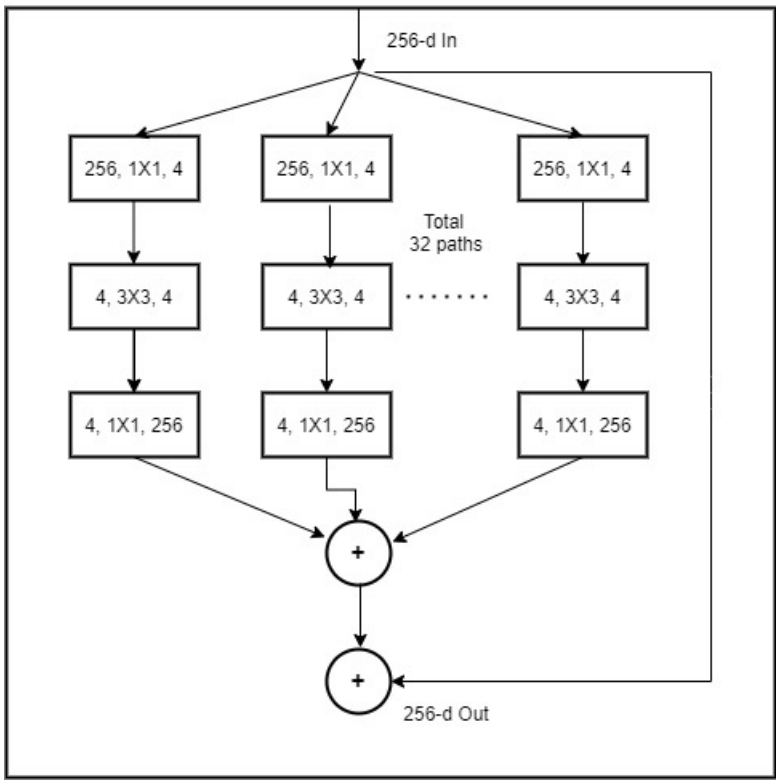

Fig. 5 A single block of ResNeXt

density map. The aggregator module is designed with two convolution layers. The primary layer consists of 128 kernels. The size of these kernels is set to 5x5 for locating the neighbourhood spatial information. The second layer utilizes just a single kernel with the size of 1x1 for producing the final density map.

As illustrated in Figure 3, the backbone convolution network that is utilized is, ResNeXt [36]. The structure of ResNeXt basically mirrors the ResNet [37]. The architecture of the ResNeXt is an extention of the deep residual network which changes the standard residual block with the one that use a strategy called "split-transform merge" which is adopted from the Inception models. Basically, as opposed to performing convolutions over the full feature map, the block's information is projected into a progression of lower dimensional portrayals of which we independently apply a couple of convolution channels prior to merging the outcomes. The single block of ResNeXt is illustrated in Figure 5.

For every path, Conv1x1, Conv3x3, Conv1x1 are done in sequence. The internal measurement for every path is meant as d (here d=4). Cardinality of the block which represents the number of paths is denoted as C (here C=32).

Table 2 Details of ResNeXt Architecture

Stage	**Output**	**ResNeXt-50 (32 X4d)**
Conv1	112 X 112	7 X 7, 64, stride 2
Conv2	56 X 56	3X3 max pool, stride 2
		$\begin{bmatrix} 1X1, 128 \\ 3X3, 128, C = 32 \\ 1X1, 256 \end{bmatrix}$ X3
Conv3	28X28	$\begin{bmatrix} 1X1, 256 \\ 3X3, 256, C = 32 \\ 1X1, 512 \end{bmatrix}$ X4
Conv4	14X14	$\begin{bmatrix} 1X1, 512 \\ 3X3, 512, C = 32 \\ 1X1, 1024 \end{bmatrix}$ X6
Conv5	7X7	$\begin{bmatrix} 1X1, 1024 \\ 3X3, 1024, C = 32 \\ 1X1, 2048 \end{bmatrix}$ X3

We can summarize the dimensions of each Conv3x3 as 128 (i.e., dxC=4x32). The dimension is expanded straightforwardly from 4 to 256, and afterward added together, and furthermore added with the path of skip connection. The number of parameters in ResNeXt is Cx(256xd+3x3xdxd+dx256), with C=32 and d=4. All the blocks details of ResNeXt is included in Table 2.

