Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 6, No. 1, 2023

DFENet: Double Feature Enhanced Class Agnostic Counting Methods

Jiakang Liu, Hua Huo

College of Information Engineering, Henan University of Science and Technology, Luoyang 471000, China

Abstract: Object counting is a basic computer vision task, which can estimate the number of each object in an image, thus providing valuable information. In dense scenes, there are huge differences in target individual scale, and the different target individual scale leads to low accuracy of target count. In addition, most of the existing target count datasets in the field require a lot of manual creation and annotation, which increases the cost and difficulty of the dataset, lack of ease of use and portability. To solve these problems, this paper proposes a class agnostic counting method Double Feature Enhancement Net based on improved Bilinear Matching Network+ (BMNet+). By introducing the feature enhancement module based on the principle of conditional random field and the adaptively spatial feature fusion module, combined with the feature similarity measurement strategy of bilinear matching network, the method can effectively extract the target features of different scales, enhance the adaptability to the targets with large scale changes, and improve the counting performance of the network. Experiments were carried out on FSC-147 data set, and the experimental results show that the proposed model has been further improved in counting accuracy. The MAE and MSE of the verification set are 15.03 and 54.53 respectively. In the test set, MAE reaches 13.65, MSE reaches 89.54, and the counting performance is at the advanced level in the field.

Keywords: Object Count; Class Agnostic Counting; Conditional Random Field; Similarity Measure.

1. Introduction

Object counting is a fundamental and important computer vision task that provides quantitative estimation of targets in images, thus providing valuable information for various fields of social production. This task is important in many practical applications, including traffic management, public safety, medical diagnosis, and agricultural monitoring. For example, in the field of traffic management, Object counting for road traffic can monitor the congestion of various roads in a city, thus assisting in traffic management; in the field of public safety, Object counting for crowds can detect the number of pedestrians on the streets in real time, and detect safety hazards in advance in order to avoid potential hazards, such as stampede incidents. However, object counting datasets often require a lot of manual creation and annotation, which not only increases the cost and difficulty of the dataset, but also restricts the diversity and versatility of the dataset, so that traditional Object counting methods can often only be used for counting specific categories of targets, and lack of ease of use and transferability. To solve this problem, Lu et al [13] first proposed the concept of Class-Agnostic Counting (CAC), which aims to create a generalized counting model that can adapt to any class of targets without the need to know the target's class information in advance. Shi et al [15] designed a network based on similarity measures based on this foundation BMNet+, by comparing the feature similarity between the sample image and the query image in order to generate a target count density map. However, when confronted with multiple targets with large scale variations, this network often fails to perform the task of feature comparison of multi-scale targets well, resulting in inaccurate counting results. Therefore, in this paper, we propose a conditional random field-based feature extraction and enhancement method for image counting task. We extract two different levels of feature maps from the backbone network, which represent the low and medium level and high level features of the image, respectively. Then we design a dual feature enhancement module, which consists of a self-similarity module and a SFEM feature enhancement module, for enhancing the feature representation of the query image by utilizing the information of the query image itself and the information between the feature maps at different levels. Finally, we use ASFF adaptive spatial fusion module and dynamic similarity metric module for dynamically fusing and comparing the features of query images and sample images at different levels according to the correlation between them and get the final counting results to improve the counting performance.

Overall, the main contributions of this work are as follows:

1. A modified network structure DFENet based on a modified Bilinear Matching Network+ (BMNet+) is proposed, which extracts, enhances & fuses features of different scales to improve the counting performance of the network.
2. A Structured Feature Enhancement Module (SFEM) is introduced, which enhances the representation of features by enabling features of different scales to transfer information to each other through an algorithm based on the principle of conditional random fields.
3. Adaptively Spatial Feature Fusion Module (ASFF) is introduced, which can adaptively learn the spatial weights of feature maps at different scales, thus improving the robustness of the feature fusion network.

2. Related Work

2.1. Object Counting Related Work

Object counting is a computer vision task involving the estimation of the number of targets in an image, which has a wide range of applications in various fields of social production, such as traffic counting, crowd density estimation, biological population counting and so on. Currently, the work related to object counting can be categorized into two types according to the different counting methods: detection-based

counting methods [1][2][3][4], and regression-based counting methods [5][6][8][9].

