Signal, Image and Video Processing (2023) 17:601–608
https://doi.org/10.1007/s11760-022-02266-4

ORIGINAL PAPER

Cross-modal collaborative representation and multi-level supervision for crowd counting

Shufang Li[1,2] · Zhengping Hu[1] · Mengyao Zhao[1] · Shuai Bi[1] · Zhe Sun[1]

Received: 12 February 2022 / Revised: 6 May 2022 / Accepted: 11 May 2022 / Published online: 9 June 2022

Abstract

Crowd features are often extracted from RGB images to complete the tasks of density estimation and crowd counting. However, RGB images will be affected in some particularly poor illumination, resulting in the inability to accurately identify semantic objects, and thermal images can help solve this problem. Considering the comprehensive utilization of optical and thermal imaging information, we propose a crowd counting method based on cross-modal coordinated representation and multi-level supervision. In order to capture the complementary features of different modalities, RGB and thermal images are used as specific steams of cross-modal cooperative learning. The missing specific information is compensated and the shared information is enhanced; both are through the aggregation and distribution calculation of specific steams and shared steam. Furthermore, in order to weaken the influence of the background and strengthen the identification of crowd regions, we combine the multi-scale crowd feature extraction and region recognition. Multiple output layers are added in the propagation process of multi-modal streams, so as to achieve the purpose of multi-level supervision. Moreover, we replace the baseline training loss with the Bayesian loss for monitoring the counting expectation of each annotation point. Finally, comprehensive experiments on the RGBT-CC benchmark show the effectiveness of the proposed method.

Keywords Crowd counting · Cross-modal collaborative representation learning · Multi-level supervision

1 Introduction

As a challenging computer vision task, crowd counting aims to obtain rich crowd features and generate the estimated density map corresponding to the crowd image. With the development of deep learning, a large number of counting models have subsequently been proposed. Although substantial progress has been made in relevant studies, there are still many factors that influence the acquisition of crowd features. Besides the problems of complex crowd background, scale variation within and between images, occlusion existing between dense crowd, some specific scenes also have problems such as the difficulty of detecting pedestrians due to too poor illumination conditions. Because most cameras cannot work normally in the dark environment, it will lead to a sharp decline in the performance of many methods [1], so other types of visual sensors (such as infrared cameras) began to be widely used as a supplement to RGB cameras to overcome this difficulty. Therefore, the research on the combination of multi-modal representation learning and crowd counting has been further developed.

Early researchers mainly used the complementarity of RGB image and depth image to complete the counting task. For example, Lian et al. collected the ShanghaiTech RGB-D dataset and utilized a depth adaptive kernel for considering head size variation to improve the quality of density maps [2]. In addition, depth sensing anchors are also used in the detection framework to better initialize the anchor size. Zhao et al. adopted the bifurcated backbone strategy to recombine multi-level features into teacher and student features, and they mined informative depth clues from channel and spatial views and finally fused RGB and depth modes in a complementary way [3]. The above methods mainly generate pixel level crowd density map through rich RGB information and depth information. However, in some unconstrained scenes, only using the above features may not be able to accurately

✉ Zhengping Hu
hzp_ysu@163.com

1 School of Information Science and Engineering, Yanshan University, West of Hebei Street No. 438, Qinhuangdao 066004, China

2 Department of Information Engineering, Hebei University of Environmental Engineering, Jingang Road No. 8, Qinhuangdao 066102, China

identify semantic objects. In the case of poor lighting conditions (such as backlight and night), it is difficult to detect pedestrians directly from RGB images. In addition, some objects with high similarity to pedestrians are easily mistaken for pedestrians only relying on optical features [4]. References [5–7] use multi-modal data to directly fuse features or input them into deep neural network in a combined way to connect the representation of all modes and supervise training, but these cannot make good use of the complementary information between different modes. Recently, Liu et al. released the RGBT-CC benchmark and proposed a cross-modal collaborative representation learning framework for crowd counting [8]. The framework includes two multi-modal specific branches, a modal shared branch and multiple information aggregation and distribution modules, which can fully capture the complementarity between different modalities. However, it does not fully consider the different characteristics of RGB data and thermal imaging data, so it is necessary to further enrich the information and enhance the expression of specific features.

