ITC 3/53
Information Technology and Control
Vol. 53 / No. 3/ 2024
pp. 785-797
DOI 10.5755/j01.itc.53.3.35805

Cross-supervised Crowd Counting via Multi-scale Channel Attention

Received 2023/12/06 | Accepted after revision 2024/06/28

HOW TO CITE: Yan, K., Luan, F., Yuan, S., Liu, G. (2024). Cross-supervised Crowd Counting via Multi-scale Channel Attention. *Information Technology and Control,* 53(3), 785-797. https://doi.org/10.5755/j01.itc.53.3.35805

Cross-supervised Crowd Counting via Multi-scale Channel Attention

Kexin Yang, Fangjun Luan, Shuai Yuan, Guoqi Liu

Shenyang Jianzhu University, School of Computer Science and Engineering, Liaoning, China;
Liaoning Province Big Data Management and Analysis Laboratory of Urban Construction, Liaoning, China;
Shenyang Branch of National Special Computer Engineering Technology Research Center, Liaoning, China;
e-mails: 1121997899@qq.com; luanfangjun@sjzu.edu.cn; reidyuan@163.com; liuguoqi@sjzu.edu.cn

Corresponding author: reidyuan@163.com

Due to the challenges posed by large-scale variability in crowd images and overlapping and occlusion of people in high-density regions, traditional CNNs with fixed-size convolution kernels or transformers lacking 2D locality and channel adaptation need to struggle to cope with this challenge. While Transformers have a global receptive field for long sequence tasks, CNNs exhibit better generalization and 2D locality. In order to combine the advantages of both approaches, this paper proposes a dual-branch multi-scale attention network (DBMSA-Net). First of all, we propose a multi-scale channel attention convolution module to extract features at different scales while enhancing channel adaptation. Furtherly, local features are augmented using a feed-forward neural network that is more suitable for visual tasks. Then an efficient lightweight multi-scale regression head is employed to predict density maps. Finally, progressive cross-head supervision is introduced as a loss function to dynamically supervise instance labels noise and mitigate its effect. Extensive experiments are conducted on three crowd counting datasets (ShanghaiTech Part A, ShanghaiTech Part B, UCF-QNRF) to validate the effectiveness of the proposed method and the results show that DBMSA-Net outperforms state-of-the-art methods.

KEYWORDS: Crowd counting, Multi-scale, Channel attention, Transformer, Computer vision.

1. Introduction

The task of crowd counting is a popular research problem in computer vision. It requires fast and accurate crowd count estimation of images collected in different scenes with different crowd levels, which is important for congestion, urban planning and traffic management. Mainstream crowd counting methods [20, 28, 47] are mostly designed using CNN and then return to the density map to predict the number of people. Recently, Transformer [38] with global attention mechanism has achieved remarkable results in natural language

processing. After Dosovitskiy et al. [8] introduced the Transformer architecture into computer vision in the form of block partitioning, the Transformer shines in visual tasks and gradually surpasses traditional CNN models. Transformer began to be gradually applied to the study of crowd counting tasks [1,41] and achieved excellent performance.

With the development of vision transformer [9, 12, 32], it has been shown that local inductive bias and global self-attention mechanisms are equally significant for visual tasks. The information entering the visual task is two-dimensional, and transformer models using fixed-size attention fail to capture rich contextual information and multi-scale features. The importance of channel attention has been confirmed in previous studies [15], but the existing transformer [2, 21] does not pay sufficient attention to channel attention. These limitations in the visual transformer are definitely not sufficient to handle large-scale variations in crowd images. To address the above issues, we propose a convolutional encoder to try to complement the limitations of the vision transformer.

In view of transformer limitations in multi-scale feature extraction, we first propose a multi-scale convolutional channel attention module (MCAM). MCAM uses multi-scale convolutional branches to obtain multi-scale context information, and small-kernel strip convolution is used instead of traditional convolution to enhance local sensitivity and reduce model complexity. Furthermore, MCAM uses a deep separable convolution with an SE module as the channel attention branch to compensate for the adaptability of channel dimensions neglected by the transformer, and complements the mesh feature of the strip convolution branch.

To address the above issues, we introduce transformer and convolutional branches, which aim to fully exploit the contextual information and multi-scale features in the feature maps. The transformer branch uses a self-attention mechanism to enable the model to globally perceive critical objects and relations in the image. At the same time, the convolution branch uses multi-scale convolution operation to enhance the perception ability of the model to objects of different scales.

First, we propose a multi-scale channel attention convolutional module (MCAM) to extract multi-scale features and enhance channel adaptability. The module can adaptively adjust the receptive field at different scales and supplement the rich channel information with the channel attention mechanism. With this design, we are able to better deal with scale variations and occlusions in crowd images. To further improve the model performance, we also introduce a feed-forward neural network specifically designed for vision tasks, which is used to enforce local feature representations. It is able to improve the modeling capability of high-density regions and enhance the expressive power of the model for crowd density estimation.

Second, we employ efficient lightweight multi-scale regression heads to predict the density maps of both branches, which are able to provide fine-grained information about the distribution and density of the crowd. The regression head not only reduces the network complexity and computational overhead while maintaining high prediction accuracy, but also helps to improve the modeling ability of the model for density maps.

Finally, due to the noise problems such as location deviation and missing labeled points often appear in the dataset, as shown in Figure 1, in the training process, the progressive cross-head supervision [10] is introduced as the loss function to supervise each other through the ground truth density map. This mutual supervision mechanism enables the model to better learn accurate density map predictions, improves the performance and robustness of the whole network, and enhances the generalization capability of the network.

In summary, we use a crowd counting model with dual branches, called dual-branch multi-scale attention network (DBMSA-Net). The main contributions are further summarized as follows:

– We propose a multi-scale convolutional channel attention module (MCAM), using a feedforward neural network (FFN) layer that is more suitable

Figure 1

Noise annotations in the dataset. The red points are the labeled annotations in the dataset, the yellow points are the missing annotations, and the green points are the offset annotations

for visual tasks and MCAM to form a convolutional encoder to extract channel attention and multi-scale contextual features.
- We use a lightweight multi-scale regression head to further probe the global scale information of the population images and obtain more accurate regression density maps.
- We conduct extensive experiments on popular datasets such as ShanghaiTech Part A, Part B, and UCF-QNRF, and show that the proposed method makes progress in counting performance.