Detection-based counting methods are methods that use techniques of target detection for counting, such as Histograms of Oriented Gradients (HOG) [10], Region-CNN (RCNN) [11], You Only Look Once (YOLO) [12], etc. The basic idea of this method is to first detect each target in the image and then count the number of targets in each category based on the detection results. The advantage of this method is that the position information and target scale information of different targets can be obtained while counting, so that the targets can be analyzed in more detail; moreover, this method is better adapted to the targets with multi-scale variations, and it can deal with targets of different sizes. However, the disadvantages of this type of method are that it has a large amount of computation and lacks real-time and convenience for the counting task; moreover, this method also tends to show poor counting performance when the target receives an occlusion, and is unable to accurately recognize the occluded target.

Regression-based counting method refers to the method that generally relies on some texture features in the image, and then uses linear regression, neural networks, and other ways to establish the mapping relationship between the features and the target, to directly obtain the statistical results of the number of targets. The advantage of the regression-based counting method is that it can directly output the number of targets without detecting the targets, so it tends to have a small amount of computation and high real-time, which is suitable for fast processing of counting tasks. However, the disadvantage is that it is less adaptable to targets of different scales, and cannot count well in the face of images with large target scales (e.g., dense crowds in perspective), because the target features in this case are often not obvious, and it is difficult to establish an effective mapping relationship.

2.2. Class Agnostic Counting Related Work

Most of the current object counting methods need to know the class information of the target, such as detection-based counting methods and regression-based counting methods. These methods have some limitations, such as requiring a large amount of labeled data, and difficulty in dealing with occlusion and overlapping. Therefore, some researchers have proposed the concept of Class-Agnostic Counting (CAC), which does not need to know the class of the target, but only how many targets are in the image. This method can be applied to a variety of scenarios, such as medical images, satellite images, UAV images, etc., with better generalization and robustness. Currently, there are several related works on class agnostic counting:

Lu et al [13] were the first to propose the concept of class agnostic counting, and they implemented class agnostic counting through a Generic Matching Network (GMN) model that was pre-trained on a labeled target tracking video dataset. The advantage of this model is that with less sample learning, it does not need to be trained with a large training set like most other convolutional neural network-based crowd density estimation methods, but only requires one pre-training session to count different objects. At the end of the paper, the authors discuss the feasibility of class agnostic counting in different domains such as cells, vehicles, crowds, etc. and show the excellent performance of the model on various datasets.

Ranjan et al [14] proposed a brand-new network architecture called Few-Shot Adaptation & Matching Network (FamNet) to solve the CAC problem. Meanwhile, the authors built an FSC-147 dataset. The dataset consists of 6135 images covering 147 different object categories ranging from kitchen utensils and office stationery to vehicles and animals.The FSC-147 dataset has a large variation in the number of objects and a large number of image categories, which are constructed specifically for the CAC problem. FamNet is an end-to-end network that consists of a feature extractor and a feature aggregator. The feature extractor uses ResNet-50 as the backbone network and outputs multi-scale feature maps at each stage. The feature aggregator uses a self-attention mechanism to fuse features from different scales and channels and outputs a global feature vector. This global feature vector is fed into a fully connected layer to predict the number of targets. The authors conducted experiments on the FSC-147 dataset and compared it with other methods, which showed that FamNet has better generalization ability and robustness.

Shi et al [15] focused on similarity modeling and constructed a generalized CAC similarity-aware framework, BMNet+, which jointly learns feature representation and similarity metrics in an end-to-end manner, and uses a composite loss function to monitor the counts and similarity metrics simultaneously. metrics are simultaneously supervised with a composite loss function, allowing it to achieve the best current performance in the CAC domain on the FSC147 dataset. Although this work proposes a dynamic similarity metric for multi-scale targets which makes the adaptability to multi-scale targets somewhat improved, there is still room for improvement. Therefore, in this paper, we propose a conditional random field-based feature enhancement method for improving the adaptability of similarity metric networks for multi-scale targets and improving the training performance of the networks.