In particular, FPN just follows up on the feature activation result by the residual block yield at each phase of ResNeXt, which are indicated as $\langle C2, C3, C4, C5\rangle$. It is a general requirement that feature maps of similar size only can be combined, so the significant upper level feature maps should be up-sampled prior to being combined with the bottom-level feature maps. Subsequently, we utilize a closest neighbor interpolation which will successfully lessen the checkerboard impact which exists in deconvolution and simple interpolation approaches. Besides, a convolution kernel of 1 x 1 size to reduce the dimensions and a 3 x 3 convolution to additional concentrate the low-layer data, and a layer of ReLU is used to acquire the non-linearity between the two convolution layers. At that point, the high level semantic features are combined with the low-level semantic features through the addition for element wise, and acquire the combined feature map with a 3 x 3 convolution and two layers of ReLu. The above procedure is repeated until the best resolution feature map is produced. At the end, a bunch of multi-scale feature maps re-

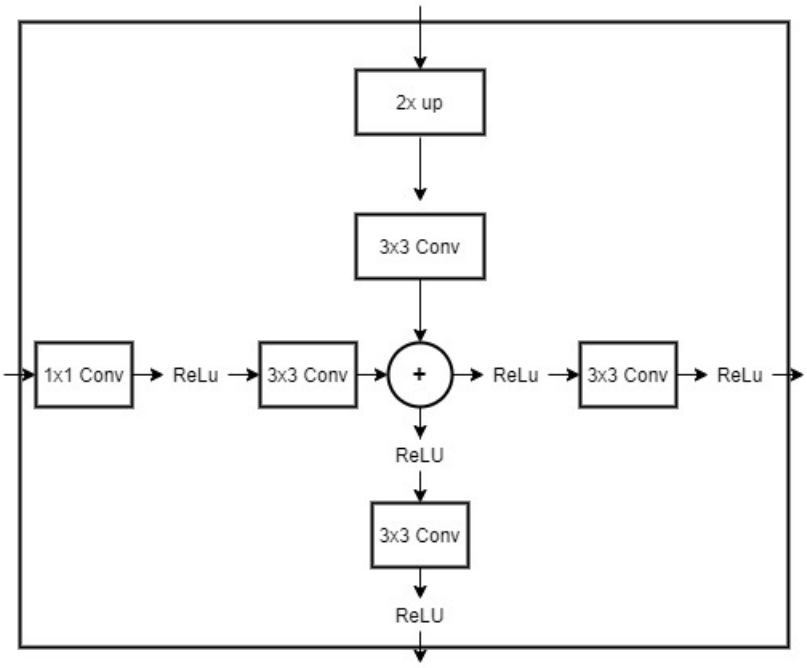

Fig. 6 The Merging Unit of FPN

lating to the merged one of each layer are produced which is characterized as $\langle P2, P3, P4, P5\rangle$. It is significant that P5 is acquired by C5 with a 1 x 1 convolution and a 3 x 3 convolution. This whole cycle of merging the down inspecting and upsampling features is shown in Figure 3 as 'Merging Unit (MU)'. The entire procedure used in 'MU' is illustrated in Figure 6.

4 Experimental Investigations

This section, first explores datasets considered for validating the proposed framework and evaluation metrics used for that validation. Then, the discussion on comparison of the proposed framework results with other frameworks by considering three popular crowd counting benchmark datasets is presented.

4.1 Datasets

The validation of the proposed framework is done with respect to the two standard datasets for the domain of crowd counting. The details of the datasets are as follows:

Table 3 Summary of crowd counting Datasets.

Dataset	Number of Images	Resolution	Average Crowd Counts
ShanghaiTech Part A	482	Varied	501.4
ShanghaiTech Part B	716	768x1024	123.6
UCF_CC_50	50	Varied	1,279.5

- ShanghaiTech dataset: Zhang et al.[38] established a new huge crowd counting dataset in the year 2016, which includes 1198 still images with a total amount of 330165 annotated heads. It is mainly divided into Part_A and Part_B. Part_A includes 482 high-density varied resolution images with average crowd counting of 501.4, Part_B includes 716 images with the fixed resolution(768X1024) images with average crowd counting of 123.6.
- UCF-CC-50: H.Idress et al.[39] introduced the small, challenging, and large variance dataset for crowd counting in the year 2013, which contains 50 images with a total of 63,974 head center annotations. It includes varied resolution images with average crowd counting of 1,279.5.

Summary of the datasets used in shown in Table 3.