As seen from the relevant literature, RGB images may not be able to accurately identify semantic objects in unconstrained scenes. Although the thermal images are not affected by light and shade, the existing hard negative objects are difficult to eliminate. Meanwhile, the depth images of outdoor scenes are rough, so they has certain limitations in application. In general, RGB information and thermal image information can be complementary, but it is not easy to capture the complementarity between multi-modal data. In order to capture the complementary features of different modalities, we refer to the idea of taking RGB and thermal images as multi-modal- specific branches proposed in [8] and add crowd region recognition design to the RGB modality, so as to realize the learning separation and coordinated representation of multi-modal streams under the constraint of region. Considering the influence of background and the judgment of crowd region, we combine multi-scale feature with region recognition to achieve the purpose of adaptive attention to different regions. And residual learning and skip connection are adopted. In dealing with reducing over fitting and facilitating gradient back propagation, we add multiple output layers and back propagation the gathered loss function in the multi-modal branch for the multi-level supervision. Besides that, we adopt the idea of [9] in the calculation of loss function for combating with the problems of occlusion, perspective effect, shape change, and so on in the scene. The density contribution probability model is constructed from the perspective of point annotation, so as to realize more reliable supervision on the counting expectation of each annotation point. To sum up, we propose the method based on cross-modal coordinated representation and multi-level supervision for crowd counting.

The rest of the article is arranged as follows. In Sect. 2, we will briefly introduce the related research of crowd counting based on deep neural network. Then, the method proposed in this paper is introduced in Sect. 3. In Sect. 4, we will report the performance of the method based on a series of experiments using images from the RGBT-CC benchmark [8]. Finally, we summarize the paper in Sect. 5.

2 Related work

At present, many crowd counting methods mainly use RGB image of crowd scene to extract features and generate estimated density map. Meanwhile, they can be divided into regression-based models [9–14] and models with additional detection [15–18]. For instance, Liu et al. proposed to use the pool pyramid to extract the features of supplementary scales and adaptively assign different weights to different scales and regions [10]. The lightweight hierarchical network structure used by Jiang et al. effectively combines the features from high level and low level [11]. The conditional random fields developed by Liu et al. can enable the features of each scale to obtain information from other scales [12]. And Liu et al. solved the problem of reduced counting accuracy in high congestion and noise scenes by adding attention mechanism and multi-scale deformable convolution in the network [13]. Dong et al. utilized the improved encoder–decoder structure and advanced loss function to mine the features between adjacent scales to cope with scale changes [14]. While Ma et al. proposed Bayesian loss which is helpful to establish density distribution probability model from point annotation [9]. When the crowd density is low, the detect-based method is more effective than the regression-based method; when the crowd density is high, the regression-based method is better than the detection-based method. Aim at giving full play to the advantages of these two methods at the same time, Liu et al. designed two sub networks for detection and regression, respectively, and fused the two by the attention mechanism [15]. However, this method is only suitable for experiments on datasets with low density. If the scene is very crowded, the cost of bounding box labeling is very high. Therefore, Liu et al. suggested changing the initialization of point annotation to the real annotation of the initial box and updating it in real time during training [16]. Moreover, Liu et al. employed additional positioning branches to detect head positions and scale adaptively in difficult to identify regions [17]. Furthermore, Rong et al. proposed a three branches network for feature extraction, regional recognition, and judgment of crowd density [18]. The method simulates a series of steps when a person actively observes the crowd scene, that is, first observing the crowd area, then paying attention to the crowd density of each area, and finally estimating the number of people.

While the above methods cannot effectively identify invisible pedestrians in the case of poor illumination conditions, multi-modal representation learning based on deep learning has attracted extensive attention because of its powerful multi-level abstract representation ability. Therefore, some researchers also use depth map or thermal image for object recognition and crowd counting. Fu et al. proposed a novel joint learning and densely cooperative fusion architecture for rgb-d salient object detection [19]. This method mainly obtains fusion features through element level multiplication or addition and cascade operation. In order to achieve fully represented shared features, Li et al. designed a cross-modal crowd counting method combining cross-modal cycle attention fusion and fine coarse supervision [20]. While Piao et al. took the estimated map and attention map as a bridge to transmit depth knowledge [21], Zhao et al. proposed an effective multi-scale cross-modal feature fusion method for rgbd salient object detection [22]. In addition, Lu et al. used the cross-modal shared feature transfer algorithm to explore the potential of modal shared information and specific features in improving the ability of re recognition [1]. These methods chiefly use the depth data as auxiliary information to assist the representation and learning of RGB data. Lately, Liu et al. proposed to utilize the complementarity of RGB image and thermal image to complete crowd counting [8].