2.3. Conditional Random Field Related Work

In the field of computer vision, Conditional Random Field (CRF) is a commonly used probabilistic graphical model that can be used to improve the features and outputs of Convolutional Neural Networks (CNNs) by using the message passing mechanism. CRFs can model the relationships between pixels or regions to improve the expressive power and prediction accuracy of CNNs. For example, Zheng et al [16] proposed a CRF-based approach that can refine the semantic segmentation map of a CNN by modeling the relationships between pixels for more accurate scene understanding. Xu et al [17] improved the performance of human pose estimation by fusing multiple features with attentional gating CRFs to produce a richer representation of contour prediction. Wang et al [18] introduced an inter-view message passing module based on CRFs to enhance image-specific features for action recognition. It can utilize the information from different viewpoints to enhance the features of each viewpoint, which improves the accuracy of action recognition. In crowd counting, Deep Structured Scale Integration Network (DSSINet) [19] for the first time utilizes CRFs to mutually refine multiple features at different scales and proves its effectiveness for this task. It improves the performance of crowd counting by utilizing the message passing mechanism of CRFs to balance features at different scales to generate finer density maps.

2.4. Metric Learning Related Work

The aim of metric learning is to embed data into a space where similar samples are brought closer together and different samples are pushed apart [22]. Similarity is measured in a fixed [23][24] or learned [25][26] manner. A common approach is to restrict the similarity between features in a pair [27] or trio [28]. This approach allows similar samples to be closer together in space and different samples to be more spread out in space. Another approach adds constraints based on the signal-to-noise ratio [29][30][31], where the similarity between pairs of positive samples is considered signal and the similarity between pairs of negative samples is considered noise. This approach allows the signal to be stronger and the noise to be weaker. BMNet+ reapplies this idea to CAC by pulling features closer between examples and target instances while pushing features away between examples and background patches. This approach allows for more similar features between examples and target instances, and more different features between examples and background patches.

3. Method

3.1. Network Structure

Aiming at the insufficient adaptability of BMNet+ to changes with multi-scale targets, this paper proposes an improved network model DFENet based on BMNet+, as shown in Fig. 1, which consists of six parts, namely, the backbone network, self-similarity module, SFEM, ASFF, and dynamic similarity metric and counting module. Our whole network structure is shown in Fig. 1.

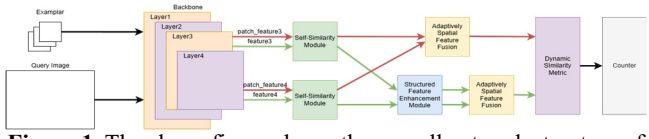

Figure 1. The above figure shows the overall network structure of this paper. The red line represents the processing route of the sample image, and the green line represents the processing route of the query image.

Backbone. The backbone network is the net structure that the model uses at the very beginning to extract the target features, and the design of the backbone network affects the results of target counting to some extent. The backbone network used in this work is the first 4 blocks of Resnet-50.

Self-Similarity Module. The Self-Similarity Module is one of the cores of the BMNet+ framework. Since the occurrence of the same type of targets is often accompanied by different scales and poses, this variation within the target group creates a great challenge for the similarity measure. BMNet+ proposes a Self-Similarity Module for enhancing the common features among each target. Specifically, the module first collects sample features $\mathrm{F}(Z)$ with each feature vector of the query feature image $F_{ij}(X)$, and put them into a feature set. Each vector in the feature set is then updated through a self-attentive mechanism, and the updated features are accompanied by a learnable γ parameter that is added to the original features. Finally, the resulting feature set is split to obtain the final $F(Z)$ and $F(X)$

SFEM. SFEM (Structured Feature Enhancement Module) was first proposed by Liu et al [19] in the network model DSSINet. Inspired by intensive forecasting work, the SFEM module mutually refines features at different scales by fully exploring their complementarity with Conditional Random Fields (CRFs) models. In this module, each scale-specific feature passes its own information to features from other scales. Specifically, it is assumed that the ensemble of features from each different sub-network is denoted as $F=\{f_1,f_2,...,f_n,\}$, our goal is to estimate a refined set of features $\hat{F}=\left\{\hat{f}_1,\hat{f}_2,...,\hat{f}_n,\right\}$,where $\hat{f}_i$ is the corresponding improvement feature of f_i. We formulate this in terms of the CRFs model. Specifically, define the conditional distributions of the original feature F and the improved feature $\hat{F}$ as:

$$P\left(\hat{F}\mid F,\Theta\right)=\frac{1}{Z(F)}\exp\left\{E\left(\hat{F},F,\Theta\right)\right\} \tag{1}$$

where $Z(\mathrm{F})=\int_{\hat{\mathrm{F}}}\exp\left\{E\left(\hat{F},F,\Theta\right)\right\}d\hat{F}$ is the partition function for normalization, and Θ is the set of parameters