The rest of the sections are organized as follows. We first briefly introduce the work related to crowd counting in Section 2. DBMSA-Net is described in detail in Section 3, and experimental results are presented in Section 4. Finally, we give a conclusion in Section 5.

2. Related Work

2.1. Crowd Counting

Currently, methods for population counting can be classified into three main types: detection [22, 23], regression [3], and density mapping [24, 25, 42]. Detection-based methods predict bounding boxes for each person in an image by constructing a detection model for counting. However, its performance is limited by the occlusion of congested regions and the need for additional annotations. Regression-based methods can directly predict the number of people based on point annotations, but the results are poorly interpreted and the information of the labeled graphs is not fully exploited. Density map-based methods can better balance the performance and annotation cost than the other two methods.

To tackle the issue of noise in counting models, numerous studies have proposed the utilization of loss functions. For instance, Ma et al. [29] devised a loss function based on Bayesian theory for instance-level supervision, which directly employs point supervision to circumvent inaccurate generation of pseudo-maps. Cheng et al. [4] introduced the maximum excess pixel loss function, wherein optimization is performed using the region with the highest loss value. Other methods [26,30] use optimal transport for divergence measurements.

2.2. Attention

Attention mechanism is a mechanism analogous to human attention, which introduces a degree of attention or importance weight on specific information to a model to make it perform better on a particular task. Spatial attention focuses on the importance weights of the input feature maps in the spatial dimension, while channel attention focuses on the importance weights of the input feature maps in the channel dimension.

Transformer proposes a self-attention mechanism and is widely used in the computer vision community. VIT [8] applies the transformer architecture to models for visual tasks and has shown excellent performance on various visual tasks. DETR [2] further improves the efficiency of the vision transformer focused on object detection. Recently, these advances have promoted the effective application of transformer in various tasks such as semantic segmentation [45, 48] and object detection [21].

However, due to the shortcomings of both convolution and self-attention, Guo et al. [13] proposed a large-kernel attention module that considers and combines the advantages of convolution and self-attention, and further proves the importance of channel dimension in visual tasks.

2.3. Multi-scale

Crowd counting tasks often suffer from issues such as image scale variations and population density variations. These issues will have a greater impact on the performance of the model. Multi-scale network [37, 39] is one of the main methods to solve such problems. The multi-scale module can help the model to better handle issues such as scale variation and population density variation, and improve the performance and generalization capabilities of the model.

Szegedy et al. [34] proposed the Inception module, which performs multi-scale feature extraction by combining convolution and pooling operations at different scales. The Inception module can process information between different scales, improving the ability to express features while reducing the amount of parameters and computations. Chen et al. [5] proposed adaptive dilated convolution and pyramid pooling models to solve the scaling problem. Adaptive dilated convolutions can process the feature informa-

tion of the target at different scales and keep the input and output sizes constant on the feature map. Pyramid pooling models can extract features from regions of different sizes and merge features to improve feature expression.

3. Method

In this section, we will elaborate on the dual-branch multi-scale attention network (DBMSA-Net) in detail, which is mainly composed of VGG-16, transformer encoder, convolution encoder and multi-scale regression head. The training part uses progressive cross-head supervision.

3.1. Framework Method

Figure 2 shows an overview of DBMSA-Net. First, we use the VGG-16 [35] network as the backbone network to extract the features of a population image I, and the extracted features are $Feature \in R^{C\times W\times H}$, where C, H and W are the number of channels, width and height, respectively. The feature maps D_{Trans} are D_{Conv} then passed into the transformer and convolutional branches for feature learning, respectively, and the multi-scale regression head is used to predict the density maps and for both branches. The training process uses a progressive cross loss so that the ground-truth density map D_{GT}, D_{Trans}, D_{Conv} and are supervised against each other to constrain the training of the entire network.

3.2. Transformer Encoder

The traditional transformer networks [8, 38] use a self-attention layer in the encoder, which gives it a global receptive field and the ability to obtain global relationships of current features. It is calculated as follows:

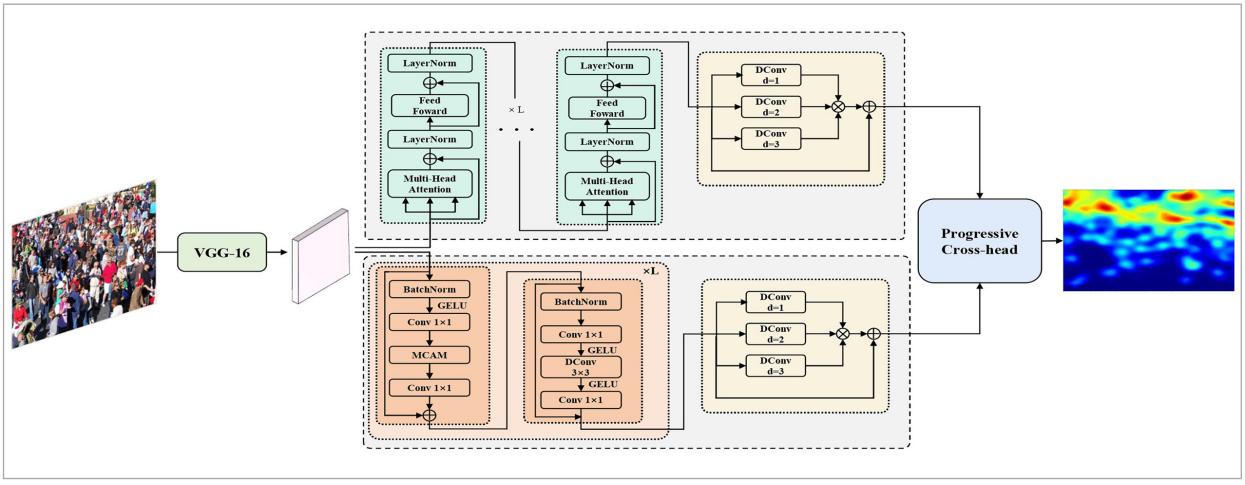
$$Attention(Q,\ K,\ V) = softmax(\frac{QK^T}{\sqrt{d_k}})V, \tag{1}$$

where $1/\sqrt{d_k}$ is a scaling factor based on the vector dimension d_k. Q, K, V, which are derived from source features, stand for the query, key, and value vectors, respectively.