For achieving the purpose of crowd counting, especially some scenes with poor illumination, we catch the modal shared information and specific features with RGB—thermal image, reduce the influence of irrelevant background through region recognition, and strengthen the supervision of point, region, and multi-scale feature extraction.

3 Proposed method

3.1 Network architecture

Aiming at the counting problem of unconditionally restricted crowd scenes, we propose the density estimation method based on multi-modal representation learning and multi-level supervision. The overall frame structure is shown in Fig. 1. In considering the learning of deep multi-modal representation, we learn from the ideas of [1] and [8], comprehensively learn the complementarity between different modes of RGB image and thermal image based on specific modal features, and make effective use of the shared information and specific information of each sample. In addition, we also consider the visual sense of human observing the scene image, that is, we will first pay attention to whether there will be someone in a part or area of the image, and then follow a series of steps to complete the density estimation processing. Therefore, the design of region recognition is added to RGB information by referring to [18]. The overall network of cross-modal collaborative representation learning is mainly composed of five modules. Each module contains two modal specific streams, one modal shared stream, and a shared specific transformation module (SSTM). The RGB modality-specific stream is represented by blue and green squares, the modality-shared stream is represented by orange squares, and the thermal modal modality-specific stream is represented by purple squares. In the design of network structure, [1] takes the network design of [23] and [24] to complete the preliminary extraction of features and combines the theory of graph convolution to determine the similarity within and between modes. [8] mainly has two backbone networks, which are based on CSRNet [25] and BL [9], and it verifies the applicability of the method combined with different classical network models. Our network is based on VGG-19 [26]. VGGNet has smaller convolution kernel and pooled sampling domain, which will bring implicit regularization results and can obtain more control parameters of image features. There followed taking the first block as an example to illustrate the components in the block. Firstly, two streams take the RGB and thermal images as inputs to extract specific modal features separately, which retain the specific information of a single modality. The shared stream takes the zero tensor as the input and aggregates the information of specific modal features in layers. The design of region recognition likes the attention mechanism, which divides each pixel in the feature into crowd and background region through the generated coarse-grained attention map. The network structure of area recognition is C (512,3)–U–C (256,3)–U–C (128,3)–U–C (64,3)–C (1,3), where C represents convolution layer and U represents up sampling. The later SSTM connected by specific streams can determine the internal similarity between RGB mode and thermal mode, as well as the similarity between them, and propagate the shared features and specific features back between the two modes at the same time to achieve the dual purpose of making up for each specific information and enhancing the shared information. In addition to the above design, the five modules are connected in turn, and the multi-scale problem is considered from the overall structure and the interior of the module.

3.2 Supervised fusion feature extraction

Region-based two-stream feature extractor For the input RGB image X^{RGB} and thermal image X^{I}, the features obtained after multi-layer convolution calculation are represented by F. Then, F is transformed into multi-scale context information I through pyramid pooling. The unified calculation process of thermal modality stream and modality-shared stream is as follows:

