Applied Intelligence (2023) 53:22602–22614
https://doi.org/10.1007/s10489-023-04721-2

An interactive network based on transformer for multimodal crowd counting

Ying Yu[1] · Zhen Cai[1] · Duoqian Miao[2] · Jin Qian[1] · Hong Tang[1]

Accepted: 21 May 2023 / Published online: 30 June 2023

Abstract

Crowd counting is a task to estimate the total number of pedestrians in an image. In most of the existing research, good vision problems, such as in parks, squares, and bright shopping malls during the day, have been addressed. However, there is little research on complex scenes in darkness. To study this problem, we propose an interactive network based on Transformer for multi-modal crowd counting. First, sliding convolutional encoding is adopted for the image to obtain better encoding features. The features are extracted through the designed primary interaction network, and then channel token attention is used to modulate the features. Then, the FGAF-MLP is used for high and low semantic fusion to enhance the feature expression and fully fuse the data in different modes to improve the accuracy of the method. To verify the effectiveness of our method, we conducted extensive ablation experiments with the latest multimodal benchmark RGBT-CC, and we verified the complementarity between multiple modal data and the effectiveness of the model components. We also verified the effectiveness of our method with the ShanghaiTechRGBD benchmark. The experimental results showed that our proposed method exhibits good results and achieves an improvement of more than 10% in terms of the mean average error and mean squared error for the RGBT-CC benchmark.

Keywords Crowd counting · Transformer · Multimodal data · Feature fusion

1 Introduction

Crowd counting is a challenging task in the field of intelligent video analysis, and it can automatically estimate the number of people in an image or a video. Crowd counting has a wide range of applications, such as congestion analysis [1]and abnormal event monitoring [2]. In particular, with the current COVID-19 pandemic, crowd counting could be employed as an effective preventive measure to maintain safe social distancing in public places. To achieve this goal, crowd counting can be used to identify the gathering of people in time and control the population density. Consequently, due to its promising application, crowd counting is receiving increasing attention from researchers worldwide.

With the rapid development of computer vision technology, an increasing number of new techniques are being applied to crowd counting [3]. Early works mainly used traditional hand-crafted methods to extract pedestrian features, and then target detection technology was employed to label each pedestrian in a scene [4]. Finally, the total number of people can be obtained by counting the number of labeled people [5]. In recent years, deep learning has made great progress, and methods based on deep learning have demonstrated significant improvements over traditional methods. This has prompted more researchers to further explore deep learning-based approaches for computer vision tasks [6, 7], which include crowd counting. Initially, convolutional neural networks (CNNs) [8] were introduced by researchers to study crowd counting. Compared with traditional hand-crafted fea-

✉ Ying Yu
yuyingjx@163.com

Zhen Cai
caizhenup@163.com

Duoqian Miao
dqmiaor@tongji.edu.cn

Jin Qian
qjqjlqyf@163.com

Hong Tang
th@ecjtu.edu.cn

[1] College of Software Engineering, East China Jiao tong University, Nanchang 330013, China

[2] Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

ture extraction methods, CNNs with global context modeling capabilities can automatically extract complex features from images, which are more suitable for complex scenes with multiscale variation [9, 10] and target occlusion [11–13]. For example, Shi et al. [14] proposed a multiscale and gated spatial attention network for crowd counting that contains two branches. The large-scale branch is used to overcome the large-scale variation of heads, and the scale-aware attention branch is used to address complex background noise in crowd scenes. Experiments have shown that this model achieves better performance than that of traditional methods. However, CNN-based crowd counting models cannot achieve perfect performance due to limited receptive fields, and researchers have continued to search for more effective methods.

Transformer was proposed by Vaswani et al. [15] , and it has been widely used in Natural Language Processing (NLP) due to its powerful feature extraction capabilities. Dosovitskiy et al. [16] proposed the Vision Transformer (ViT) for image classification, demonstrating the powerful feature representation capability of Transformer in computer vision tasks. Since then, various improved versions of Transformer have been developed for computer vision tasks.

The Transformer has three obvious advantages over CNNs. First, its cascading self-attention module can capture dependencies among different inputs so that the local and global features in an image can be extracted automatically. Second, self-attention can also help the model pay more attention to the key regions in an image to extract more important and critical information for more accurate judgment. Finally, the input of Transformer is flexible, and it can easily encode information of any modality, so it can perform well in multimodal information fusion.

In the past, we usually studied crowd counting based on single-modal image data. However, for certain scenarios with special environments, such as low illumination, multimodal crowd counting is necessary. For example, due to insufficient light at night, RGB images may be blurred and cannot provide sufficient discriminative information. If crowd counting is based solely on RGB images, the results may be inaccurate due to the poor quality of images. In this case, if image information from other modalities could be fused, such as thermal images or infrared images that could provide additional auxiliary information, a better counting result will be obtained compared with that using only the traditional single-modal image. Therefore, researchers are increasingly interested in how to combine multisource image information to improve the accuracy of pedestrian counts.

Motivated by the superior performance of Transformer and the strong demand for multimodal crowd counting, we propose a novel interactive multimodal crowd counting network based on Transformer (IMMNet-T). Its overall performance with the RGBT-CC benchmark is shown in Fig. 1.

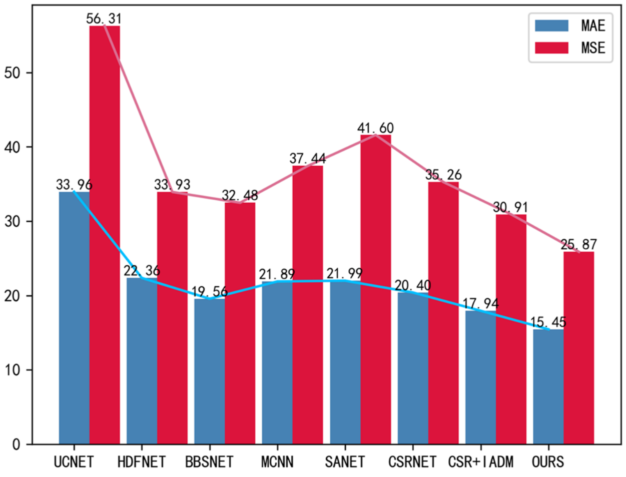


Fig. 1 The performance with the RGBT-CC benchmark. The height of the blue column indicates the Mean Absolute Error (MAE) and the red column indicates the Mean Square Error (MSE). Note that CSR+IADM in the abscissa in the figure is CSRNet+IADM. The abscissa means recent different methods

The main contributions of this work can be summarized as follows.

(1) We introduce Transformer into the field of crowd counting and propose a transformer-based interactive network, IMMNet-T, to solve the multimodal crowd counting problem, simultaneously modeling both local and global correlations among different inputs.

(2) Incorporating self-attention and the proposed token attention into the IMMNet-T model can help the model focus more on the crowd regions, which helps to improve counting accuracy.

(3) We design an effective sliding convolutional encoding (SCE) module and a feature grouping alignment multilevel fusion module with multilayer perceptron (FGAF-MLP). With these blocks, we can strengthen the captured features and capture the complementarities between different modalities.

(4) We conduct experiments on two crowd counting benchmarks, RGBT-CC [17] and ShanghaiTechRGBD [18]. The experimental results demonstrate that the proposed model is effective for multimodal crowd counting.