Transformer leverages a self-attention mechanism to capture long-range dependencies in time or space, learn global contextual information, and flexibly process different features, thus achieving superior performance in crowd counting tasks. However, it suffers from certain shortcomings in processing information in densely populated areas, namely the lack of channel attention and 2D locality. In crowd counting tasks, due to the high crowd density, there are a large number of overlapping and occluded regions in the image, so it is

Figure 2

The framework of the dual-branch multi-scale attention network (DBMSA-Net). Firstly, the crowd images are fed to VGG-16, then the output feature maps are passed to the transformer encoder and convolutional encoder, respectively, and finally the multi-scale regression head prediction density map is passed in. The progressive cross-head supervises the predicted density map during training. DConv denotes the void convolution and d denotes the void rate



particularly important to explicitly model the importance of each channel and process overlapping people.

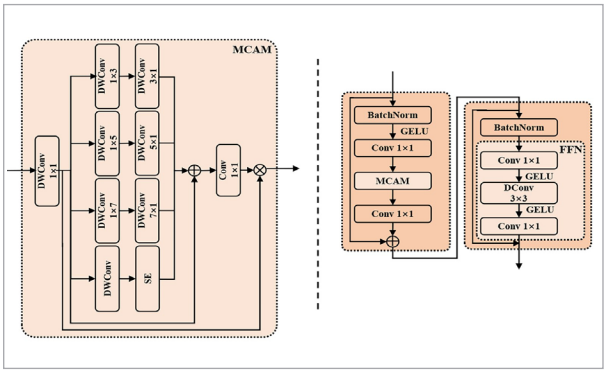
Therefore, we propose to adopt a dual-branch approach, add convolutional neural networks to process crowd images, and characterize and model each spatial location in more detail, thereby improving the ability to express features.

3.3. Convolutional Encoder

For image-intensive scenarios such as high-density crowd counting, although the Transformer model has superior performance for processing sequence modeling tasks, it has certain limitations due to its weak ability to model local features in images. Therefore, in the task of high-density population images, introducing appropriate local models or modeling methods based on local features can further improve the prediction accuracy and efficiency of the models. In this paper, we design a convolutional encoder. The structure is shown in Figure 3 and it forms a dual-branch structure with the transformer to compensate for the shortcomings of the transformer in local modeling.

Figure 3

Schematic diagram of MCAM and convolutional encoder, where DWConv is the depth convolution and SE is the channel attention module



We first use a 1×1 depth convolution to combine local characteristic information. It is then fed into a multi-scale branch to capture multiple context information, while a channel attention branch is added to extract channel features and supplement the grid information. Then use 1×1 convolution to model the relationship between multiple branches. Finally, the output is weighted with the input to obtain the final output result.

Our multi-scale branch uses three convolutional kernels of different sizes to extract image features individually. This structure can effectively cope with multi-scale variations and capture head information of different sizes. We adopt a structure similar to SegNext [14], so that the method of eliciting spatial attention by multiplying elements can obtain more efficient spatial information encoding than ordinary convolution and self-attention mechanisms. But the difference is that instead of continuing with the large kernel convolution formalism, we use small kernel convolutions. First, a small convolutional kernel can reduce the number of parameters, reduce the complexity of the model, make the model lighter, accelerate the training and inference of the model, and can be applied even when computational resources are limited. Second, in cases where the transformer branch provides a global receptive field, small convolutional kernels can capture more detailed feature information, thereby improving the sensitivity and modeling ability of neural networks to local images and better extracting feature information in dense population images. In the experiments presented later in this paper, it is demonstrated that using small kernel convolution in this model leads to better results than large kernel convolution. In later experiments, we demonstrate that using small kernel convolution in this model leads to better results than large kernel convolution.

In the crowd counting task, there are a large number of overlapping people and irregular shapes in the images, which makes it difficult to obtain accurate features using ordinary convolutions. Therefore, in some studies [16, 33], it is proposed to use strip convolution that emphasizes spatial positional relationships for feature extraction, and at the same time, this method can further lighten the network structure. We replace the original 2D convolution with a strip convolution, and replace the original convolution kernel of $k \times k$ with a pair of convolution kernels of $k \times 1$ and $1 \times k$. On the basis of the above model, we have added a deeply separable convolution branch with a channel attention SE module [17]. Such a choice not only compensates for the lack of channel adaptation in the transformer branch, but also enables our model to better process channel features and enhance the expressive power. It can also help to extract basic grid information for the model reference during feature

extraction. MCAM can be expressed mathematically as follows:

$$F = DWConv_{1\times1}(Feature), \tag{2}$$

$$Attention = Conv_{1\times1}(F + DWConv_{SE}(F) + \sum_{i=0}^{2} Scale_i(DWConv(F))), \tag{3}$$

$$Output = Attention \otimes Feature. \tag{4}$$

Feed-forward neural networks also play an important role in crowd counting models. They can adaptively learn the relationship between features, enhance features, and effectively prevent overfitting. However, the FFN structure in the NLP field lacks effective learning of two-dimensional local features, but this is essential for visual tasks [11,13]. To better meet the needs of crowd counting tasks, this paper uses an FFN structure that can better extract two-dimensional locality. The construction is shown in Figure 3.

3.4. Multi-scale Regression Head

In the population counting task, the regression head is considered as a crucial component. Its task is to predict the number of people in an image and output a predicted density map. We have extracted sufficient global and local context information in the previous stage, therefore we use a lightweight multi-scale regression head to regress the predicted density map.