$$I = \mathrm{Conv}_{1\times1}(P^1(F) \oplus P^2(F) \oplus P^3(F)) \quad (1)$$

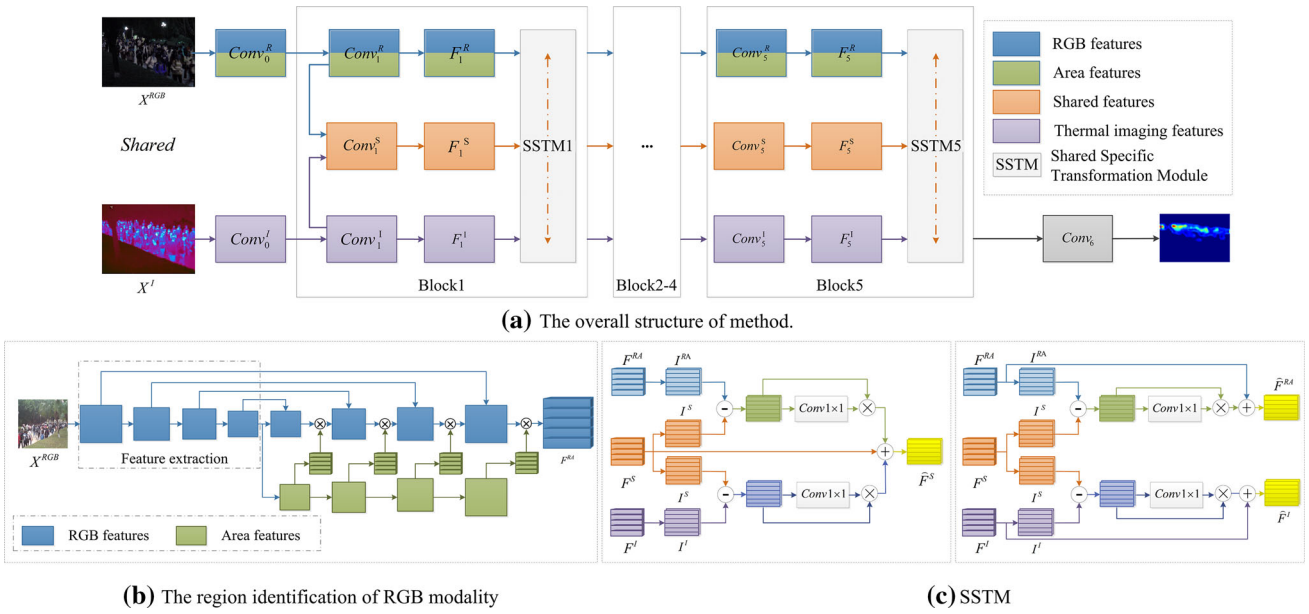


Fig. 1 **a** Overall structure of method. **b** Region identification of RGB modality. **c** Shared specific transformation module (SSTM)

where $\text{Conv}_{1\times1}$ expresses 1×1 convolutional layer and P represents the maximum pool layer with different size. The output sizes of $P(F)$ are 1, 1/2 and 1/4 of the original input size.

For the RGB modal stream, it is necessary to calculate the regional information while extracting the crowd characteristics. The region map generated by region recognition gives different weights to the pixels in different regions of the image. It is similar to the attention mechanism, which is mainly realized by multiple full connections and sigmoid functions, so that the transformation of the feature information of the part of interest can be highlighted. The updated feature F^{RA} is obtained by multiplying the crowd feature and the regional information F^A and then adding it to the original feature. The calculation process is as follows:

$$I^R = F^{R'} \oplus F^{R''} + F^{R'} \tag{2}$$

Shared-specific feature transfer (SSTM) As shown in Fig. 1, the residual calculation and gate function are carried out for I^{RA}, I^S and I^I obtained above to complete the aggregation and distribution of information. First, the residual information between the three is obtained, and then, complementary information is propagated to refine modality-shared features F^S, through 1×1 convolution adaptively. The calculation formula of enhanced features $\hat{F}^S$ is:

$$\text{FRS} = \text{Conv}_{1\times1}(I^{RA} - I^S) \tag{3}$$
$$\text{FIS} = \text{Conv}_{1\times1}(I^I - I^S) \tag{4}$$
$$\hat{F}^S = F^S + (I^{RA} - I^S) \otimes \text{FRS} + (I^I - I^S) \otimes \text{FIS} \tag{5}$$

Next, new modalities are assigned to share feature information, and the specific features of each modality are refined, respectively. The context information $\hat{I}^S$ corresponding to the enhancement feature $\hat{F}^S$ is dynamically propagated into $\hat{F}^R$ and $\hat{F}^I$. $\hat{F}^R$ and $\hat{F}^I$ are calculated as follows:

$$\hat{F}^R = F^R + (\hat{I}^S - I^{RA}) \otimes \text{Conv}_{1\times1}(\hat{I}^S - I^{RA}) \tag{6}$$
$$\hat{F}^I = F^I + (\hat{I}^S - I^I) \otimes \text{Conv}_{1\times1}(\hat{I}^S - I^I) \tag{7}$$

$\hat{F}^{RA}$, $\hat{F}^S$, and $\hat{F}^I$ complete the representation learning through five blocks in turn. Finally, the estimated density map is obtained through multiple convolution layers, and then, the density map is summed pixel by pixel to obtain the estimated counting.

4 Experiments and results

In this chapter, we will give the evaluation indicators and experimental details based on the newly proposed RGBT-CC benchmark. The training and evaluation are performed on Intel Core i7-7800 @ 3.50GHz processor and 31.1G memory. Experimental environment adopts PyTorch [27] framework and Adam [28] optimizer. At the same time, we use and start with training each model for 400 epochs and set the learning rate to 1e-5.