2 Related work

2.1 Crowd counting

With the widespread application of intelligent video analysis, crowd counting has attracted extensive attention from many researchers worldwide. Early researchers relied mainly on traditional computer vision methods to extract hand-crafted features, and then obtained the number of pedestrians in an image by means of target detection [19] or regression [20]. Traditional methods have limitations in extracting higher-level semantic features from images. Due to the challenges of multiscale variation, occlusion and dense crowds, traditional crowd counting methods usually cannot perform as well as we expect. In recent years, deep learning techniques have developed rapidly, and the focus of crowd counting research has shifted from traditional methods to deep learning-based approaches. Compared to traditional methods, deep learning-based approaches can automatically extract complex semantic features from scenes, which are useful for improving counting accuracy. First, CNN-based target detection methods were used for crowd counting [21], CNN was used to construct a detection model to predict the bounding box for a person as a whole or the local area of a person, and the number of boxes was regarded as the number of persons. However, detection-based counting methods cannot work well in crowded scenes, as it is difficult to accurately label each person. To solve this problem, researchers employed a point in the center of each head to represent a person rather than a box, and then a density map was generated from these points using a Gaussian kernel function. Currently, density mapping-based methods have become the mainstream crowd counting approaches, and the quality of the density map is the key to improving counting accuracy. Therefore, how to obtain a high-quality density map is an important research direction in the field of crowd counting. To improve the quality of density maps, Yang et al. [22] perceived the area of dense crowds by combining them with depth maps. Jiang et al. [23] used an attention mechanism to generate higher quality density maps. Although the density map-based method is widely used, it is essentially a fuzzy estimation, and the results may be inaccurate. To address this, Ma et al. [24] proposed a point-to-point Bayesian loss function to obtain more accurate counts. Recent studies have shown the need for multimodal crowd counting. Liu et al. [17] proposed a general information aggregation distribution module to capture the complementary information of different modalities and constructed a RGBT-CC benchmark for multimodal crowd counting.

2.2 Vision transformer

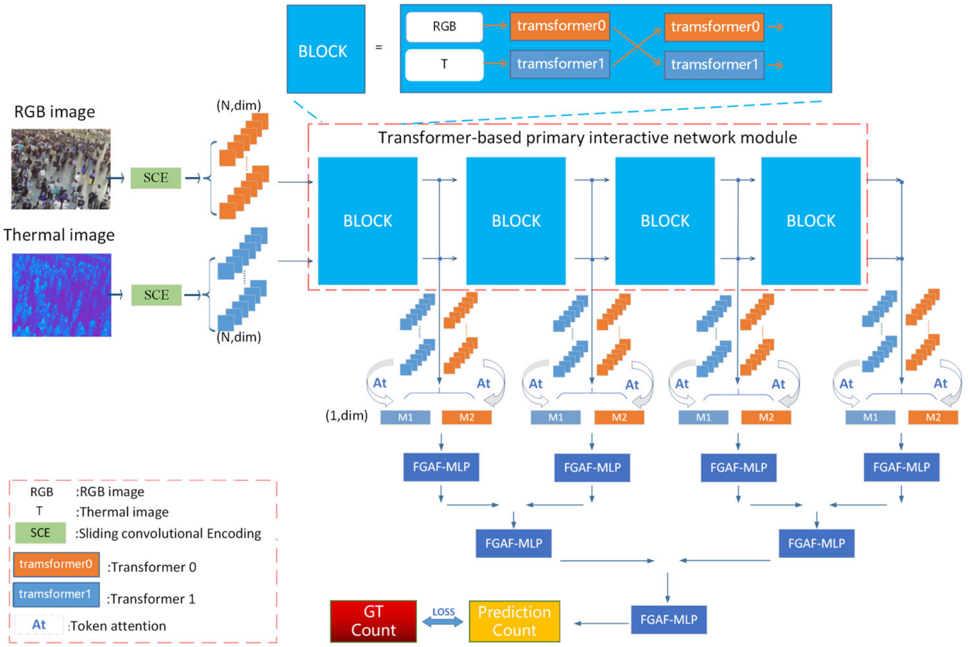
Transformer [15] is a promising neural network that encodes the input data as powerful features via an attention mechanism, and has obtained great progress in various machine learning tasks. [16] proposed the Vision Transformer (ViT) model for image recognition. It is a breakthrough work that extended Transformer from NLP to computer vision. Subsequently, various transformer-based computer vision models were successively proposed. Carion et al. proposed an end-to-end object detection model called DERT [25]. It uses a transformer encoder-decoder architecture as the main component of the framework and views object detection as a direct set prediction problem. He et al. [26] proposed a transformer architecture for fine-grained recognition, namely, TransFG. It integrates all raw attention weights of a transformer into an attention map to guide a network to effectively and accurately select discriminative image patches. Basically, visual transformers first divide the input images into several local patches and then calculate both representations and their relationship. Thus, the attention inside the local patches is ignored, but it is also essential for building visual transformers with high performance. Therefore, Han et al. [27] proposed a Transformer-in-Transformer (TNT) architecture for visual recognition. It further embeds a sub-transformer into the architecture for excavating the features and details of smaller visual words. In addition, Swin-Transformer [28] is a general-purpose hierarchical transformer, and it has the flexibility to model at various scales.

Motivated by the good performance of Transformer applied in other fields, researchers began to employ the Transformer to solve single-modal crowd counting problems. Liang et al. [29] proposed a transformer for weakly supervised crowd counting. It uses a Transformer to directly regress the number of people. Gao et al. [30] ntroduced a window-based vision transformer into crowd localization, and proposed a Dilated Convolutional Swin Transformer (DCST). The designed dilated convolution module is inserted in different stages to enhance the context information, effectively improve the representation learning ability and achieve excellent performance. Due to the inherent structural properties of the Transformer, it has advantages in handling multimodal data. However, perhaps due to the lack of multimodal experimental data, there is still less work on multimodal crowd counting based on the Transformer.

3 Methods

In this section, we introduce the architecture of the proposed model, IMMNet-T, in detail, and its overall network architecture is shown in Fig. 2. IMMNet-T is a transformer-based network for multimodal crowd counting , which mainly consisting of four parts: a sliding convolutional encoding module, a transformer-based primary interactive network module, a token attention module, and a feature grouping alignment fusion module with MLP (FGAF-MLP). The sliding convolution encoding module is responsible for encoding the input image, the main interactive network module is responsible for extracting the features from two modalities, the token attention module assigns different weights to different tokens, and the multilevel feature fusion module can fuse the image information of different modalities.