The energy functions $\mathrm{E}\left(\hat{F},F,\Theta\right)$ in CRFs is defined as:

$$\mathrm{E}\left(\hat{F},F,\Theta\right)=\sum_{\mathrm{i}}\phi\left(\hat{f}_i,f_i\right)+\sum_{i,j}\psi\left(\hat{f}_i,\hat{f}_j\right) \tag{2}$$

where the unitary potential $\phi\left(\hat{f}_i,f_i\right)$ representing the similarity between the original and improved features is defined as:

$$\phi\left(\hat{f}_i,f_i\right)=-\frac{1}{2}\|\hat{f}_i-f_i\|^2 \tag{3}$$

We model the correlation between two improved features with a bilinear potential function, and thus define the pairwise potentials as:

$$\psi\left(\hat{f}_i,\hat{f}_j\right)=\left(\hat{f}_i\right)^T\omega_j^i\hat{f}_j \tag{4}$$

where ω_j^i is the learnable parameter used to calculate the correlation between $\hat{f}_{\mathrm{i}}$ and $\hat{f}_{\mathrm{j}}$.

This is a typical CRF formula that we solve using mean field extrapolation. feature $\hat{f}_i$ is calculated by the formula:

$$\hat{f}_i=f_i+\sum_{j\neq i}\omega_j^i\hat{f}_j \tag{5}$$

where the unary term is the feature f_i itself, and the second term denotes the information received from other features at different scales. The parameter ω_j^i determines the information content passed from f_j to f_i .Since $\hat{f}_{\mathrm{i}}$ and $\hat{f}_{\mathrm{j}}$ are

interdependent in $\hat{f}_i = f_i + \sum_{j \neq i} \omega_j^i \hat{f}_j$, we iteratively obtain each improvement feature using the following formula:

$$h_i^0 = f_i, h_i^t = f_i + \sum_{j \neq i} \omega_j^i h_j^{t-1}, t = 1...n, \hat{f}_i = h_i^n \quad (6)$$

Where n is the total number of iterations and h_i^t is the intermediate feature of the t^{th} iteration. The formula:

$$h_i^0 = f_i, h_i^t = f_i + \sum_{j \neq i} \omega_j^i h_j^{t-1}, t = 1...n, \hat{f}_i = h_i^n \quad (7)$$

can be easily implemented in the SFEM module. Specifically, the SFEM module uses a 1x1 convolutional layer to pass complementary information from f_j to f_i, ω_j^i is the learning parameter of the convolutional layer and is shared for all iterations.

ASFF. ASFF [20] is a module for fusing feature maps that have been enhanced by the SFEM module. Unlike previous methods of integrating multilevel features using element summation or concatenation, the key idea of ASFF (Adaptively Spatial Feature Fusion) is to adaptively learn the spatial weights of the fused feature maps at each scale. The process consists of two steps: constant scaling and adaptive fusion. The ASFF structure is shown in Fig. 2

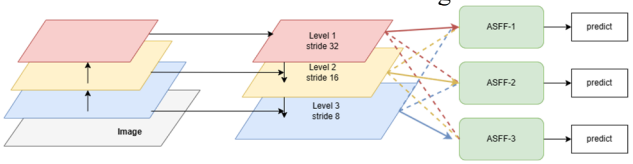

Figure 2. Illustration of the adaptively spatial feature fusion mechanism

Feature Resizing. We denote the features of the resolution at level l as x^l. For level l, we resize the features x^n at the other level $n(n \neq l)$ to the same shape as that of x^l. For up-sampling, we first apply a 1×1convolution layer to compress the number of channels of features to that in level l, and then interpolated to increase the resolution respectively. For 1/2-ratio down-sampling, a 3×3 convolution layer with a 2-step length is used, while modifying the number of channels and resolution. For 1/4 ratio down-sampling, a 2-step maximum pooling layer is added before the 2-step convolution.