4.2 Performance Metrics

Regularly the evaluation of crowd counting models was done by Mean Absolute Error and Mean Squared Error metrics.

$$MeanAbsoluteError(MAE) = \frac{1}{N}\sum_{i=1}^{N} \mid z_i - \hat{z_i} \mid \tag{1}$$

$$MeanSquaredError(MSE) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mid z_i - \hat{z_i} \mid^2} \tag{2}$$

Where N is the count of test images, Z_i is the ground truth of the persons in the i^{th} image, Z_i is the predicated count in the i^{th} image. MAE specifies the accuracy of the evaluation, and MSE specifies the robustness of the evaluation. Lower value is better for both MAE and MSE.

4.3 Discussion on the Results

The performance of the proposed ResNeXtFP network is demonstrated with a four challenging benchmark crowd counting datasets and compared with general CNN architectures. The implementation of the proposed network is done in python because of the wide availability of the libraries and frameworks for deep learning. To build the FPN and deep learning architectures, Keras and TensorFlow are used in the backend. Experiments were done on DELL Power Edge R740 Server with 2 X Intel Xeon Gold 6226R- 2.9G, 16 C, 32T, 22 M

Table 4 Comparison of MAE and MSE on ShanghaiTech-Part-A dataset.

Framework	MAE	MSE
Multi-Column CNN [41]	110.2	173.2
Cascaded-MTL [42]	101.3	152.4
Switching-CNN[43]	90.4	135.0
CP-CNN [44]	73.6	106.4
Proposed (ResNeXt+FPN)	69.3	104.7

Cache, NVIDIA Quadro RTX6000, 24 GB GDDR6.

The evaluation of the projected framework is done with earlier well-known networks on the ShanghaiTech dataset. The efficiency of the projected framework is compared with four CNN based networks. The experimental results are demonstrated in Table 4 and Table 5 for Part-A and Part-B datasets of shanghaiTech dataset respectively. The results show that the proposed framework achieves the lowest MAE in both Part A and Part B compared to other CNN based frameworks. On the part-A dataset the existing multi-column CNN [41] has 110.2 and 173.2 as MAE and MSE respectively. The cascaded-MTL framework [42] achieves 101.3 and 152.4 as MAE and MSE respectively. The switching-CNN [43] got 90.4 and 135 MAE and MSE respectively. The proposed model performs superior than CP-CNN [44] which implements VGG-16 as backbone has 73.6 and 106.4 MAE and MSE respectively. Finally the proposed ResNeXtFP network achieves 69.3 and 104.7 MAE and MSE which are better when compared to the four existing frameworks. On the part-B dataset the existing multi-column CNN [41] has 26.4 and 41.3 as MAE and MSE respectively. The cascaded-MTL framework [42] achieves 20 and 31.1 as MAE and MSE respectively. The switching-CNN [43] got 21.6 and 33.4 MAE and MSE respectively. The proposed model performs better than CP-CNN [44] which has 20.1 and 30.1 MAE and MSE respectively. Finally the proposed ResNeXtFP network achieves 14.3 and 21.9 MAE and MSE which are better when compared to the four existing frameworks.

The comparison of the proposed framework is done with previous state-of-art networks on UCF-CC-50 dataset. For this dataset also the performance of the proposed framework is compared with four CNN based networks. The detailed results are illustrated in Table 6. The results show that the proposed framework attains the lowest MAE when compared to other CNN based frameworks. On UCF-CC-50 dataset the existing multi-column CNN [41] has 377.6 and 509.1 as MAE and MSE respectively. The cascaded-MTL framework [42] achieves 322.8 and 397.9 as MAE and MSE respectively. The switching-CNN [43] got 318.1 and 439.2 MAE and MSE respectively. The proposed model achieves better than CP-CNN [44] has 295.8 and 320.9 MAE and MSE re-

Table 5 Comparison of MAE and MSE on ShanghaiTech-Part-B dataset.

Framework	MAE	MSE
Multi-Column CNN [41]	26.4	41.3
Cascaded-MTL [42]	20.0	31.1
Switching-CNN[43]	21.6	33.4
CP-CNN [44]	20.1	30.1
Proposed (ResNeXt+FPN)	14.3	21.9

Table 6 Comparison of MAE and MSE on UCF-CC-50 dataset.

Framework	MAE	MSE
Multi-Column CNN [41]	377.6	509.1
Cascaded-MTL [42]	322.8	397.9
Switching-CNN[43]	318.1	439.2
CP-CNN [44]	295.8	320.9
Proposed (ResNeXt+FPN)	269.6	312.3

spectively. Finally the proposed ResNeXtFP network achieves 269.6 and 312.3 MAE and MSE which are better when compared to the four existing frameworks.