Fig. 2 The overall architecture of our method



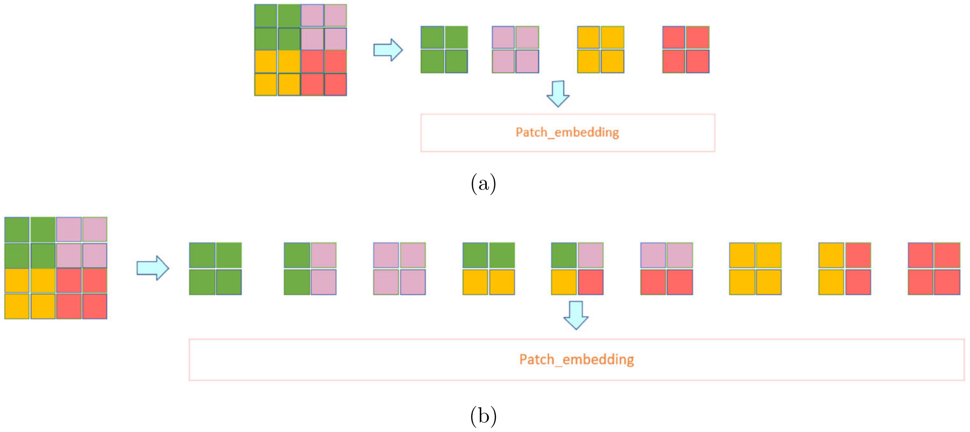
3.1 Sliding convolutional encoding

In the image preprocessing phase of the ViT [16] model, each image is split into a sequence of fixed-size tokens, also known as patches, and then multiple Transformer encoders are applied to model the global relation among these tokens for image classification. Although ViT proves that the full-transformer architecture is promising for vision tasks, its performance is not always superior to that of CNN-based models. Because the ViT tends to ignore important local structures, such as edges and lines, when directly splitting images into several nonoverlapping blocks with fixed lengths for encoding, as shown in Fig. 3 (a), it requires significantly more training samples than CNNs to achieve similar performance.

To overcome the above limitation, Li et al. [31] changed the architecture of the simple tokenization (hard split) used in the ViT and developed a Tokens-to-Token Vision Transformer (T2T-ViT) model, which splits an image into tokens with overlap (soft split). Thus, important local structure information can be encoded for each token.

Inspired by the validity of the T2T-ViT, we apply soft split in IMMNet-T to reduce the information loss in generating tokens from the image. As shown in Fig. 3 (b), the sliding convolutional encoding module splits the image into overlapping patches via a sliding window. Each patch is correlated with surrounding patches, and thus, local information can be aggregated from surrounding pixels and patches. The sliding window is similar to the convolution operation without convolution filters, which is equivalent to introducing a

Fig. 3 Block encoding in ViT(a) and sliding convolutional encoding(b)



convolution into the transformer to learn local information. After obtaining the overlapping patches, we flatten each patch and map it to *dim* dimensions with a trainable linear projection. We refer to the output of this projection as the patch embeddings.

Unlike traditional block coding used in the ViT, the sliding convolutional encoding is more flexible, and the step size of the convolution can be adjusted according to the task.

When conducting the soft split, the number of overlapping patches N of the image is calculated as follows.

$$N = N_h \times N_k \tag{1}$$

where,

$$N_h = \frac{H-K}{S} + 1 \tag{2}$$
$$N_k = \frac{W-K}{S} + 1 \tag{3}$$

H and W denote the height and width of the input image, respectively, K is the size of the sliding convolution kernel, and S indicates the sliding step. After a sliding convolution operation, an image with resolution $H \times W$ is is encoded as (N, *dim*), where N is calculated as in (1) to represent the number of generated patches, and *dim* is a hyperparameter indicating the dimension of the output of encoding operation.

3.2 Transformer-based primary interactive network module

To make progress in understanding the multiple modalities in the world around us, multimodal machine learning (MML) [32]was proposed. It aims to imitate human perception, such as sight, sound and touch, and build intelligent models that can extract and relate information from multiple modalities. Multimodal machine learning has enabled various applications, including audio-visual speech recognition and image captioning, and it is a vibrant research field with extraordinary potential. For crowd counting, it is also necessary to develop multimodal counting methods. In some special scenarios, such as insufficient illumination, the image information of a single modality cannot provide sufficient information for crowd counting. We need to leverage multimodal image information extracted from the same scenario to perform more robust predictions. These multiple modalities might allow us to capture complementary information that is not visible in a single modality on their own.

Due to the intrinsic advantages and scalability in modeling different modalities, Transformer is well positioned to handle multimodal tasks, allowing easy encoding of data in different modalities with a unified framework. For example, ViLBERT [33] is a multimodal co-attentional transformer model that can learn joint visual-linguistic representations. It extends the popular BERT [34] language model to a multimodal two-stream model, processing both visual and textual inputs in separate streams that interact through co-attentional transformer layers. This structure can accommodate the differing processing needs of each modality and provides interactions between modalities at varying representation depths.

Inspired by ViLBERT's success in learning joint representations from paired multimodal data, we develop an analogous Transformer-based interactive module and training tasks to learn joint representations from multimodal crowd image data. Specifically, we consider jointly representing RGB images and thermal images, and the structure of the proposed module is shown in Fig. 2. It consists of four blocks, where each block has two parallel Transformer-based streams operating over the RGB image and thermal image of the same scene. Within each block, the two data streams interact through the exchange of different modal information. As shown in Fig. 2, the RGB image is input to Transformer 0, and then the output of Transformer 0 is input to Transformer 1, while the thermal image is processed in the opposite way at the same time. After the processing of four stacked blocks, the features of two kinds of modal images can be extracted. It is worth mentioning that the proposed feature extraction module mainly extracts the global features of two modal images, and captures the semantic correlation between different modalities through interaction. Then, channel token attention is utilized to assign weights to the features.

3.3 Token attention module

An attention mechanism can be regarded as a mechanism for resource allocation and is widely used in various research fields. In the field of natural language processing, Ayetiran et al.[35] proposed an attention-based CNN-embedded model that learns in both sentiment and textual directions and allows for the capture of high-level semantic and contextual features. In the field of computer vision, classical attention models include CBAM [36] and DANet [37], which greatly improve the performance of computer vision tasks. In addition, there are also attention models for cross-cutting research fields, such as BERT[40] and ViLBERT [33], which can be used for joint textual and visual learning tasks.

To extract valid crowd information for counting, we introduce an attention mechanism in multimodal crowd counting. Although existing crowd counting models also introduce attention mechanisms, such as CBAM, to improve the counting accuracy, most of the existing models are based on CNNs, so these attention mechanisms are not suitable for direct migration to the Transformer-based crowd counting networks. In this paper, we refer to the architecture of CBAM to construct the token attention extraction module shown in Fig. 4. The specific process consists of four stages. First, N *dim*-dimensional vectors output by the Transformer-based

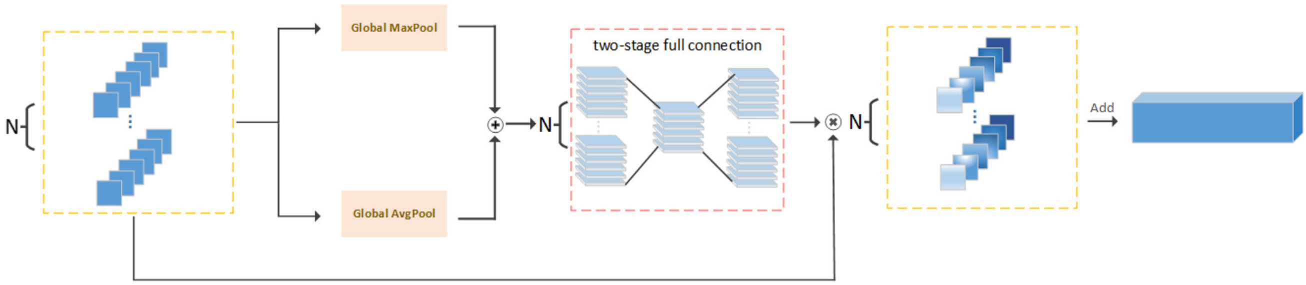


Fig. 4 The architecture of token attention module

primary interactive network module are simultaneously input to pixel-level global max pooling and global average pooling for information compression to obtain two N-dimensional vectors. Then, two vectors are added together to generate a N-dimensional vector which goes through a two-stage full connection to obtain an N-dimensional attention vector. Finally, the obtained N-dimensional attention vector is multiplied with the original N *dim*-dimensional feature vectors so that each token is given a certain weight, and then N *dim*-dimensional vectors with weight are added to get a *dim*-dimension vector. The calculation method of token attention *TAt* is shown in Formula (4), and the final output is calculated as Formula (5).