As shown in Figure 2, we used three dilated convolutions with different dilation rates and convolution nuclei [24]. For a dilated convolution with a convolution nucleus size of $k\times k$ and an expansion rate of r, the receptive field size of each pixel is $(k-1)\times r+1$. Compared with traditional convolution methods, the use of dilated convolution can reduce the amount of parameters in the model, and make the model lighter while keeping the size of the receptive field unchanged. In addition, dilated convolutions can also enlarge the effective receptive area without changing the size of the receptive field, which can better accommodate the needs of large-scale and high-resolution image tasks.

There are heads of different scales in the images for the crowd counting task, so we chose a multi-scale form with different receptive fields. And since some dense regions have relatively small multi-scale, the multi-scale branch used in this paper tries to choose a small convolution kernel and dilation rate to avoid losing detailed information. We spliced the feature map output from each branch $B_i, i\in(0,3)$, and finally used two 1×1 convolutional layers to predict the regression of the density map.

3.5. Progressive Cross-head Supervision

Images collected in crowd counting tasks are typically annotated based on points. Since such manually labeled annotated points make up a relatively small fraction of the human head, there are spatial errors where the annotated points are not located at the exact location. At the same time, the lack of point annotations is also a non-negligible error when there is high density, occlusion, etc.

During supervision, noisy images are often used as a reference for training, but sometimes the labels predicted by the trained model are more correct than the manually labeled ones. Considering this situation, we use a progressive cross-head supervision method mentioned in CHS-Net [10] as the loss function during training. The construction is shown in Figure 4.

In progressive cross-head supervision, each branch is supervised using the weights of the predicted density map and the ground truth density map of the other branch, and the weights of the complementary branches are gradually increased as the training progresses. Taking the volume integral branch as an example here, the density map $\bar{D}_{Conv}$ obtained after supervision is defined as follows:

$$\bar{D}_{Conv} = \alpha F_{Trans}(X) + (1-\alpha)D_{GT}, \tag{5}$$

where α is the weight coefficient assigned to the complementary branch, D_{GT} is the ground truth density map, and F_{Trans} is the feature map output by the transformer encoder. To deal with the noise of the ground truth density map, mask M_{Conv} is used to supervise the density map $\bar{D}_{Conv}$. The formulation is stated as follows:

$$D_{Conv} = M_{Conv} \odot \bar{D}_{Conv} + (1-M_{Conv}) \odot D_{GT}, \tag{6}$$

$$M_{Conv} = \begin{cases} 0, \varepsilon \geq t_\delta \\ 1, \varepsilon < t_\delta \end{cases}, \tag{7}$$

where $\varepsilon = \left|F_{Conv}(X) - D_{GT}\right| \in R^{W\times H}$, t_δ is the mask threshold value after the error ε is arranged in descending order, and $\delta \in [0,1]$. It follows that the final

expression of the density map generated by the supervised volume integral branch is given by:

$$D_{Conv} = \alpha M_{Conv} \odot F_{Trans}(X) + (1 - \alpha M_{Conv}) \odot D_{GT}. \quad (8)$$

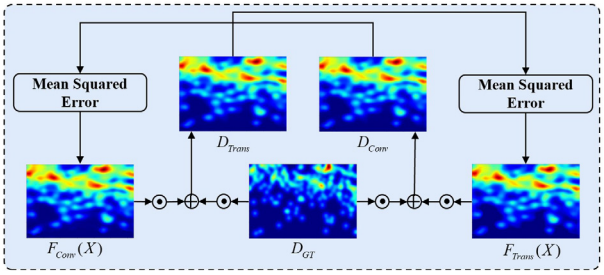
The density map D_{Trans} obtained by the supervised Transformer branch is similar to the one obtained by the convolutional branch, that is, the predicted density map $F_{Conv}(X)$ of the convolutional branch and the ground truth density map D_{GT} are used to supervise it. Thus, the overall optimized loss function of the model is formulated as follows:

$$Loss = \left\| F_{Conv}(X) - D_{Conv} \right\|_2^2 + \left\| F_{Trans}(X) - D_{Trans} \right\|_2^2, \quad (9)$$

Since the density map predicted by the model in the early stage is not reliable and the accuracy is low, a progressive supervision method is introduced, where the noise ratio δ_i and the coefficient α_i are gradually increased to the maximum as the training round i. It is formulated as follows:

$$\delta_i = \delta_{max} \frac{i}{T}, \alpha_i = \alpha_{max} \frac{i}{T}. \quad (10)$$

Figure 4
Overview of the progressive cross-head supervision



4. Experiments

4.1. Implement Details

Network Structure: VVG-16 [35] is used as the backbone network for feature extraction, and its pre-trained weights on ImageNet are used. We refer to the encoder part in [8], as the structure of our transformer branch, and use our proposed convolutional encoder to form a dual branch structure with it, and finally pass in the multi-scale regression head to predict the density map.

Training Details: We chose the same data preprocessing method as BL [29]. Each training image is randomly scaled, cropped, and horizontally flipped to enhance the data, and then a 512×512-sized image block is cropped. Using Adam's algorithm with a learning rate of 10-5 for parameter optimization and a cosine learning rate dispatcher for hyperparameters supervised by progressive crossheads, the hyperparameters are consistent with CHS-Net.

4.2. Datasets and Evaluation Metrics

The evaluation experiments were performed on three population counting datasets: ShanghaiTech Part A [47], ShanghaiTech Part B [47], and UCF-QNRF [19]. These data are described as follows:

ShanghaiTech Part A [47] dataset contains 482 randomly captured images of people from the Internet, including multiple scenes and different density levels, with a total of 244, 167 tokens, of which 300 images are used for training and the remaining 182 images are used as a test set.

ShanghaiTech Part B [47] dataset, compared to Part A, has a relatively small population density and contains 85,998 labeled points, 400 training images, and 316 test images. The images were taken on the streets of Shanghai.

UCF-QNRF [19] dataset contains 1535 high-resolution images with an average resolution of approximately 2013 x 2902, of which 1201 training and 334 test sets contain 1,251, 642 annotated points. It is very diverse in terms of adaptability, image resolution, crowd density, and scenarios where crowds are present.