4.1 Evaluation metrics

Referring to the previous work, we adopt the Grid Average Mean Absolute Error (GAME [29]) and Mean Squared Error

(MSE) as the evaluation indicators of performance. Compared with Mean Absolute Error (MAE), GAME not only evaluates the overall area, but also includes the evaluation of areas with different sizes. The calculation formula of GAME is:

$$\mathrm{GAME}(l) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{4^l}|\mathrm{CE}_i^j - \mathrm{CG}_i^j| \tag{8}$$

where N represents the number of images, and l is a specific level. The image can be divided into $4l$ nonoverlapping regions according to l, and the corresponding regional error measurement can be carried out. The values of l are 0, 1, 2, and 3, respectively. When the value of l is 0, the value of GAME is the same as that of MAE. CE_i^j and CG_i^j represent the estimated count of the lth region and the corresponding ground-truth count, respectively. And the definition of MSE is as follows:

$$\mathrm{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathrm{CE}_i - \mathrm{CG}_i)^2} \tag{9}$$

where CG_i is the ground truth of testing image and CE_i is the corresponding estimation.

4.2 Loss function

The published datasets used for training generally provide point annotation for each training image. Many counting methods first use Gaussian kernel to convert the point annotation of each training image into ground truth map and then train the depth neural network model by regression calculation of each pixel's value in the density map. In contrast, the Bayesian loss function [9] adopted in this paper constructs a density contribution probability model from the perspective of point annotation and then calculates the expectation of each annotation point by summing the product of the contribution probability and the estimated density of each pixel. The calculation formula of loss function is:

$$L = \sum_{n=1}^{N} F(1 - E[C_n]) \tag{10}$$

where $F(\cdot)$ is distance function. The ground truth C_n of each annotation point is 1, and $E[_n]$ is the expectation of C_n. Compared with the loss function that limits the density value of each pixel, the Bayesian loss function monitors the count expectation of each annotation point.

4.3 Performance on comparison

RGBT-CC benchmark [8] contains 2030 pairs of RGB and thermal images from different scenes, in which 1013 pairs are from bright scenes, 1017 pairs are from dark scenes, and there are 138,389 marked pedestrians. The size of all images used for training and testing is uniformly set to 640 × 480. A total of 1030 pairs were used for training, 200 pairs were used for verification, and the remaining 800 pairs were used for testing.

Experimental results Table 1 shows the comparison between our method and other methods on RGBT-CC benchmark [8]. Each method in the table considers capturing enough scene details in various ways to complete the task of recognition and counting. HDFNet [30] and BBSNet [7] fully integrate and make use of cross-modal information (RGB image with depth optical information) to facilitate the task of target detection. The multi-view crowd counting proposed by MVMS [31] uses the information from multiple camera views to predict the scene level density map on the 3D world ground plane. Compared with CSRNet [25] and BL [9], CSRNet + IADM [8] and BL + IADM [8] combine RGB information and thermal imaging information for density map estimation. Both use the cross-modal collaborative representation learning framework to fully capture the complementary information of different modalities. Besides cross-modal collaborative representation learning, we add a multi-level supervision mechanism, which not only integrates multi-modal information, but also considers the extraction of coarse and fine-grained features. From the results, our method is better than other methods in the values of GAME and MSE.

Comparison of different condition of illumination Table 2 shows the testing comparison between our method and BL + IADM [8] on the bright and dark images given in [8]. And the table also shows the comparison of test results using RGB information, thermal imaging information, and RGB-T information, respectively. It can be seen that optical and thermal imaging information can complement each other, multi-modal information can obtain more detailed features than single-modal information, and thermal image can help to extract crowd features. Moreover, both our method and BL + IADM [8] improve the quality of density map by using the knowledge characteristics of comprehensive information and obtain more accurate estimation results. Meanwhile, it can achieve better results in brightness and darkness. In general, our method is better than BL + IADM [8] in MAE and MSE under the different condition of illumination.

4.4 Ablation experiments

In this part, we will conduct further experimental comparison and discuss the details of model design. At the same time, the effect of cross-scene is also preliminarily considered.