$$TAt(F_{N\times K}) = \sigma(FC(GMaxPool(F)_{N\times 1} + GAvgPool(F)_{N\times 1})) \tag{4}$$

$$Fout_{1\times K} = Add(F_{N\times K} \otimes TAt(F)_{N\times 1})) \tag{5}$$

Where, F is the original $N \times dim$ feature matrix, σ is a sigmoid function, and $N \times 1$ is the dimension of token attention, $\otimes$ represents multiplying by weight, Add means to aggregate the information and output the Fout, which is a $1 \times dim$ dimension feature vector.

3.4 Feature grouping alignment multilevel fusion module with multilayer perceptron

In the field of computer vision, information fusion is an effective means to improve the performance of algorithms. It can merge information from multiple sources, which can generate improved information with better quality and higher accuracy for decision-making than those of a single source alone [38]. Based on the level of fusion, information fusion can be further subdivided into data fusion, feature fusion, and decision fusion [39]. Data fusion is the lowest level, where the original multisource data are fused directly to obtain a roughly unified representation. It preserves the original information, so there is little information loss, but it is computationally intensive, and the fused data may have redundant information. Decision fusion is the highest level, which is based on model fusion to make comprehensive decisions. It is less computationally intensive, but the loss of information brings about a decrease in accuracy. The feature fusion belongs to the intermediate level, which reduces the redundancy of the original data and only retains the necessary information. Moreover, feature fusion can realize mutual complementarity of different features, thus obtaining more robust and accurate prediction results. Therefore, feature fusion is currently the optimal choice.

There are many ways to implement feature fusion, and skip connections are an effective method. Deep neural networks, such as ResNet [40], often employ skip connections to fuse multiscale features to improve the performance. Features extracted from lower layers have high resolution and contain more details but less semantics, while features extracted from higher layers have low resolution and contain more abstract semantic information but are less perceptive of details. Therefore, the advantages can be complemented by fusing the features extracted from the high and low layers through skip connections.

In the multimodal crowd counting task, the extracted multimodal features reflect different characteristics of the image. If the complementarity between different modalities can be exploited and the features of different modalities are combined into one more discriminative feature, the robustness of the counting model can be improved. Therefore, inspired by ResNet, we propose a feature grouping alignment multilevel fusion module with a multilayer perceptron (FGAF-MLP), the architecture of which is shown in Fig. 5. It fuses the features of different modalities through skip connections.

As shown in Fig. 2, two multimodal feature maps M1 and M2 output from each block are input to the token attention module, and then two *dim*-dimensional multimodal feature vectors are generated. Subsequently, the operations are the same for both modal vectors, as shown in Fig. 5. First, they are simultaneously partitioned into k groups. The first group of both modalities is directly input into a multilayer perceptron (MLP), and the output is a vector that is twice as long as the input. This output vector is then divided equally into two parts. The first parts of the two modalities are concatenated and fed into the MLP to obtain the multimodal fusion output f_1 of the first group. The second part is concatenated

Algorithm 1 Feature grouping alignment multi-level fusion with MLP

```
Require: Input : $M$, k
Ensure: Output : $R$
$M \leftarrow Multimodal\ feature$
k $\leftarrow Number\ of\ feature\ blocks$
for $i \leftarrow 1$ to 2 do
    $m_{i1},\ m_{i2},\ m_{i3} \ldots m_{ik} = Split\ (M_i)$
end
for $j \leftarrow 1$ to $k$ do
    for $i \leftarrow 1$ to 2 do
        if $j == 1$ then
            $t_{ij} = MLP_{(len->2len)}(m_{ij})$
            $t_{ij}^{top}, t_{ij}^{down} = Split(t_{ij})$
        else if $j == k$ then
            $t_{ij} = Concat(m_{ij}, t_{(3-i)(j-1)}^{down})$
            $t_{ij}^{top} = MLP_{(2len->len)}(t_{ij})$
        else
            $t_{ij} = Concat(m_{ij}, t_{(3-i)(j-1)}^{down})$
            $t_{ij} = MLP_{(len->len)}(t_{ij})$
            $t_{ij}^{top}, t_{ij}^{down} = Split(t_{ij})$
        end if
    end
    $f_j = MLP_{(len->len)}(Concat(t_{1j}^{top}, t_{2j}^{top}))$
end
$R = Concat(f_j)$
return $R$
```

with the second group of other modalities, and then the concatenated vectors perform the same operation as those of the first group. This process is repeated continuously, and finally, the other multimodal fusion output $f_j(j = 2, ..., k)$ of each group is obtained. It should be mentioned here that Group k is special. After it is concatenated with the feature map generated by Group k-1, it is fed into the MLP. However, since it is the last group, the information output from MLP is not split like other groups. Finally, each group of multimodal fusion information $f_j(j = 1, ..., k)$ is concatenated to obtain the final output . In this way, the cross-fusion of two modal information between adjacent groups is achieved, and finer features are extracted while maintaining the original feature mapping, and the algorithm flow is shown in Algorithm 1.

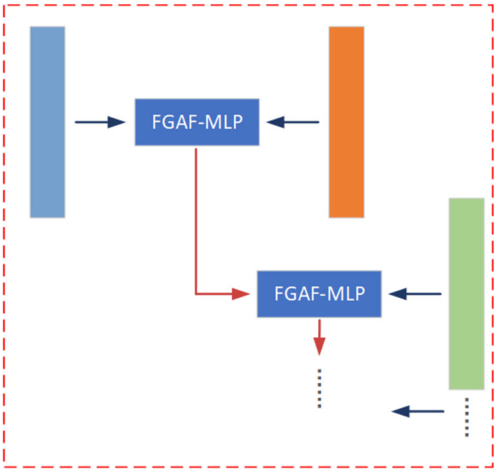


Fig. 6 The extended process of FGAF-MLP

As seen from the structure of the FGAF-MLP module, it is a highly pervasive module that can be easily extended to the fusion of information from more than two modalities. The specific process is shown in Fig. 6, where different colors represent different modal information.