Adaptive Fusion. Let $x_{ij}^{n \to l}$ denote the feature vector at the position (i, j) on the feature maps resized from level n to level l ,then the feature fusion corresponding to level l is as follows:

$$y_{ij}^l = \alpha_{ij}^l \cdot x_{ij}^{1 \to l} + \beta_{ij}^l \cdot x_{ij}^{2 \to l} + \gamma_{ij}^l \cdot x_{ij}^{3 \to l} \quad (8)$$

where y_{ij}^l implies the (i, j)-th vector of the output feature maps y^l among channels. α_{ij}^l, β_{ij}^l and γ_{ij}^l refer to the spatial importance weights for the feature maps at three different levels to level l, which are adaptively learned by the network. Note that α_{ij}^l, β_{ij}^l and γ_{ij}^l can be simple scalar variables, which are shared across all the channels. We forced $\alpha_{ij}^l + \beta_{ij}^l + \gamma_{ij}^l = 1$ and $\alpha_{ij}^l, \beta_{ij}^l, \gamma_{ij}^l \in [0,1]$, and define:

$$\alpha_{ij}^l = \frac{e^{\lambda_{\alpha_{ij}}^l}}{e^{\lambda_{\alpha_{ij}}^l} + e^{\lambda_{\beta_{ij}}^l} + e^{\lambda_{\gamma_{ij}}^l}} \quad (9)$$

Here the SoftMax function is used to define α_{ij}^l, β_{ij}^l and γ_{ij}^l with $\lambda_{\alpha_{ij}}^l$, $\lambda_{\beta_{ij}}^l$ and $\lambda_{\gamma_{ij}}^l$ each as a control parameter respectively. We use 1×1 convolutional layers to compute weight scalar mappings for λ_α^l, λ_β^l, and λ_γ^l from $x^{1 \to l}$, $x^{2 \to l}$, and $x^{3 \to l}$ respectively, such that they can be learned by standard backpropagation.

Dynamic Similarity Metrics. Given that humans may base their categorization judgments on the typical features of a category, for example, if someone describes an organism as having four limbs and hair-covered auditory organs, you might infer that it is a canine. Therefore, Shi et al [15] proposed a Dynamic Similarity Metric (DSM) that adaptively learns the category patterns of target samples without being restricted to specific categories. Specifically, this is an integrated feature selection module, similar to SENet [32], it learns the dynamic channel attention weights a conditioned on $Qz + b_z$ such that the similarity S can be computed by

$$S_{ij}(x, z) = \left[\left(Px_{ij} + b_x\right)\right]^T \left[a \circ \left(Qz + b_x\right)\right] \quad (10)$$

where $\circ$ denotes the Hadamard product.

Counter module. The Counter module consists of several convolutional and bilinear upsampling layers used to regress a density map of the same size as the query image. Specifically, the counting module receives the channel-level connections between query feature map $F(X)$ and similarity maps S, and then predicts a density map D_{pr}. The final count results in an integral of D_{pr}.

3.2. Experiment Details

Data Pre-processing. In this work, we adopt the same data preprocessing method as BMNet+ to ensure the consistency and comparability of the data. The data preprocessing method consists of two main steps: image scaling and image cropping. The purpose of image scaling is to resize the query image to a suitable range to accommodate image inputs with different resolutions. We use an aspect ratio-based scaling method, i.e., scaling the longest side of the query image to a preset interval, i.e., [384,1584], while keeping the original aspect ratio of the query image unchanged. This avoids the problem of distortion or deformation of the image during the scaling process. The purpose of image cropping is to unify the size of the sample images to 128 × 128 for subsequent feature extraction and similarity metrics. We use a center-based cropping method, i.e., starting from the center of the sample image, 128 pixels are cropped out horizontally and vertically along the horizontal and vertical directions, respectively, to obtain a square region as the final sample image. This preserves the most important parts of the sample image while reducing the interference of irrelevant information.