Table 1 Results comparison of different methods on the RGBT-CC benchmark

Method	GAME(0)	GAME(1)	GAME(2)	GAME(3)	MSE
HDFNet [30]	22.36	27.79	33.68	42.48	33.93
CSRNet [25]	20.40	23.58	28.03	35.51	35.26
MVMS [31]	19.97	25.10	31.02	38.91	33.97
BBSNet [7]	19.56	25.07	31.25	39.24	32.48
BL [9]	18.70	22.55	26.83	34.62	32.67
[25] + IADM [8]	17.94	21.44	26.17	33.33	30.91
[9] + IADM [8]	15.61	19.95	24.69	32.89	28.18
Our method	14.10	18.71	23.42	31.81	23.96

Table 2 Performance of different methods on the RGBT-CC benchmark under different illumination conditions

Illumination	Method	Input data	GAME(0)	GAME(1)	GAME(2)	GAME(3)	MSE
Brightness	BL + IADM [8]	RGB	41.01	56.22	68.07	80.22	63.59
	BL + IADM [8]	T	51.52	52.36	54.10	56.90	73.93
	BL + IADM [8]	RGBT	20.04	23.77	29.12	37.31	33.26
	Our method	RGB	45.48	51.90	58.44	65.69	72.87
	Our method	T	55.93	56.75	58.35	61.19	77.79
	Our method	RGBT	17.47	22.11	28.15	37.43	31.00
Darkness	BL + IADM [8]	RGB	59.58	69.94	71.38	85.60	94.7
	BL + IADM [8]	T	48.80	50.97	52.42	55.61	81.22
	BL + IADM [8]	RGBT	17.43	22.24	26.58	34.35	32.14
	Our method	RGB	73.67	85.36	97.79	107.53	94.16
	Our method	T	35.50	39.25	42.76	48.58	58.57
	Our method	RGBT	16.96	22.14	26.48	32.70	31.17

Architecture learning Besides comparing with the references, we also consider different schemes in the design of the network structure. The results of various schemes are compared as shown in Fig. 2. BL [9] mainly adopts VGG-19 network, and BL + IADM [8] adds cross-modal cooperative representation learning on the basis of BL [9]. Moreover, BL + IADM + TSF tries to take crowd region recognition as a specific stream to participate in the subsequent calculation process. While BL + IADM + RS1 adds region recognition for the extraction of RGB and thermal imaging feature, BL + IADM + RS2 only recognizes the region of RGB features. From the curve changing trends in the chart, BL + IADM + RS2 is lower than other schemes in GAME value and MSE value. The results show that more detailed crowd features can be obtained through the comprehensive knowledge extraction of cross-modal features, which is helpful to improve the accuracy of crowd counting. More than two modal representations are difficult to learn, while the coarse- and fine-grained feature extraction of multi-modal features is helpful to obtain high-quality density map.

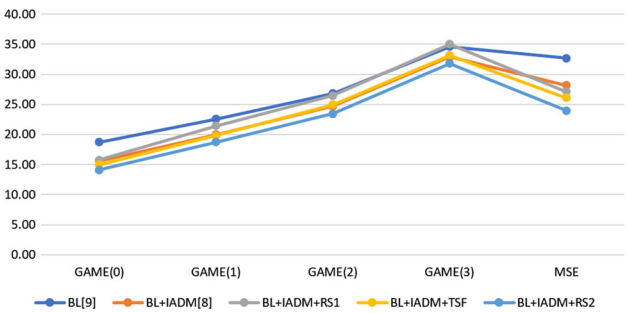
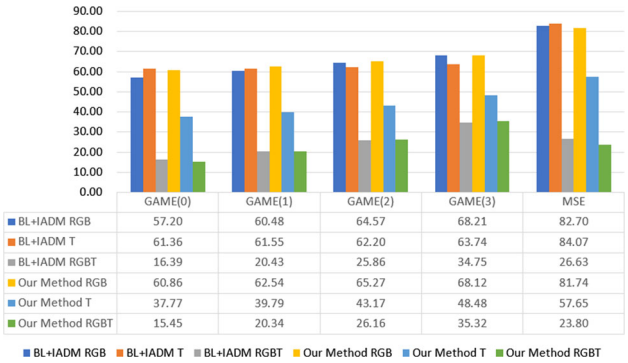