Fig. 5 The structure of FGAF-MLP

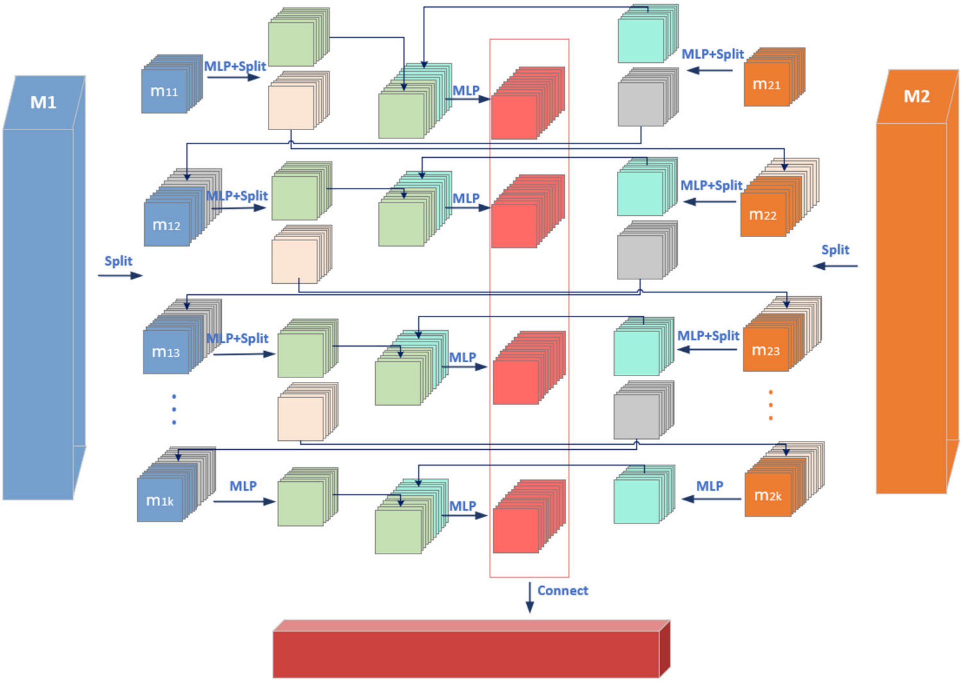


Table 1 The Information of RGBT-CC benchmark [17]

	Training	Validation	Testing
Bright	510/65.66	97/63.02	406/73.39
Dark	520/62.52	103/67.74	394/74.88
ALL	1030/64.07	200/65.45	800/74.12
Scene	malls, streets, train/metro station, etc.		

The training, validation and testing sets of RGBT-CC benchmark. In each grid, the first value is the number of images, while the second value implies the average number of people per image

4 Experimental details

Our proposed approach is implemented based on the Pytorch framework. The parameter k of FGAF-MLP module is set to 4 and the fusion process is divided into three levels, as shown in Fig. 2. The first level contains four fusion modules, the second level has two fusion modules, while the third level has only one fusion module. During the training process, a pair of images is fed into the network simultaneously. The size of all input images is fixed, with the height and width are both set to 256. The learning rate is set to 1e-5, and the Adam is used to optimize our model. Similar to the IADM+ [17] algorithm, we use BAYESIAN+ [41] as the loss function.

4.1 Experimental benchmarks

RGBT-CC [17]. It is a large-scale RGBT Crowd Counting (RGBT-CC) benchmark, which contains 2,030 pairs of RGB-thermal images with 138,389 annotated pedestrians. All the images are captured from different scenarios, such as shopping malls, streets, railway stations, subway stations, etc., and each image is 640 × 480 in size. Among these samples, 1,013 pairs are captured in the light and 1,017 pairs are in the darkness. The RGBT-CC crowd counting dataset is randomly divided into three parts. As shown in Table 1, 1030 pairs are used for training, 200 pairs are for validation and 800 pairs are for testing, respectively. Figure 7 shows some representative images from the RGBT-CC dataset. It can be seen that in some cases, the information of the two different modalities, RGB image and thermal image, is complementary, and this complementary information can greatly help to recognize pedestrians.

ShanghaiTechRGBD [18]. This large-scale RGB-D crowd counting dataset contains 2,193 images, of which 1193 images are randomly selected as the training set, accounting for 54.4% of the total, and the remaining images are used as the test set. A total of 144,512 heads are labeled for the entire dataset, with an average of 66 heads labeled per image. Each image in ShanghaiTechRGBD has been appropriately cropped to a size of 1920 × 1080. These images were captured by a stereo camera (ZED3) with an effective depth range from 0 to 20 meters. The scenarios in this benchmark include busy streets and crowded parks with illumination conditions ranging from very bright to very dark, and each scene contains both RGB image and Depth image modal information. Figure 8 shows some representative images from the ShanghaiTechRGBD dataset.

4.2 Evaluation

Mean Absolute Error (MAE) and Mean Square Error (MSE) are commonly used as evaluation metrics to measure the performance of crowd counting algorithms, and their calculation formula is as follows.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|C_i - C_i^{GT}| \tag{6}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(C_i - C_i^{GT})^2 \tag{7}$$

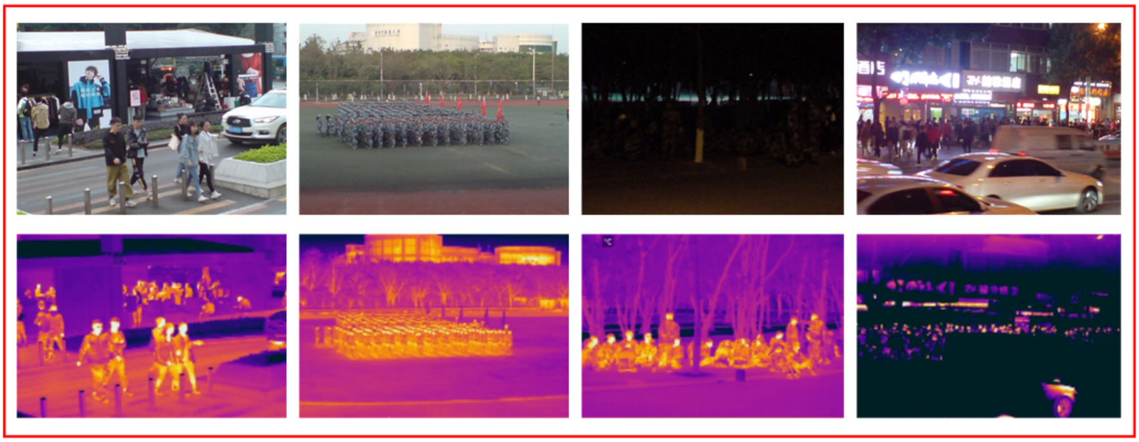


Fig. 7 Some images from the RGBT-CC benchmark

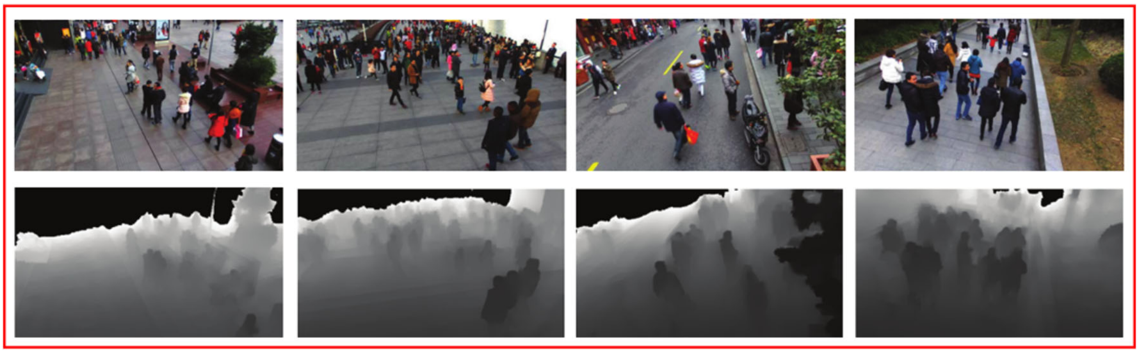


Fig. 8 Some images from the ShanghaiTechRGBD benchmark

where m represents the total number of images in the dataset, Ci represents the predicted number of pedestrians in the i-th image, and Ci^{GT} represents the actual number of pedestrians in the i-th picture. *MAE* is usually related to the accuracy of the model, while *MSE* is related to the stability of the model. The smaller these two values, the better performance of the model.