With respect to the three datasets the proposed ResNeXtFP Network outperforms some of the existing CNN based networks with minimum MAE and MSE.

5 Conclusion

In this work, we introduce a feature pyramid network named ResNeXtFP network for counting the individuals in medium or high-level crowd visible in a still image. The convolutions in of the background network are utilized to extract the multi-scale features, creating density maps with unaltered resolution. By utilizing the benefit of skip associations, our network can diminish excess features and total multi-scale data. The projected network further uses features at various scales, contemplating global semantics. Results on three datasets show that our projected ResNeXtFP Network can accomplish best in class exhibitions. In future work, profound measurement learning approaches might be worried to more readily recognize heads and the excess foundation data.

Conflict of interest

The authors declare that they have no conflict of interest.

Contributions

Each author has equally contributed in conceptualization, model building, simulation, and writing of the article.

Corresponding author

Correspondence to G. Kalyani

Ethical approval

This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Helbing,D.,Brockmann,D.,Chadefaux,T.,Donnay,K.,Blanke,U.,Woolley-Meza,O.,Moussaid, M., Johansson,A., Krause,J., Schutte,S., et al., 2014. Savinghuman lives: what Complexity science and information systems can contribute.J.Stat. Phys.,1–47.
2. Illiyas,F.T., Mani,S.K., Pradeepkumar,A., Mohan,K., 2013. Human stampedes during religious festivals: a comparative review of mass gathering emergencies in india.Int.J.DisasterRiskReduct.5,10–18.
3. Wang,J., Lo,S., Wang,Q., Sun,J.,Mu,H., 2013. Risk of large-scale evacuation based on the effectiveness of rescue strategies under different crowd densities.Risk Anal. 33(8),1553–1563.
4. Krausz, B.,Bauckhage,C.,2012. Love parade 2010: automatic video analysis of a crowd disaster.Comput.Vis.Image Underst.116(3),307–319.
5. Helbing, D.,Mukerji,P.,2012. Crowd disasters as systemic failures: analysis of the love parade disaster.EPJDataSci.1(1),1–40.
6. Shah, M.,Javed,O.,Shafique,K.,2007. Automated visual surveillance in realistic scenarios. IEEE Multimedia 14(1),30–39.
7. Hu, W.,Tan,T.,Wang,L.,Maybank,S.,2004. A survey on visual surveillance of object motion and behaviors.IEEE Trans.Syst. ManCybern.PartC Appl.Rev.34 (3), 334–352.
8. Aggarwal,J.,Ryoo,M.S.,2011. Human activity analysis: a review. ACM Comput. Surv.43(3),16.
9. Junior, S.J.,etal.,2010.Crowdan alysis using computer vision techniques. IEEE Signal Process.Mag.27(5),66–77.
10. Dittrich, F.,Koerich,A.,Oliveira,L.,2012.People counting in crowded scenes using multiple cameras.In:201219th International Conference on Systems, Signals and ImageProcessing(IWSSIP).IEEE,pp.138–141.
11. Chen, T.-Y.,Chen,C.-H.,Wang,D.-J.,Chen,T.-J.,2010.Real-time counting method for a crowd of moving people.In:2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP). IEEE, pp.643–646.
12. Abdelghany, A., Abdelghany, K., Mahmassani, H., Alhalabi, W., 2014. Modeling framework for optimal evacuation of large-scale crowded pedestrian facilities. European Journal of Operational Research 237, 1105–1118.