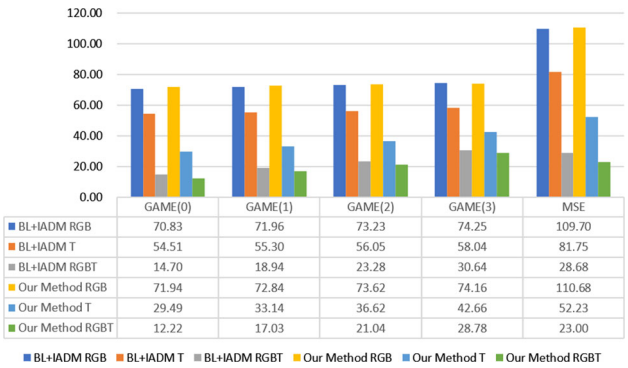
Fig. 2 Results comparison of different structural schemes

Effect of cross-scene In addition to considering the training and testing on different illumination, we also preliminarily verify the effect of cross-scene. Firstly, the model is trained on the whole dataset and then tested on the brightness and darkness sets divided by BL + IADM [8]. As seen from the data distribution in Figs. 3 and 4, the effect of using RGB-T multi-modal information is better than using RGB or T feature information alone. When using RGB information alone to extract features, our method is not as effective as BL + IADM [8]. But when T information or RGB-T information can be used to extract features, our method is better than BL + IADM [8], and the MAE and MSE values are increased by 38%, 31% and 6%, 11%, respectively.



	GAME(0)	GAME(1)	GAME(2)	GAME(3)	MSE
BL+IADM RGB	57.20	60.48	64.57	68.21	82.70
BL+IADM T	61.36	61.55	62.20	63.74	84.07
BL+IADM RGBT	16.39	20.43	25.86	34.75	26.63
Our Method RGB	60.86	62.54	65.27	68.12	81.74
Our Method T	37.77	39.79	43.17	48.48	57.65
Our Method RGBT	15.45	20.34	26.16	35.32	23.80

Fig. 3 Comparison of cross-scene testing and the illumination condition is brightness



	GAME(0)	GAME(1)	GAME(2)	GAME(3)	MSE
BL+IADM RGB	70.83	71.96	73.23	74.25	109.70
BL+IADM T	54.51	55.30	56.05	58.04	81.75
BL+IADM RGBT	14.70	18.94	23.28	30.64	28.68
Our Method RGB	71.94	72.84	73.62	74.16	110.68
Our Method T	29.49	33.14	36.62	42.66	52.23
Our Method RGBT	12.22	17.03	21.04	28.78	23.00

Fig. 4 Comparison of cross-scene testing and the illumination condition is darkness

5 Conclusion

We propose a method based on coordinated representation and multi-level supervision for the estimated density map and crowd counting in the unconstrained crowd scene. The whole network includes five blocks and density map generator. Pairs of RGB-thermal images are first input into the two-stream feature extractor to obtain shared features and specific features, in which the RGB stream contains both crowd features and regional information. Then, the multi-modal features determine the similarity within and between modalities through SSTM module and transmit shared and specific features between modalities. In addition, the five blocks and the later density map generator extract multi-scale global and local features and form a three-level supervision mechanism of point, region, and multi-scale. Meanwhile, the multi-scale feature map between multiple modules is combined with region recognition and combined with residual calculation to achieve the purpose of adaptive attention to different regions and extracting different detail features. Finally, the estimated density map is generated through the density map generator. We conduct experiments on the RGBT-CC benchmark to verify the effectiveness of the method. And we will further consider the application of unsupervised method in crowd counting in the future work.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant Nos. 61771420 and 62001413, the National Natural Science Foundation of Hebei Province under Grant No. F2020203064, as well as the China Postdoctoral Science Foundation under Grant No. 2018M641674, the Doctoral Foundation in Yanshan University under Grant No. BL18033 and Science and Technology Research and Development Program of Qinhuangdao under Grant No. 202101A004.