5 Experiments

5.1 Ablation studies

Verify the effectiveness of multimodal data: First, we conduct a comparison experiment with the RGBT-CC benchmark to verify the effect of multimodal inputs on the counting performance under different illumination conditions. The input information is divided into three types: RGB images, thermal images and RGB-thermal images. As shown in Table 2, in both bright and dark scenarios, the crowd counting performance is significantly better when the input information is multimodal, such as RGB-thermal images, than when the input information is unimodal, such as RGB images or thermal images, indicating that the complementarities provided by the multimodal information facilitate the counting performance. Furthermore, we can also observe that the counting performance is better with thermal images than with RGB images in both bright and dark environments, indicating that thermal images are better for identifying potential pedestrians from the cluttered background. It is worth mentioning that the worst performance occurs when counting in dark scenarios using only RGB images, mainly because it is not easy to effectively recognize pedestrians in dark environments. In this case, pedestrian information can only be captured with the help of information from other modalities, such as thermal images. Here, we deliberately compare our proposed model with another multimodal crowd counting model, CSRNet+IADM [17], and as seen in Table 2, the overall trend of both models is similar, but our model performs significantly better than the other model for the three kinds of input.

Verify the effectiveness of model components: To verify the effectiveness of each component in our proposed framework, we conduct sufficient ablation studies with the RGBT-CC benchmark. The three components that need to

Table 2 The performance under different illumination conditions with the RGBT-CC benchmark

Method	Illumination	Input Data	MAE	MSE
CSRNet+IADM [17]	Brightness	RGB	23.49	45.40
		thermal	25.21	40.60
		RGB+thermal	20.36	32.57
	Darkness	RGB	44.72	87.81
		thermal	17.97	33.74
		RGB+thermal	15.44	29.11
OURS	Brightness	RGB	16.34	27.57
		thermal	16.61	26.75
		RGB+thermal	15.81	26.47
	Darkness	RGB	18.64	34.02
		thermal	16.49	27.10
		RGB+thermal	16.03	27.52

Table 3 Performance of different components of IMMNet-T with the RGBT-CC benchmark

Transformer Backbone	SCE	Token Attention	FGAF-MLP	Illumination	MAE	MSE
✓				Brightness+Darkness	16.85	29.08
✓	✓			Brightness+Darkness	15.96	27.97
✓		✓		Brightness+Darkness	16.36	28.33
✓			✓	Brightness+Darkness	16.05	26.49
✓	✓	✓		Brightness+Darkness	15.84	27.45
✓	✓		✓	Brightness+Darkness	15.52	26.20
✓		✓	✓	Brightness+Darkness	16.01	27.62
✓	✓	✓	✓	Brightness+Darkness	15.45	25.87

All the methods in this table utilize both RGB images and thermal images to estimate the crowd counts

be validated are SCE, token attention and FGAF-MLP. As shown in the previous sections, the backbone of our proposed model consists of multiple Transformer encoders. We refer to a backbone that only contains the Transformer encoder block without any components as a baseline, such as the Transformer based main interaction network module in Fig. 2. The experimental procedure is divided into two steps. First, we test the performance of the baseline. Then, we add a single component or combinations of different components to the baseline to verify whether they can improve the performance of the model. As shown in Table 3, a total of 7 ablation experiments are performed on the RGBT-CC benchmark, with the baseline performing the worst. In the single-component ablation experiments, it can be seen that SCE contributes the most to the model, followed by FGAML-MLP and Token Attention. The addition of SCE to the backbone network results in a large improvement in counting performance. In the two-component combination ablation experiments, it can be seen from Table 3 that when SCE is combined with FGAML-MLP or Token Attention, the counting performance is significantly improved compared to adding only a single component, and the combination of SCE and FGAML-MLP is better than the combination of SCE and Token Attention. Each component can play a role in improving the model performance. Overall, the more components that are added, the better the performance of the model, and the counting performance of the model is optimal when all three components are added to the baseline, namely, the proposed IMMNet-T counting model. We have also discovered that SCE is not negligible compared to other components. The counting model performs well when SCE is added to the baseline alone or together with other components. By analyzing the experimental data, it can be seen that the number of annotations per image in the RGBT-CC benchmark is small, with only 68 annotations in an image on average. It is difficult to achieve good learning performance with such sparsely labeled images, if they are directly fed into the Transformer network, while human head features can be better extracted with sliding overlap coding of images via the SCE component, which can better identify pedestrians. The above analysis shows that all three components are essential to our proposed method.

5.2 Comparison with state-of-the-art methods

To illustrate the validity of our model, we compare the proposed method with several state-of-the-art models on two large-scale crowd counting benchmarks. The models involved in the comparison can be divided into two categories. The first category includes models such as MCNN [4], SANet [43], CSRNet [44], and MVMS [45], which are specifically designed for crowd counting. The second category contains a few of best-performing models dedicated to multimodal learning, including UCNet [46], HDFNet [47], and BBSNet [48]. In order to employ them for multimodal crowd counting, some adjustments are made to these multimodal learning models. In addition, CSRNet+IADM[17], a classical multimodal crowd counting model, is also involved in the comparison.

Firstly, we conduct a comparison experiment on the RGBT-CC dataset. As shown in Table 4, the proposed model performs better both when compared to popular crowd counting models and to modified multimodal crowd counting models. For example, compared with those of the popular CSRNet, the proposed model increases by 24.2% in MAE and 26.6% in MSE. This may be due to the fact that the extracted multimodal features complement each other, which results in more comprehensive crowd features and improved counting performance. Compared with those of the effective CSRNet +IADM, the MAE of our proposed model increased by 13.88%, and the MSE increased by 16.31%. This is due to the effective cross-transformer feature extraction module, the improved data encoding and the design of FGAF-MLP for more detailed fusion of multimodal features. Thus, it can be seen that the proposed idea of fusing two types of modal information for counting pedestrians and making full use of their complementarity to enhance the counting effect is feasible, and the final counting effect is also impressive. The training process of the experiment is shown in Fig. 9 (left).