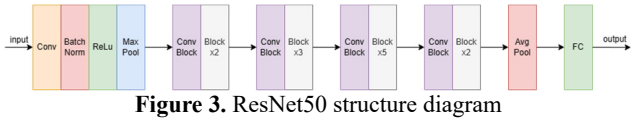

Figure 3. ResNet50 structure diagram

Network Design. To extract effective features from images,

we employ a deep residual network (ResNet-50) based backbone network as a feature extractor. The schematic diagram of ResNet50 structure is shown in Fig. 3. The backbone network consists of the first four residual blocks of ResNet-50, and each residual block contains multiple convolutional layers and jump connections. We extract two different levels of feature maps from the backbone network, which are the outputs of the third residual block (layer3) and the fourth residual block (layer4), and their channel counts are 1024 and 2048, respectively. These two feature maps represent the low and medium level and high level features of the image, respectively, which can capture different details and semantic information of the image. For each query image, we use a 1×1 convolutional layer to reduce the number of channels of the two feature maps to 256 to reduce the computation and memory consumption. For each sample image, we first perform a Global Average Pooling (GAP) operation on the two feature maps to compress each feature map into a single vector, and then use a fully-connected layer to stitch the two vectors together and map them into a 256-dimensional feature vector. In this way, we obtain two different forms of feature representations, i.e., feature maps and feature vectors, for each query image and sample image. In order to further enhance the feature representation of the query image, we design a dual feature enhancement module, which consists of a Self-Similarity Module and a Structured Feature Enhancement Module for SFEM. The role of the Self-Similarity Module is to enhance the feature map of the query image by utilizing the information of the query image itself, while the role of the SFEM Feature Enhancement Module is to further enhance the feature map of the query image by utilizing the information between the feature maps of the query images at different levels. Specifically, we send the query image feature maps and sample image feature vectors output from the same layer of the backbone network to the self-similarity module for the first feature enhancement to get the enhanced query image feature maps; then we send the query image feature maps from different layers processed by the self-similarity module to the SFEM feature enhancement module for the second feature enhancement to get the final query image feature maps. The SFEM feature enhancement module is based on the Conditional Random Field (CRF) method, which can effectively utilize the correlation between spatially adjacent pixels to optimize the feature value of each pixel point.

After obtaining the feature representations of the final query image and the sample image, we need to fuse and compare them to obtain a similarity measure between them. For this purpose, we use the ASFF Adaptively Spatial Fusion Module and Dynamic Similarity Metric Module, The role of the ASFF Adaptive Spatial Fusion Module is to dynamically fuse the query and sample images at different levels based on the correlation between them. The function of ASFF adaptive spatial fusion module is to dynamically fuse the features of query images and sample images at different levels according to their correlations, so as to obtain a more comprehensive and accurate fusion features. The role of the dynamic similarity metric module is to calculate the similarity between each pixel and the sample image according to the fused feature map, and dynamically adjust the similarity threshold according to the distribution of similarity, so as to obtain a more stable and robust similarity metric. Finally, we feed the output of the dynamic similarity metric module into a regression layer to get the final counting result.

Loss function. The loss function in this work consists of two parts, namely Similarity Loss and Count Loss. We will assume that the size of S is 1/r of X, i.e., each position in the similarity map corresponds to an r×r block in the query image. For each location in S, a positive label is assigned if its corresponding r×r block contains multiple targets, and a negative label is assigned if it does not contain targets. Then, we can derive the similarity loss based on the signal-to-noise ratio:

$$L_{sim} = -\log \frac{\sum_{i \in pos} \exp(S_i)}{\sum_{i \in pos} \exp(S_i) + \sum_{i \in neg} \exp(S_j)} \tag{11}$$

where pos and neg denote the positive labeling position and negative labeling position in S.

Regarding the count loss, we choose to use the traditional L2 loss as our count loss function:

$$L_{count} = \left\| D_{pr}(X,Z) - D_{gt}(X,Z) \right\|_2^2 \tag{12}$$

where D_{gt} denotes the Ground Truth density map.

Thus, our overall loss function is:

$$L = L_{count}(D_{pr}, D_{gt}) + \alpha \cdot L_{sim}(S) \tag{13}$$

where α is used to balance the values of the two loss functions.