13. Almeida, J.E., Rosseti, R.J., Coelho, A.L., 2013. Crowd simulation modeling applied to emergency and evacuation simulations using multiagent systems. arXiv preprint arXiv:1303.4692 .
14. Chow, W.K., Ng, C.M., 2008. Waiting time in emergency evacuation of crowded public transport terminals. Safety Science 46, 844–857.
15. Gustafson, S., Arumugam, H., Kanyuk, P., Lorenzen, M., 2016. Mure: fast agent based crowd simulation for vfx and animation, in: ACM SIGGRAPH 2016 Talks, ACM. p. 56.
16. Perez, H., Hernandez, B., Rudomin, I., Ayguade, E., 2016. Task-based crowd simulation for heterogeneous architectures, in: Innovative Research and Applications in Next-Generation High Performance Computing. IGI Global, pp. 194–219.
17. Barr, J.R., Bowyer, K.W., Flynn, P.J., 2014. The effectiveness of face detection algorithms in unconstrained crowd scenes, in: Applications of Computer Vision (WACV), 2014 IEEEWinter Conference on, IEEE. pp. 1020–1027.
18. Wang, C., Zhang, H., Yang, L., Liu, S., Cao, X., 2015. Deep people counting in extremely dense crowds, in: Proceedings of the 23rd ACM international conference on Multimedia, ACM. pp. 1299–1302.
19. Fu, M., Xu, P., Li, X., Liu, Q., Ye, M., Zhu, C., 2015. Fast crowd density estimation with convolutional neural networks. Engineering Applications of Artificial Intelligence 43, 81–88.
20. Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.
21. Sermanet, P., Chintala, S., LeCun, Y., 2012. Convolutional neural networks applied to house numbers digit classification, in: Pattern Recognition (ICPR), 2012 21st International Conference on, IEEE. pp. 3288–3291.
22. Zhang, C., Li, H., Wang, X., Yang, X., 2015. Cross-scene crowd counting via deep convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 833–841.
23. Zhang, Y., Zhou, C., Chang, F., Kot, A. C. (2019). Multi-resolution attention convolutional neural network for crowd counting. Neurocomputing, 329, 144-152.
24. Ma, J., Dai, Y., Tan, Y. P. (2019). Atrous convolutions spatial pyramid network for crowd counting and density estimation. Neurocomputing, 350, 91-101.
25. Chen, J., Xiu, S., Chen, X., Guo, H., Xie, X. (2021). Flounder-Net: An efficient CNN for crowd counting by aerial photography. Neurocomputing, 420, 82-89.
26. Wang, P., Gao, C., Wang, Y., Li, H., Gao, Y. (2020). MobileCount: An efficient encoder-decoder framework for real-time crowd counting. Neurocomputing, 407, 292-299.
27. Zhu, F., Yan, H., Chen, X., Li, T., Zhang, Z. (2021). A multi-scale and multi-level feature aggregation network for crowd counting. Neurocomputing, 423, 46-56.
28. Gao, J., Wang, Q., Yuan, Y. (2019). SCAR: Spatial-/channel-wise attention regression networks for crowd counting. Neurocomputing, 363, 1-8.
29. L. Wen, D. Du, P. Zhu, Q. Hu, Q. Wang, L. Bo, S. Lyu, Drone-based joint density map estimation, localization and tracking with space-time multi-scale attention network (2019). arXiv:1912.01811.
30. Y. Yang, G. Li, Z. Wu, S. Li, N. Sebe, Reverse perspective network for perspective-aware object counting, in: IEEE Conference on Computer Vision and Pattern Recognition, 2020.
31. M. A. Hossain, M. Hosseinzadeh, O. Chanda, Y. Wang, Crowd counting using scale-aware attention networks, in: IEEE Winter Conference on Applications of Computer Vision, 2019.
32. Y. Fang, B. Zhan, W. Cai, S. Gao, B. Hu, Locality-constrained spatial transformer network for video crowd counting, in: IEEE International Conference on Multimedia and Expo, 2019.
33. C. Zhang, H. Li, X. Wang, X. Yang, Cross-scene crowd counting via deep convolutional neural networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 833–841.
34. Z. Shen, Y. Xu, B. Ni, M. Wang, J. Hu, X. Yang, Crowd counting via adversarial cross-scale consistency pursuit, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018.

35. Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125).
36. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K. (2017). Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1492-1500).
37. He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
38. Y. Zhang, D. Zhou, S. Chen, S. Gao and Y. Ma, "Single-Image Crowd Counting via Multi-Column Convolutional Neural Network," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 589-597, doi: 10.1109/CVPR.2016.70.
39. H. Idrees, I. Saleemi, C. Seibert and M. Shah, "Multi-source Multi-scale Counting in Extremely Dense Crowd Images," 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, pp. 2547-2554, doi: 10.1109/CVPR.2013.329.
40. Cong Zhang, Hongsheng Li, X. Wang and Xiaokang Yang, "Cross-scene crowd counting via deep convolutional neural networks," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 833-841, doi: 10.1109/CVPR.2015.7298684.
41. Zhang, Yingying, Desen Zhou, Siqin Chen, ShenghuaGao, and Yi Ma. (2016),Single-image crowd counting via multi-column convolutional neural network.In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 589-597.
42. V. A. Sindagi, V. M. Patel, Cnn-based cascaded multi-task learning of high-level prior and density estimation for crowd counting, in: IEEE International Conference on Advanced Video and Signal Based Surveillance,2017, pp. 1–6.
43. Deepak Babu Sam, Shiv Surya, and R. VenkateshBabu. Switching convolutional neural network for crowd counting.In IEEE Conference on Computer Vision and Pattern Recognition, pages 4031–4039, 2017.
44. Vishwanath A Sindagi and Vishal M Patel. Generating high-quality crowd density maps using contextual pyramid cnns. In IEEE International Conference on Computer Vision, pages 1879–1888, 2017

Figures

Figure 1

Images of various crowded scenes.(a) Parade (b) Sports Stadium (c) Musical Concert (d) Political Rally.