Evaluation Metrics: We evaluate the count level of the model using two commonly used metrics, MAE and MSE. In particular, MAE is a better measure of the accuracy of the model count and MSE is a better measure of the robustness of the model. The lower the value of both, the better the model performance [47]. MAE and MSE are defined as follows:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| C_i^{GT} - C_i \right|, \quad (11)$$

$$MSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(C_i^{GT} - C_i \right)^2}, \quad (12)$$

where N is the number of sample images, C_i^{GT} and C_i are the number of heads of the predicted image and the ground truth map.

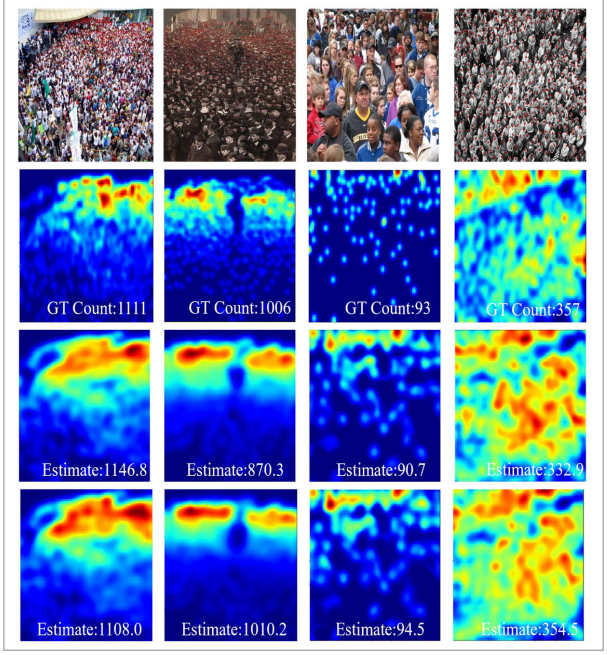
Table 1

Comparison of results on ShanghaiTech Part A, ShanghaiTech Part B and UCF-QNRF datasets. The best performer is indicated in bold, and the next best is indicated with a dash. The CHS-Net* results are obtained using the same conditions as DBMSA-Net

Method \ Dataset	Venue	SHT_Part A		SHT_Part B		UCF_QNRF	
		MAE	MSE	MAE	MSE	MAE	MSE
CSRNet [24]	CVPR 18	68.2	115.0	10.6	16.0	-	-
CAN [27]	CVPR 19	62.3	100.0	7.8	12.2	107.0	183.0
BL [29]	ICCV 19	62.8	101.8	7.7	12.7	88.7	154.8
DSSINet [25]	ICCV 19	60.6	96.0	6.9	10.3	99.1	159.2
LSC-CNN [36]	TPAMI 20	66.4	117.0	8.1	12.7	120.5	218.2
NoisyCC [43]	NIPS 21	61.9	99.6	7.4	11.3	85.8	150.6
GL [44]	CVPR 21	61.3	**95.4**	7.3	11.7	84.3	147.5
MCC [50]	ICASSP 22	71.4	110.4	9.6	15.0	-	-
GauNet(CSRNet) [7]	CVPR 22	61.2	97.8	7.6	12.7	84.2	152.4
CHS-Net [10]	ICASSP 23	59.2	97.8	7.1	12.1	**83.4**	144.9
CHS-Net*	ICASSP 23	61.9	104.0	7.6	11.7	85.8	147.9
DBMSA-Net		**58.8**	102.6	**6.8**	**11.3**	85.2	**144.1**

Figure 5

The visualization showcases ShanghaiTech Part A dataset, where the first row displays the input image, the second row presents the ground truth density map, followed by the third row illustrating the predicted density map of CHS-Net, and finally, in the fourth row lies our model's pre-density map. Warmer colors indicate higher crowd density



4.3. Comparison with Existing Methods

DBMSA-Net is evaluated on the above three datasets, and eleven recent methods are listed for comparison. The results are shown in Table 1. DBMSA-Net shows superior counting performance on different datasets. Our method achieves 58.8 MAE on the SHA dataset and 6.8 and 11.3 MAE and MSE on the SHB dataset. Compared to other approaches, the model in this paper achieves better accuracy and robustness. A visualization of the model is shown in Figure 5.

4.4. Ablation Studies

Ablation experiment on the model: The results from Table 2 demonstrate that employing a dual-branch structure yields superior outcomes compared to utilizing a single branch alone. Furthermore, when employing progressive cross loss for supervision, the Mean Absolute Error (MAE) attains optimal performance. The efficacy of this model can be further substantiated through result evaluation.

We perform ablation experiments on different branches of the convolutional encoder in ShanghaiTech Part A, and the results are shown in Table 3. The contribution of the strip convolution and the deep separable convolution with attention module in this

Table 2
Ablation experiments of the model

	Progressive Cross Loss	MAE	MSE
Conv.	×	62.8	104.7
Tran.	×	63.1	105.6
DBMSA-Net	×	59.6	106.1
DBMSA-Net	√	58.8	102.6

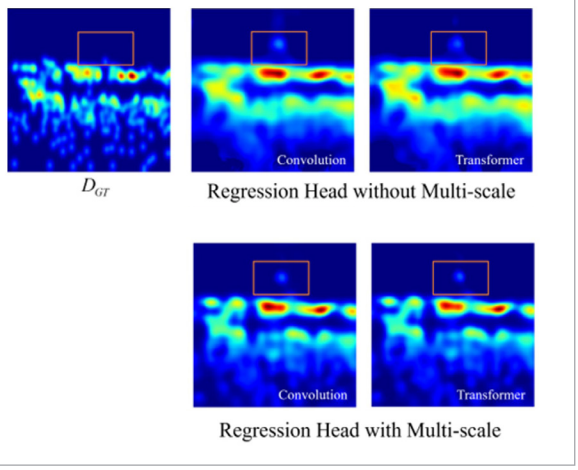
model were tested and the MAE was reduced by 4.6% and 5.9%, respectively. Subsequently, ablation experiments were further designed to demonstrate that convolution kernels of different sizes in the multi-scale branch are effective for extracting multiple populations in this model. We replace the original attention branch in the model with ordinary deep separable convolution and SE channel attention module, which has degraded the performance to varying degrees. It can be seen that the addition of mesh features and channel attention are important for the performance improvement of the model.