4. Experimental Results & Analysis

4.1. Evaluation Metrics

This study provides an in-depth exploration and analysis of the target counting problem based on the FSC-147 dataset. This dataset was constructed by Ranjan et al [14] and contains high-resolution images of 147 different scenes covering a wide range of densities, illumination, occlusion and background complexity. In order to objectively assess the performance of the method proposed in this study, two widely used evaluation metrics are used, namely mean absolute error (MAE) and mean square error (MSE). MAE measures the average deviation between the predicted value and the true value, and MSE measures the average squared difference between the predicted value and the true value. These two metrics reflect the accuracy and stability of the target counting method.

$$\mathrm{MAE} = \frac{1}{m} \sum_{i=1}^{m} \left| \left(y_i - \hat{y}_i \right) \right| \tag{14}$$

$$\mathrm{MSE} = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - \hat{y}_i \right)^2 \tag{15}$$

4.2. Experiment Results

Table 1. Experimental results and comparison of FSC-147 dataset

Methods	Val MAE	Val MSE	Test MAE	Test MSE
GMN[4]	29.66	89.81	26.52	124.57
FamNet[4]	24.32	70.94	22.56	101.54
FamNet+[4]	23.75	69.07	22.08	99.54
CFOCNet[4]	21.19	61.41	22.10	112.71
BMNet[4]	19.06	67.95	16.71	103.31
BMNet+[4]	15.74	58.53	14.62	91.83
DEFNet (Ours)	**15.03**	**54.53**	**13.65**	**89.54**

Table 1 demonstrates the experimental results and comparisons of the counting method of this work on the FSC-

147 dataset. According to the indexes shown in the following table, the proposed method of this paper is compared with other related methods in the experiments on the FSC-147 dataset, and the results show that the method of this paper achieves the highest accuracy rate so far on this dataset, which proves the effectiveness and superiority of the method of this paper.

Figure 4 demonstrates the prediction effect of this paper's method on the FSC-147 dataset. It can be seen that the method of this paper can effectively deal with targets of different scales and predict the number of targets more accurately.

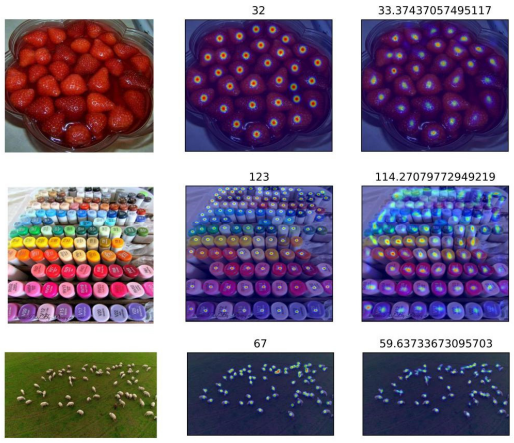

Figure 4. Prediction effect of the proposed method on FSC-147 data set

5. Conclusion

In this paper, we propose a class agnostic counting network structure, which is based on the BMNet+ framework and is improved for both feature enhancement and feature fusion. We introduce a feature enhancement module based on the conditional random field principle, which achieves further optimization of the target features by modeling the probability distribution of each pixel in the image. Secondly, we added an adaptive spatial feature fusion module, which realizes the enhancement of the feature extraction effect for targets of different scales through the dynamic weight allocation mechanism. Through experimental validation on the FSC-147 dataset, our method can effectively realize the class agnostic counting effect and reach the excellent level in the field in all evaluation indexes. There are still many shortcomings in this work, for example, the problem of increasing computation brought by adding modules has not been solved, which may affect the operation speed and efficiency of the network, and we will try to solve this problem in our future work.

References

[1] WU B, NEVATIA R.Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors[J].International Journal of Computer Vision, 2007, 75(2):247-266.

[2] LIN S F, CHEN J Y, CHAO H X.Estimation of number of people in crowded scenes using perspective transformation [J]. IEEE Transactions on Systems, Man, and Cybernetics, Part A:Systems and Humans, 2001, 31:645-654.