Figure 2

Pictures before crowd crush of Water Festival stampede and Love Parade music event 2010.

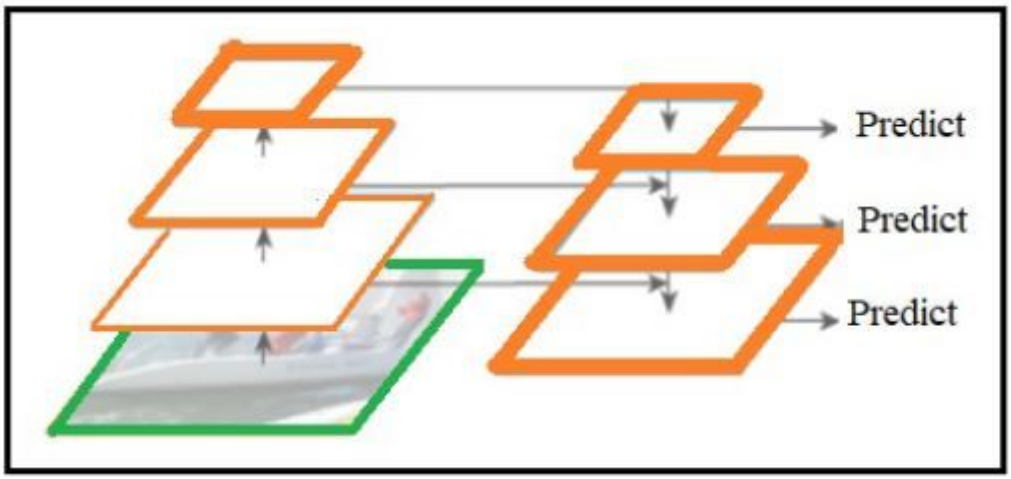