Table 4 Performance of different methods with the proposed RGBT-CC benchmark

Method	Journal / Conference	MAE	MSE
MCNN [42]	CVPR 2016	21.89	37.44
SANet [43]	ECCV 2018	21.99	41.60
CSRNet [44]	CVPR 2018	20.40	35.26
MVMS [45]	CVPR 2019	19.97	33.97
UCNet [46]	CVPR 2020	33.96	56.31
HDFNet [47]	ECCV 2020	22.36	33.93
BBSNet [48]	ECCV 2020	19.56	32.48
CSRNet+IADM [17]	CVPR 2021	17.94	30.91
OURS	-	**15.45**	**25.87**

All the methods in this table utilize both RGB images and thermal images to estimate the crowd counts

Table 5 Performance of different methods with the ShanghaiTechRGBD benchmark

Method	Journal / Conference	MAE	MSE
MCNN [42]	CVPR 2016	11.12	16.49
SANet [43]	ECCV 2018	5.74	8.66
CSRNet [44]	CVPR 2018	4.92	7.41
UCNet [46]	CVPR 2020	10.81	15.70
HDFNet [47]	ECCV 2020	8.32	13.01
BBSNet [48]	ECCV 2020	6.26	9.26
DetNet [49]	CVPR 2018	9.74	13.14
CL [50]	ECCV 2018	7.32	10.48
RDNet [18]	CVPR 2019	4.96	7.22
CSRNet+IADM [17]	CVPR 2021	4.38	7.06
OURS	-	**3.97**	**6.56**

All the methods in this table utilize both RGB images and Depth images to estimate the crowd counts

The blue line indicates the relationship between the MSE and training epochs, and the yellow line indicates the relationship between the MAE and training epochs. During the training process, the MAE and MSE show an overall decreasing trend, and eventually, the two evaluation indicators gradually stabilize.

Second, we also verify the validity of the proposed model with another multimodal crowd counting benchmark, ShanghaiTechRGBD [18]. We utilize the RGB images and depth maps as a pair of multimodal inputs, and the depth map provides additional information about the localization of heads. The specific experimental results are shown in Table 5. As seen from the table, the performance of the proposed model is the best. It has great advantages not only compared with multimodal learning models, such as UCNet [46], HDFNet[47], and CSRNet+IADM [17], but also compared with traditional crowd counting models, such as MCNN [42], CSRNet [44]. The training process of the experiment is shown in Fig. 9 (right). As in the left figure, the blue and yellow lines represent the relationship between the MSE, MAE, and training epochs, respectively. During the training process, the MAE and MSE gradually decrease and finally stabilize. This experiment shows that our proposed method is also effective for other types of multimodal crowd datasets.

6 Conclusion

In this paper, we propose a transformer-based interactive network to process different modal data for multimodal crowd counting. Through token attention and the FGAF-MLP module, the complementary information between different modal data is captured. We conduct extensive ablation experiments and complete comparative experiments with two large-scale multimodal crowd counting benchmarks. The experimental results show that the proposed model performs well, and our proposed multimodal counting approach is feasible and effective. It should also be noted that the design of the network structure gives our proposed model some advantages

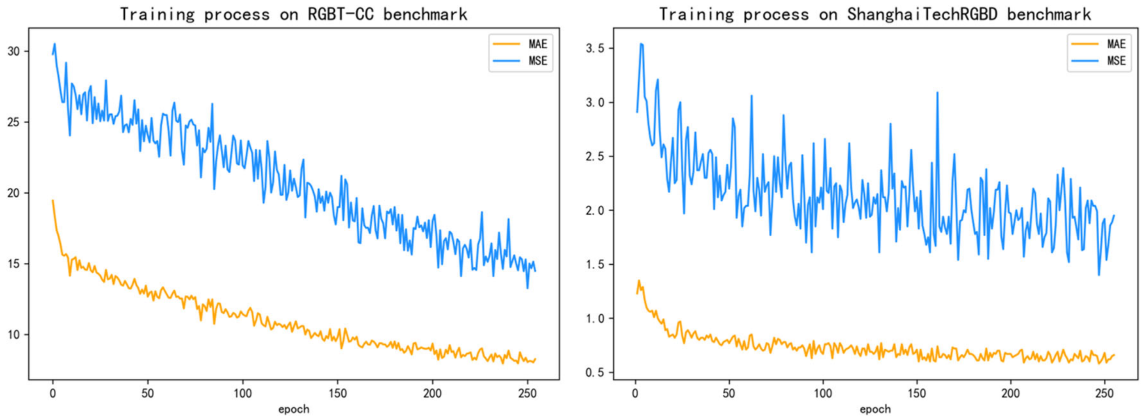


Fig. 9 The training process of RGBT-CC benchmark(left) and ShanghaiTechRGBD benchmark(right)

in multimodal crowd counting, but its ability in crowd localization is lacking, which would be an important direction of our future research. In the future, more researchers will focus on the research field of multimodal crowd counting, and the counting scenarios would be more abundant and diverse. In summary, future research in multimodal crowd counting could be carried out from the following directions.

(1) Continue to explore more effective multimodal crowd counting approaches. For example, RGB images and depth images can be combined to divide the scene to alleviate background interference and crowd scale variation.

(2) Continue to explore more potential applications of multimodal crowd counting in conjunction with crowd localization, such as nighttime infrared crowd localization and event detection.

Acknowledgements This paper was supported by the National Natural Science Foundation of China (No. 62163016, 62066014, 62202165), the Natural Science Foundation of Jiangxi Province (20212ACB202001, 20202BABL202018), and the Double Thousand Plan of Jiangxi Province of China.