[3] MIN L, ZHANG Z, HUANG K, et al.Estimating the number of people in crowded scenes by MID based foreground segmentation and head-shoulder detection[C]// 2008 19th International Conference on Pattern Recognition, 2009.

[4] XU T, CHEN X, WEI G, et al.Crowd counting using accumulated HOG[C]// 2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery(ICNC-FSKD), 2016:1877-1881.

[5] CHAN A B, LIANG Z S J, VASCONCELOS N.Privacy preserving crowd monitoring:counting people without people models or tracking[C]// 2008 IEEE Conference on Computer Vision and Pattern Recognition, 2008:1-7.

[6] RYAN D, DENMAN S, FOOKES C, et al.Crowd counting using multiple local features[C]// 2009 Digital Image Computing:Techniques and Applications, 2009:81-88.

[7] KE C, CHEN C L, GONG S, et al.Feature mining for localised crowd counting[C]// British Machine Vision Conference, 2012.

[8] PARAGIOS N, RAMESH V.A MRF-based approach for real-time subway monitoring[C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[9] MCDONALD G C.Ridge regression[J].Wiley Interdisciplinary Reviews:Computational Statistics, 2009, 1(1):93-100.

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection[C]//2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.

[11] Girshick R , Donahue J, Darrell T ,et al.Rich feature hierarchies for accurate object detection and semantic segmentation tech report (v5)[J]. 2017.DOI:10. 1109/ cvpr. 2014.81.

[12] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[13] Erika Lu, Weidi Xie, Andrew Zisserman. Class-Agnostic Counting [C]//Asian Conference on Computer Vision (ACCV), 2018. arXiv:1811.00472.

[14] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, Minh Hoai. Learning To Count Everything[C]//Computer Vision and Pattern Recognition(CVPR), 2021. arXiv:2104.08391.

[15] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, Zhiguo Cao. Represent, Compare, and Learn: A Similarity-Aware Framework for Class-Agnostic Counting[C]// Computer Vision and Pattern Recognition(CVPR), 2022. arXiv: 2203. 08354.

[16] Zheng S, Jayasumana S, Romera-Paredes B, et al. Conditional random fields as recurrent neural networks[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1529-1537.

[17] Dan Xu, Wanli Ouyang, Xavier Alameda-Pineda, Elisa Ricci, Xiaogang Wang, and Nicu Sebe. Learning deep structured multi-scale features using attention-gated crfs for contour prediction. In NIPS, pages 3961–3970, 2017.

[18] Wang D, Ouyang W, Li W, et al. Dividing and aggregating network for multi-view action recognition[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 451-467.

[19] Liu L, Qiu Z, Li G, et al. Crowd counting with deep structured scale integration network[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1774-1783.

[20] Liu S, Huang D, Wang Y. Learning spatial fusion for single-shot object detection[J]. arXiv preprint arXiv:1911.09516, 2019.

[21] Yang S D, Su H T, Hsu W H, et al. Class-agnostic few-shot object counting[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2021: 870-878.

[22] Musgrave K, Belongie S, Lim S N. A metric learning reality check [C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. Springer International Publishing, 2020: 681-699.

[23] Atallah M J. Faster image template matching in the sum of the absolute value of differences measure[J]. IEEE Transactions on image processing, 2001, 10(4): 659-663.

[24] Lewis J P. Fast template matching[C]//Vision interface. 1995, 95(120123): 15-19.

[25] Sun Y, Cheng C, Zhang Y, et al. Circle loss: A unified perspective of pair similarity optimization[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6398-6407.

[26] Wang H, Wang Y, Zhou Z, et al. Cosface: Large margin cosine loss for deep face recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 5265-5274.

[27] Hadsell R, Chopra S, LeCun Y. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06). IEEE, 2006, 2: 1735-1742.

[28] Weinberger K Q, Saul L K. Distance metric learning for large margin nearest neighbor classification[J]. Journal of machine learning research, 2009, 10(2).

[29] Sohn K. Improved deep metric learning with multi-class n-pair loss objective[J]. Advances in neural information processing systems, 2016, 29.

[30] Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv: 1807. 03748, 2018.

[31] Yuan T, Deng W, Tang J, et al. Signal-to-noise ratio: A robust distance metric for deep metric learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 4815-4824.

[32] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.