Figure 3

Overview of FPN

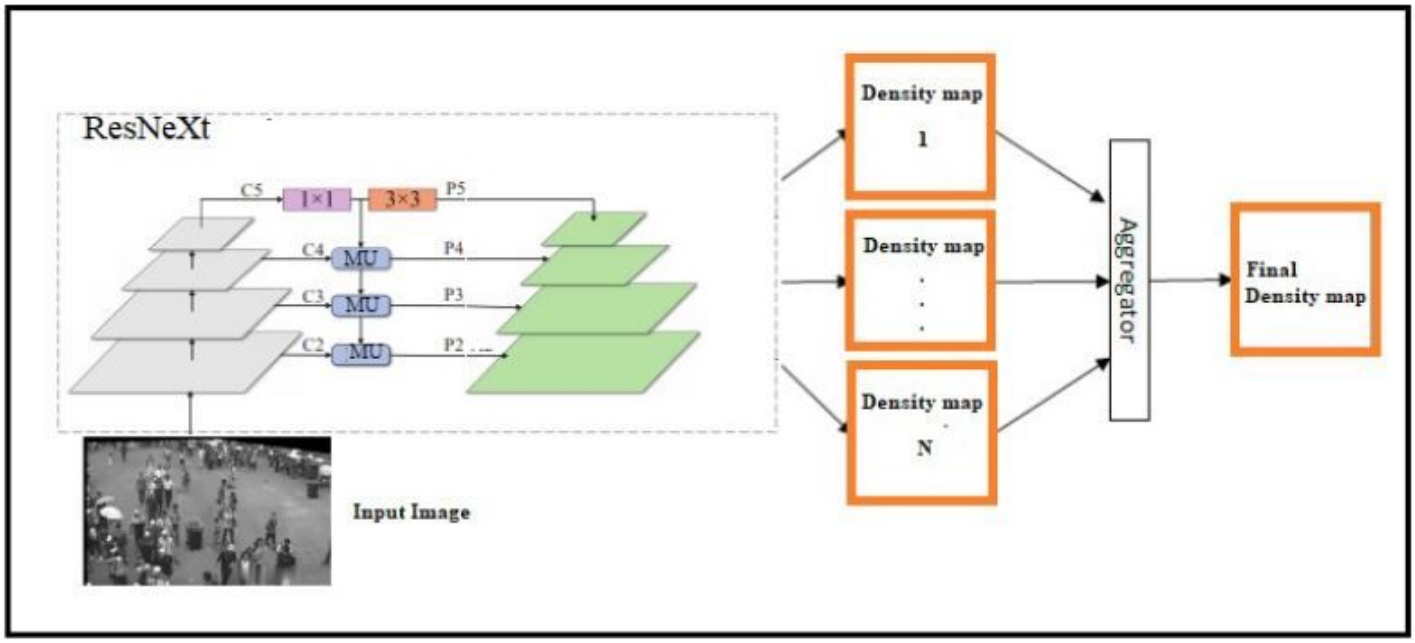

Figure 4

Proposed ResNextFPNet Architecture