Statements and Declarations

We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

References

1. Kumar N, Raubal M (2021) Applications of deep learning in congestion detection, prediction and alleviation: A survey. Transp Res C Emerg Technol 133:103432. https://doi.org/10.1016/j.trc.2021.103432. Get rights and content
2. Bamaqa A, Sedky M, Bosakowski T et al (2022) SIMCD: SIMulated crowd data for anomaly detection and prediction. Expert Syst Appl 203:117475. https://doi.org/10.1016/j.eswa.2022.117475. Get rights and content
3. Fan Z, Zhang H, Zhang Z et al (2022) A survey of crowd counting and density estimation based on convolutional neural network. Neurocomputing 472:224–251. https://doi.org/10.1016/j.neucom.2021.02.103
4. Topkaya I S, Erdogan H, Porikli F (2014) Counting people by clustering person detector outputs. In: Proc of the 11th IEEE Int Conf on Advanced Video and Signal Based Surveillance, IEEE, Piscataway, NJ, pp 313–318. https://doi.org/10.1109/AVSS.2014.6918687
5. Idrees H, Saleemi I, Seibert C et al (2013) Multi-source multi-scale counting in extremely dense crowd images. In: Proc of the IEEE Conf on Computer Vision and Pattern Recognition. IEEE, Piscataway, NJ, pp 2547–2554. https://doi.org/10.1109/CVPR.2013.329
6. Delussu R, Putzu L, Fumera G (2022) Scene-specific crowd counting using synthetic training images. Pattern Recog 124:108484. https://doi.org/10.1016/j.patcog.2021.108484
7. Yue X, Zhang C, Fujita H et al (2021) Clothing fashion style recognition with design issue graph. Appl Intell 51(6):3548–3560. https://doi.org/10.1007/s10489-020-01950-7
8. Lecun Y, Bottou L, Bengio Y et al (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
9. Yu Y, Zhu H, Wang L et al (2021) Dense crowd counting based on adaptive scene division. Int J Mach Learn Cybern 12(4):931–942. https://doi.org/10.1007/s13042-020-01212-5
10. Liang L, Zhao H, Zhou F et al (2022) SC2Net: scale-aware crowd counting network with pyramid dilated convolution. Appl Intell 1–14. https://doi.org/10.1007/s10489-022-03648-4
11. Wang K, Liu M (2022) YOLOv3-MT: A YOLOv3 using multi-target tracking for vehicle visual detection. Appl Intell 52(2):2070–2091. https://doi.org/10.1007/s10489-021-02491-3
12. Xie J, Gu L, Li Z et al (2022) HRANet: Hierarchical region-aware network for crowd counting. Appl Intell 1–15. https://doi.org/10.1007/s10489-021-03030-w
13. Wang W, Liu Q, Wang W (2022) Pyramid-dilated deep convolutional neural network for crowd counting. Appl Intell 52(2):1825–1837. https://doi.org/10.1007/s10489-021-02537-6
14. Shi Y, Sang J, Wu Z et al (2022) MGSNet: A multi-scale and gated spatial attention network for crowd counting. Appl Intell 1–11. https://doi.org/10.1007/s10489-022-03263-3
15. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. Adv Neural Inf Process Syst 5998–6008. https://doi.org/10.1609/aaai.v34i07.6693
16. Dosovitskiy A, Beyer L, Kolesnikov A et al (2021) An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations. https://openreview.net/forum?id=YicbFdNTTy. Accessed 13 Jan 2021
17. Liu L, Chen J, Wu H et al (2021) Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4823–4833. https://doi.org/10.1109/CVPR46437.2021.00479
18. Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Sheng hua Gao (2019) Density map regression guided detection network for RGB-D crowd counting and localization. In: CVPR, pp 1821–1830. https://doi.org/10.1109/CVPR.2019.00192
19. Gavrila D M, Philomin V (1999) Real-time object detection for "smart" vehicles. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, vol 1. IEEE, Kyoto, pp 87–93. https://doi.org/10.1109/ICCV.1999.791202
20. Zhang C, Li H , Wang X et al (2015) Cross-scene crowd counting via deep convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE, Boston, pp 833–841. https://doi.org/10.1109/CVPR.2015.7298684
21. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05), vol 1. IEEE, pp 886–893. https://doi.org/10.1109/CVPR.2005.177
22. Yang S D, Su H T, Hsu W H et al (2019) DECCNet: Depth enhanced crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. https://doi.org/10.1109/ICCVW.2019.00553
23. Jiang X, Zhang L, Xu M et al (2020) Attention scaling for crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4706–4715. https://doi.org/10.1109/CVPR42600.2020.00476
24. Ma Z, Wei X, Hong X et al (2019) Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6141–6150. https://doi.org/10.1109/ICCV.2019.00624. IEEE
25. Carion N, Massa F, Synnaeve G et al (2020) End-to-end object detection with transformers. In: European Conference on Computer

Vision, Springer, Cham, pp 213–229. https://doi.org/10.1007/978-3-030-58452-8_13

26. He J, Chen JN, Liu S et al (2022) TransFG: A transformer architecture for fine-grained recognition. Proc AAAI Conf Artif Intel. 36(1):852–860. https://doi.org/10.1609/aaai.v36i1.19967
27. Han K, Xiao A, Wu E et al (2021) Transformer in transformer. Adv Neural Inf Process Syst 34:15908–15919
28. Liu Z, Lin Y, Cao Y et al (2021) Swin Transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986
29. Liang D, Chen X, Xu W et al (2022) Transcrowd: weakly-supervised crowd counting with transformers. Sci China Inf Sci 65(6):160104. https://doi.org/10.1007/s11432-021-3445-y
30. Gao J, Gong M, Li X (2022) Congested crowd instance localization with dilated convolutional Swin transformer. Neurocomputing 513:94–103. https://doi.org/10.1016/j.neucom.2022.09.113
31. Yuan L, Chen Y, Wang T et al (2021) Tokens-to-Token ViT: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp 558–567. https://doi.org/10.1109/ICCV48922.2021.00060
32. Baltrušaitis T, Ahuja C, Morency L P. Multimodal machine learning: A survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443. https://doi.org/10.1109/TPAMI.2018.2798607
33. Lu J, Batra D, Parikh D et al (2019) ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems, pp 13–23. https://dl.acm.org/doi/10.5555/3454287.3454289. Curran Associates Inc., Red Hook, NY, United States
34. Devlin J, Chang M W, Lee K, Toutanova K (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp 4171–4186. https://doi.org/10.18653/v1/N19-1423. Association for Computational Linguistics
35. Ayetiran EF (2022) Attention-based aspect sentiment classification using enhanced learning through CNN-BiLSTM networks. Knowl-Based Syst 252:109409. https://doi.org/10.1016/j.knosys.2022.109409
36. Woo S, Park J, Lee J Y et al (2018) CBAM: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV), pp 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
37. Fu J, Liu J, Tian H et al (2019) Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 3146–3154. https://doi.org/10.1109/CVPR.2019.00326
38. Zhang P, Li T, Wang G et al (2021) Multi-source information fusion based on rough set theory: A review. Inf Fusion 68:85–117. https://doi.org/10.1016/j.inffus.2020.11.004
39. Li S, Kang X, Fang L et al (2017) Pixel-level image fusion: A survey of the state of the art. Inf Fusion 33:100–112. https://doi.org/10.1016/j.inffus.2016.05.004
40. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEEconference on computer vision and pattern recognition, pp 770–778. https://doi.org/10.1109/CVPR.2016.90
41. Antoni BC, Nuno V (2009) Bayesian Poisson regression for crowd counting. 2009 IEEE 12th international conference on computer vision. IEEE, Kyoto, pp 545–551
42. Zhang Y, Zhou D, Chen S et al (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. https://doi.org/10.1109/CVPR.2016.70
43. Cao X, Wang Z, Zhao Y et al (2018) Scale aggregation network for accurate and efficient crowd counting. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 734–750. https://doi.org/10.1007/978-3-030-01228-1-45
44. Li Y, Zhang X, Chen D (2018) CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1091–1100. https://doi.org/10.1109/CVPR.2018.00120
45. Zhang Q, Chan A B (2019) Wide-area crowd counting via ground-plane density maps and multi-view fusion CNNs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 8297–8306. https://doi.org/10.1109/CVPR.2019.00849
46. Zhang J, Fan D P, Dai Y et al (2020) UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8582–8591. https://doi.org/10.1109/CVPR42600.2020.00861
47. Pang Y, Zhang L, Zhao X et al (2020) Hierarchical dynamic filtering network for RGB-D salient object detection. In: European Conference on Computer Vision. Springer, Cham, pp 235–252. https://doi.org/10.1007/978-3-030-58595-2_15
48. Fan D P, Zhai Y, Borji A et al (2020) BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In: European Conference on Computer Vision. Springer, Cham, pp 275–292. https://doi.org/10.1007/978-3-030-58610-2_17
49. Liu J, Gao C, Meng D et al (2018) DecideNet: Counting varying density crowds through attention guided detection and density estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5197–5206. https://doi.org/10.1109/CVPR.2018.00545
50. Idrees H, Tayyab M, Athrey K et al (2018) Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European Conference on Computer Vision (ECCV), pp 532–546. https://doi.org/10.1007/978-3-030-01216-8-33

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.