Table 3
Ablation experiments on ShanghaiTech Part A. In Table, S_k and D_k denote the strip convolution branch or hollow convolution with convolution kernel k, respectively, and SE-DWC is the deep separable convolution with SE module

S_3	S_5	S_7	SE-DWC	MAE	MSE
√	√	√		62.5	107.1
D_3	D_5	D_7	√	61.6	104.5
	√	√	√	61.2	110.7
√		√	√	60.8	108.2
√	√		√	59.6	106.0
√	√	√	DWC	61.9	112.7
√	√	√	SE	61.4	107.3
√	√	√	√	58.8	102.6

We performed a visual analysis of the proposed density map prediction using a lightweight multi-scale regression head, which is visualized in Figure 6. As can be seen from the visual result plots, after adding the multi-scale regression head, the quality of the model for generating density maps is further improved and the detailed information of image features can be better presented. Meanwhile, the visualization results demonstrate the effectiveness of progressive cross-head supervision for instance noise region supervision.

Figure 6
Visualization of whether to use a lightweight multi-scale regression head. The left plot shows the original image and the ground truth density map. The first behavior on the right uses the visualization of a tradition-al regression head, and the second behavior uses the visualization of a lightweight multi-scale regression head. The square boxes mark the regions where the points of the original image have noise



Comparison of training costs: Table 4 shows the performance comparison of the models using the big kernel convolution and the conventional dilated convolution. Here are all calculated based on the input 512×512 image. As can be observed from the data in Table, better accuracy can be achieved with the use of small-kernel strip convolutions in population counting. The MAE is reduced by 2.5 per cent compared to the use of large kernel-strip convolution, by 4.6 per cent compared to the use of ordinary dilated convolution, and the use of small-kernel strip convolution reduces the number of model parameters. It can be seen that this can make the model lighter, better meet the

Table 4
Comparison of MAE, MSE and model parameter quantity

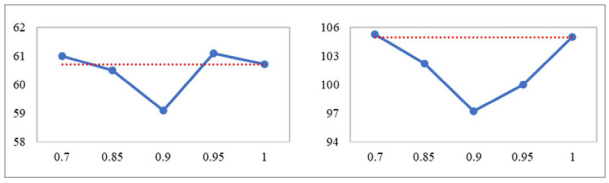
	MAE	MSE	Params (M)
$D_3 + D_5 + D_7$	61.6	104.5	143.9
$S_7 + S_{11} + S_{21}$	60.3	104.7	57.1
Ours	**58.8**	**102.6**	**56.9**

requirements of re source constraints, and improve the usability of the algorithm.

Effect of θ: We compare the MAE and MSE computed at different thresholds θ on the ShanghaiTech Part A dataset and the results are shown in Figure 7, where $\theta = 1 - \delta_{\max}$. We take $\theta = 1$, the case where no marked points are considered as noise, as the baseline. We observe that when $0.8 \leqslant \theta \quad 0.9$, the performance of the model gradually decreases as θ decreases, which may be due to the treatment of too many labeled points as noise, resulting in insufficient available training data. When the choice of θ is too large, the performance of the model is also degraded due to the increased noise in the selected training data. According to the experimental observation, the model performance reaches its best when $\theta = 0.9$, that is, the labeled data with the largest deviation from the prediction of 10 percent is considered to be masked by the noisy data, thereby reducing the effect of human noise on the model performance.

Figure 7

The effect of threshold θ on model performance. The dashed line indicates that the marked points are not considered as noise



5. Conclusion

In this study, we successfully combined the dual-branch structure of Transformer and convolution to enhance the performance of the population counting model. By incorporating a multi-scale channel attention module, we addressed the limitations of the Transformer structure regarding 2D locality and channel adaptation. Furthermore, we adopted a lightweight multi-scale regression approach to improve feature capture and regression accuracy. In order to mitigate the impact of label noise on model training, we designed progressive cross-head supervision as a loss function to supervise the generated density map and the ground truth density map. The experimental results on all three datasets illustrate the excellent performance of DBMSA-Net.

Our work investigates the feasibility of combining transformer and convolution as two distinct structures according to their own strengths and weaknesses in visual tasks. Our findings provide meaningful suggestions and insights in this research direction.

Acknowledgement

This work was supported by National Natural Science Foundation of China (62073227) and Liaoning Provincial Science and Technology Department Foundation (2023JH2/101300212).

References

1. Chen, P., Gao, J., Yuan, Y., Wang, Q. MAFNet: A Multi-Attention Fusion Network for RGB-T Crowd Counting. arXiv preprint arXiv:2208.06761, 2022.
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Ki-rillov, A., Zagoruyko, S. End-to-end Object Detection with Transformers. Proceedings of the 16th European Conference on Computer Vision (ECCV), August 23-28, 2020, 213-229. https://doi.org/10.1007/978-3-030-58452-8_13
3. Chan, A. B., Vasconcelos, N. Bayesian Poisson Re-gression for Crowd Counting. Proceedings of the IEEE/CVF International Conference on Computer Vi-sion (ICCV), Kyoto, Japan, September 29-October 02, 2009, 545-551. https://doi.org/10.1109/ICCV.2009.5459191
4. Cheng, Z. Q., Li, J. X., Dai, Q., Wu, X., Hauptmann, A. G. Learning Spatial Awareness to Improve Crowd Counting. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Ko-rea (South), October 27-November 02, 2019, 6152-6161. https://doi.org/10.1109/ICCV.2019.00625
5. Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. Pro-ceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, September 8-14, 2018, 833-851. https://doi.org/10.1007/978-3-030-01234-2_49

6. Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, R. Masked-attention Mask Transformer for Universal Image Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, June 18-24, 2022, 1290-1299. https://doi.org/10.1109/CVPR52688.2022.00135
7. Cheng, Z. Q., Dai, Q., Li, H., Song, J., Wu, X., Hauptmann, A. G. Rethinking Spatial Invariance of Convolutional Networks for Object Counting. Pro-ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orle-ans, USA, June 18-24, 2022, 19638-19648. https://doi.org/10.1109/CVPR52688.2022.01902
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissen-born, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszoreit, J., Houlsby, N. An Image is Worth 16x16 Words: Trans-formers for Image Recognition at Scale, International Conference on Learning Representations (ICLR), May 3-7, 2021. https://iclr.cc/virtual/2021/oral/3458
9. Dai, Z., Liu, H., Le, Q. V., Tan, M. CoAtNet: Marry-ing Convolution and Attention for All Data Sizes. Advances in Neural Information Processing Systems, 2021, 34, 3965-3977.
10. Dai, M., Huang, Z., Gao, J., Shan, H., Zhang, J. Cross-Head Super-vision for Crowd Counting with Noisy Annotations. ICASSP 2023-2023 IEEE Inter-national Conference on Acoustics, Speech and Sig-nal Processing (ICASSP), Rhodes Island, Greece, June 04-10, 2023, 1-5. https://doi.org/10.1109/ICASSP49357.2023.10095636
11. Fu, L., Tian, H., Zhai, X. B., Gao, P., Peng, X. In-cepFormer: Efficient Inception Transformer with Pyramid Pooling for Semantic Segmentation. arXiv preprint arXiv:2212.03035, 2022.
12. Guo, J., Han, K., Wu, H., Gao, P., Peng, X. CMT: Convolutional Neural Networks Meet Vision Trans-formers. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, June 18-24, 2022, 12175-12185. https://doi.org/10.1109/CVPR52688.2022.01186
13. Guo, M. H., Lu, C. Z., Liu, Z. N., Cheng, M. M., Hu, S. M. Visual Attention Network. Computational Visu-al Media, 2023, 9(4), 733-752. https://doi.org/10.1007/s41095-023-0364-2
14. Guo, M. H., Lu, C. Z., Hou, Q, Liu, Z., Cheng, M. M., Hu, S. M. Segnext: Rethinking Convolutional Atten-tion Design for Semantic Segmentation. Advances in Neural Information Processing Systems, 2022, 35, 1140-1156.
15. Hu, J., Shen, L., Sun, G. Squeeze-and-Excitation networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, June 18-23, 2018, 7132-7141. https://doi.org/10.1109/CVPR.2018.00745
16. Hou, Q., Zhang, L., Cheng, M. M., Feng, J. Strip Pooling: Rethinking Spatial Pooling for Scene Parsing. Proceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition (CVPR), Seattle, USA, June 13-19, 2020, 4003-4012. https://doi.org/10.1109/CVPR42600.2020.00406
17. Howard, A., Sandler, M., Chu, G., Chen, L. C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., Le, Q. V., Adam, H. Searching for Mobilenetv3. Proceedings of the IEEE/CVF International Confer-ence on Computer Vision (ICCV), Seoul, Korea (South), October 27-November 02, 2019, 1314-1324. https://doi.org/10.1109/ICCV.2019.00140
18. Huo, Z., Wang, Y., Qiao, Y., Wang, J., Luo, F. Do-main Adaptive Crowd Counting via Dynamic Scale Aggregation Network, IET Computer Vision, 2023, 17(7), 814-828. https://doi.org/10.1049/cvi2.12198
19. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M. Composition Loss for Counting, Density Map Estimation and Lo-calization in Dense Crowds. Proceedings of the Euro-pean Conference on Computer Vision (ECCV), Mu-nich, Germany, September 8-14, 2018, 532-546. https://doi.org/10.1007/978-3-030-01216-8_33
20. Jiang, X., Zhang, L., Xu, M., Zhang, T., Lv, P., Zhou, B., Yang, X., Pang, Y. Attention Scaling for Crowd Counting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, June 14-19, 2020, 4706-4715. https://doi.org/10.1109/CVPR42600.2020.00476
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. Swin Transformer: Hierarchical Vi-sion Transformer Using Shifted Windows. Proceed-ings of the IEEE/CVF International Conference on Computer Vision (CVPR), Montreal, Canada, October 10-17, 2021, 10012-10022. https://doi.org/10.1109/ICCV48922.2021.00986
22. Liu, J., Gao, C., Meng, D., Hauptmann, A. G. De-cideNet: Counting Varying Density Crowds Through Attention Guided Detection and Density Estimation. Proceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition (CVPR), Salt Lake City, USA, June 18-23, 2018, 5197-5206. https://doi.org/10.1109/CVPR.2018.00545

23. Liu, Y., Shi, M., Zhao, Q., Wang, X. Point in, Box out: Beyond Counting Persons in Crowds. Proceed-ings of the IEEE/CVF Conference on Computer Vi-sion and Pattern Recognition (CVPR), Long Beach, USA, June 15-20, 2019, 6469-6478. https://doi.org/10.1109/CVPR.2019.00663

24. Li, Y., Zhang, X., Chen, D. CSRNet: Dilated Convo-lutional Neural Networks for Understanding the High-ly Congested Scenes. Proceedings of the IEEE Con-ference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, USA, June 18-22, 2018, 1091-1100. https://doi.org/10.1109/CVPR.2018.00120

25. Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L. Crowd Counting with Deep Structured Scale Integra-tion Net-work. Proceedings of the IEEE/CVF Interna-tional Conference on Computer Vision (ICCV), Seoul, Korea (South), October 27-November 02, 2019, 1774-1783. https://doi.org/10.1109/ICCV.2019.00186

26. Lin, H., Hong, X., Ma, Z., Wei, X., Qiu, Y., Wang, Y., Gong, Y. Direct Measure Matching for Crowd Counting. In IJ-CAI, 2021. https://doi.org/10.24963/ijcai.2021/116

27. Liu, W., Salzmann, M., Fua, P. Context-aware Crowd Counting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, USA, June 16-20, 2019, 5099-5108. https://doi.org/10.1109/CVPR.2019.00524