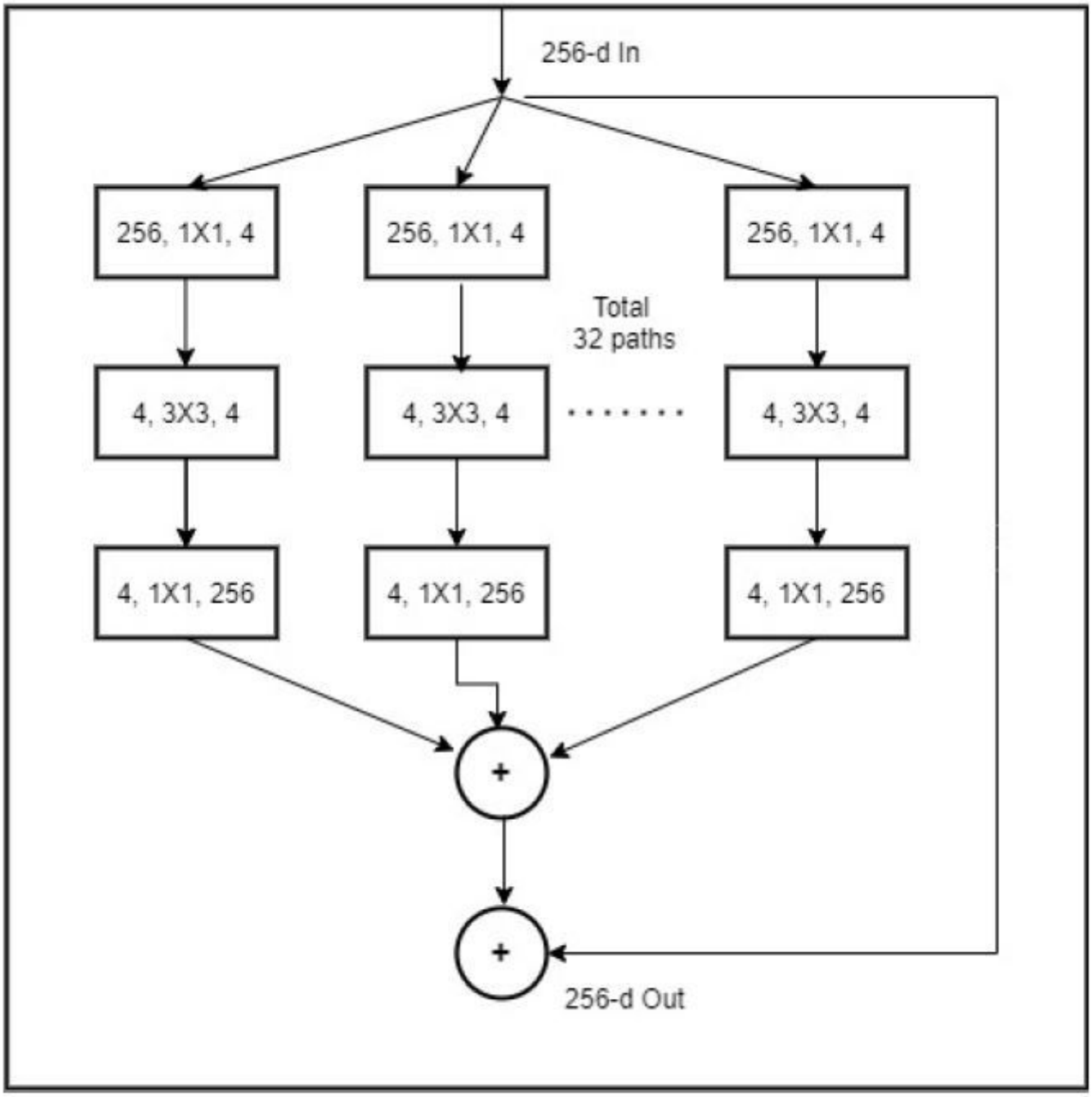

Figure 5

A single block of ResNeXt

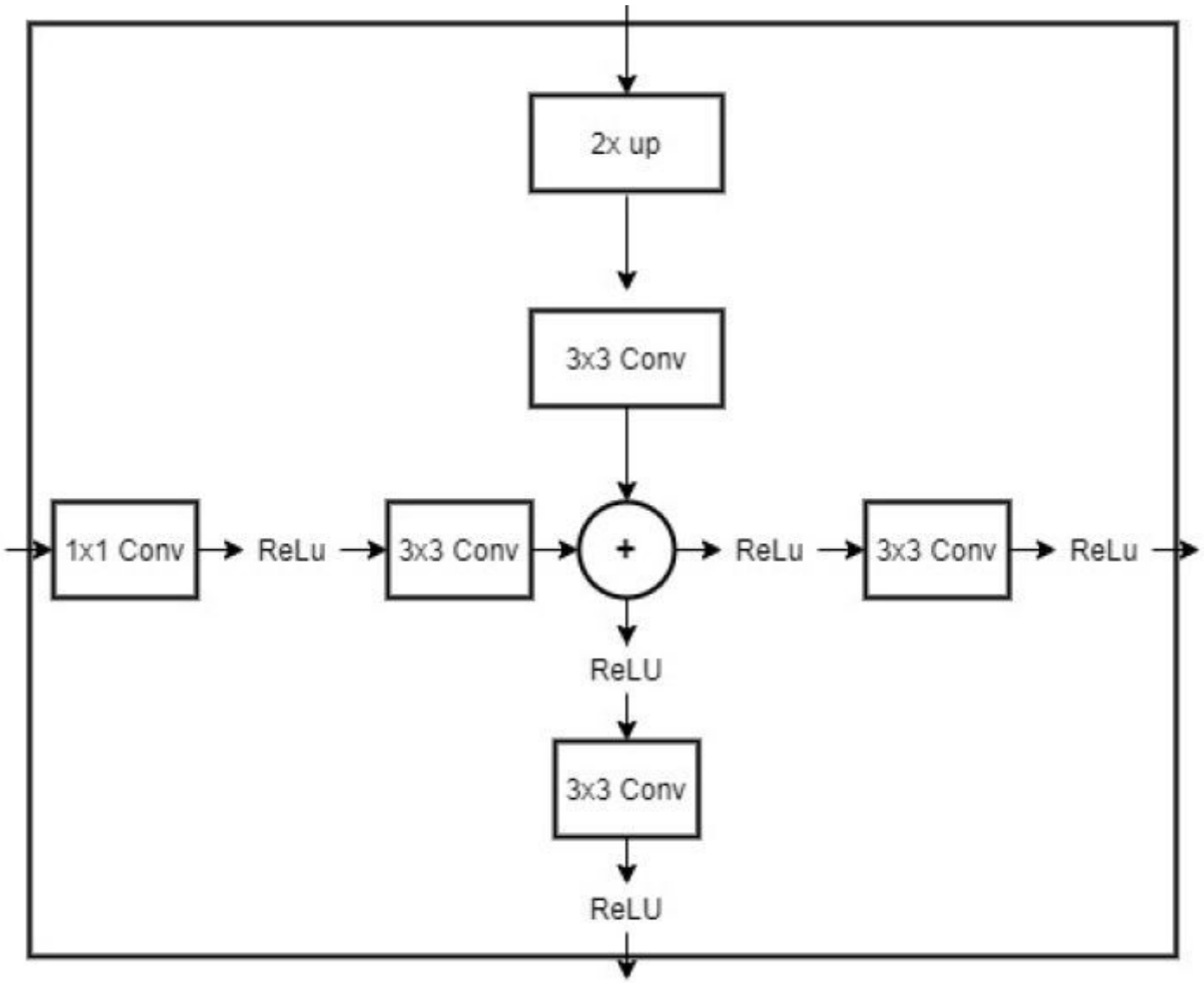

Figure 6

The Merging Unit of FPN