28. Ma, Z., Hong, X., Wei, X., Qiu, Y., Gong, Y. To-wards a Universal Model for Cross-Dataset Crowd Count-ing. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, October 10-17, 2021, 3205-3214. https://doi.org/10.1109/ICCV48922.2021.00319

29. Ma, Z., Wei, X., Hong, X., Gong, Y. Bayesian Loss for Crowd Count Estimation with Point Supervision. Pro-ceedings of the IEEE/CVF International Confer-ence on Computer Vision (ICCV), Seoul, Korea (South), October 27-November 2, 2019, 6142-6151. https://doi.org/10.1109/ICCV.2019.00624

30. Ma, Z., Wei, X., Hong, X., Lin, H., Qiu, Y., Gong, Y. Learning to Count via Unbalanced Optimal Trans-port. Proceedings of the AAAI Conference on Artificial In-telligence, Vancouver, Canada, February 2-9, 2023, 2319-2327. https://doi.org/10.1609/aaai.v35i3.16332

31. Mehta, S., Rastegari, M. Separable Self-attention for Mobile Vision Transformers. arXiv preprint arX-iv:2206.02680, 2022.

32. Peng, Z., Huang, W., Gu, S., Xie, L., Wang, Y., Jiao, J., Ye, Q. Conformer: Local Features Coupling Glob-al Representations for Visual Recognition. Pro-ceed-ings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, October 10-17, 2021, 367-376. https://doi.org/10.1109/ICCV48922.2021.00042

33. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J. Large Kernel Matters-Improve Semantic Segmentation by Global Convolutional Network. Proceedings of the IEEE/CVF Conference on Computer Vision and Pat-tern Recog-nition (CVPR), Honolulu, USA, July 21-26, 2017, 4353-4361. https://doi.org/10.1109/CVPR.2017.189

34. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., An-gurlov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. Go-ing Deeper with Convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recog-nition (CVPR), Boston, USA, June 07-12, 2015, 1-9. https://doi.org/10.1109/CVPR.2015.7298594

35. Simonyan, K., Zisserman, A. Very Deep Convolu-tion-al Networks for Large-scale Image Recognition. arXiv preprint arXiv:1409.1556, 2014.

36. Sam, D. B., Peri, S. V., Sundararaman, M. N., Ka-math, A., Babu, R. V. Locate, Size, and Count: Ac-curately Re-solving People in Dense Crowds via De-tection. IEEE Transactions on Pattern Analysis and Machine Intel-ligence (TPAMI), 2020, 43(8), 2739-2751. https://doi.org/10.1109/TPAMI.2020.2974830

37. Tian, Y., Lei, Y., Zhang, J., Wang, J. Z. PaDNet: Pan-Densi-ty Crowd Counting. IEEE Transactions on Im-age Pro-cessing, 2019, 29, 2714-2727. https://doi.org/10.1109/TIP.2019.2952083

38. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. Attention is All You Need. Proceedings in Neural Information Pro-cessing Systems (NIPS), 2017.

39. Varior, R. R., Shuai, B., Tighe, J., Modolo, D. Multi-Scale Attention Network for Crowd Counting. arXiv preprint arXiv:1901.06026, 2019.

40. Vaswani, A., Ramachandran, P., Srinivas, A, Parmar, N., Hechtman, B., Shlens, J. Scaling Local Self-Attention for Parameter Efficient Visual Backbones. Proceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition (CVPR), Nash-ville, USA, June 20-25, 2021, 12894-12904. https://doi.org/10.1109/CVPR46437.2021.01270

41. Wang, F., Liu, K., Long, F., Sang, N., Xia, X., Sang, J. Joint CNN and Transformer Network via Weakly Supervised Learning for Efficient Crowd Counting. arXiv preprint arXiv:2203.06388, 2022.

42. Wang, L., Li, Y., Peng, S., Tang, X., Yin, B. Mul-ti lev-el Feature Fusion Network for Crowd Counting. IET

Computer Vision, 2021, 15(1), 60-72. https://doi.org/10.1049/cvi2.12012

43. Wan, J., Chan, A. Modeling Noisy Annotations for Crowd Counting. Advances in Neural Information Processing Systems, 2020, 33, 3386-3396.
44. Wan, J., Liu, Z., Chan, A. B. A Generalized Loss Function for Crowd Counting and Localization. Pro-ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 19-25, 2021, 1974-1983. https://doi.org/10.1109/CVPR46437.2021.00201
45. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., Luo, P. SegFormer: Simple and Efficient De-sign for Semantic Segmentation with Transformers. Advances in Neural Information Processing Systems, 2021, 34, 12077-12090.
46. Xia, Z., Pan, X., Song, S., Li, L. E., Huang, G. Vision Transformer with Deformable Attention. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, June 18-24, 2022, 4794-4803. https://doi.org/10.1109/CVPR52688.2022.00475
47. Zhang, Y., Zhou, D., Chen, S., Gao, S., Ma, Y. Sin-gle-image Crowd Counting via Multi-column Convo-lutional Neural Network. Proceedings of the IEEE Conference on Computer Vision and Pattern Recog-nition (CVPR), Las Vegas, USA, June 1-26, 2016, 589-597. https://doi.org/10.1109/CVPR.2016.70
48. Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H.S., Zhang, L. Rethinking Semantic Segmentation from a Se-quence-to-sequence Perspective with Transformers. Proceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition (CVPR), June 19-25, 2021, 6881-6890. https://doi.org/10.1109/CVPR46437.2021.00681
49. Zhang, P., Dai, X., Yang, J., Xiao, B., Yuan, L., Zhang, L., Gao, J. Multi-Scale Vision Longformer: A New Vision Transformer for High-Resolution Image Encoding. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, October 10-17, 2021, 2998-3008. https://doi.org/10.1109/ICCV48922.2021.00299
50. Zand, M., Damirchi, H., Farley, A., Molahasani, M., Greenspan, M., Etemad, A. Multiscale Crowd Counting and Localization by Multitask Point Supervision. ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, May 23-27, 2022, 1820-1824. https://doi.org/10.1109/ICASSP43922.2022.9747776