References

1. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: IEEE Conference on Computer Vision and Pattern Recognition, 14–19 June 2020, Seattle, WA, USA, pp. 13379–13389. IEEE (2020)
2. Lian, D., Li, J., Zheng, J., Luo, W., Gao, S.: Density map regression guided detection network for RGB-D crowd counting and localization. In: IEEE Conference on Computer Vision and Pattern Recognition, 16–20 June 2019, Long Beach, CA, USA, pp. 1821–1830. IEEE (2019)
3. Zhai, Y., Fan, D., Yang, J., Borji, A., Shao, L., Han, J., Wang, L.: Bifurcated backbone strategy for RGB-D salient object detection. IEEE Trans. Image Process. **30**, 8727–8742 (2021)
4. Finn, C., Abbeel, P., Levine, S.: Model-agnostic meta-learning for fast adaptation of deep networks. In: IEEE International Conference on Machine Learning, 18–21 Dec 2017, Taipei, Taiwan, pp. 1126–1135. IEEE (2017)
5. Zhou, D., He, Q.: Cascaded multi-task learning of head segmentation and density regression for RGBD crowd counting. IEEE Access **8**, 101616–101627 (2020)
6. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-Induced Multi-Scale Recurrent Attention Network for Saliency Detection. In: IEEE/CVF International Conference on Computer Vision, 20–26 Oct 2019, Seoul, South Korea, pp. 7253–7262. IEEE (2019)
7. Fan, D., Zhai, Yi., Borji, A., Yang, J., Shao, L.: Bbs-net: RGB-D salient object detection with a bifurcated backbone strategy network. In: European Conference on Computer Vision, 23–28 Aug 2020 (2020)
8. Liu, L., Chen, J., Wu, H., Li, G., Li, C., Lin, L.: Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting. In: IEEE International Conference on Computer Vision, 20–25 June 2021, Nashville, TN, USA. IEEE (2021)
9. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: IEEE International Conference on Computer Vision, 20–26 Oct 2019, South Korea, pp. 6142-6151. IEEE (2019)
10. Liu, W., Salzmann, M., Fua, P.: Context-aware crowd counting. In: IEEE Conference on Computer Vision and Pattern Recognition, 16–20 June 2019, Long Beach, CA, USA, pp. 5099-5108. IEEE (2019)
11. Jiang, X., Xiao, Z., Zhang, B., Zhen, X., Cao, X., Doermann, D., Shao, L.: Crowd counting and density estimation by trellis encoder–decoder networks. In: IEEE Conference on Computer Vision and Pattern Recognition, 16–20 June 2019, Long Beach, CA, USA, pp. 6133–6142. IEEE (2019)
12. Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L.: Crowd counting with deep structured scale integration network. In: IEEE International Conference on Computer Vision, 20–26 Oct 2019, South Korea, pp. 1774–1783. IEEE (2019)

13. Liu, N., Long, Y., Zou, C., Niu, Q., Pan, L., Wu, H.: Adcrowdnet: an attention-injective deformable convolutional network for crowd understanding. In: IEEE Conference on Computer Vision and Pattern Recognition, 16–20 June 2019, Long Beach, CA, USA, pp. 3225–3234. IEEE (2019)
14. Dong, Z., Zhang, R., Shao, X., Li, Y.: Scale-recursive network with point supervision for crowd scene analysis. Neurocomputing **384**, 314–324 (2019)
15. Liu, J., Gao, C., Meng, D., Hauptmann, A.G.: Decidenet: counting varying density crowds through attention guided detection and density estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, 18–22 June 2018, Salt Lake, UT, USA, pp. 5197–5206. IEEE (2018)
16. Liu, Y., Shi, M., Zhao, Q., Wang, X.: Point in, box out: beyond counting persons in crowds, In: IEEE Conference on Computer Vision and Pattern Recognition, 16–20 June 2019, Long Beach, CA, USA, pp. 6469–6478. IEEE (2019)
17. Liu, C., Weng, X., Mu, Y.: Recurrent attentive zooming for joint crowd counting and precise localization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 16–20 June 2019, Long Beach, CA, USA, pp. 1217–1226. IEEE (2019)
18. Sam, D.B., Peri, S.V., Sundararaman, M.N., Kamath, A., Radhakrishnan, V.B.: Locate, size and count: accurately resolving people in dense crowds via detection. IEEE Trans. Pattern Anal. Mach. Intell. **43**, 2379–2751 (2020)
19. Fu, K., Fan, D., Ji, G., Zhao, Q.: Jldcf: joint learning and densely-cooperative fusion frame work for RGB-D salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, 14–19 June 2020, Seattle, WA, USA, pp. 3052–3062. IEEE (2020)
20. Sun, T., Di, Z., Che, P., Liu, C., Wang, Y.: Leveraging crowd sourced GPS data for road extraction from aerial imagery. In: IEEE Conference on Computer Vision and Pattern Recognition, 14–19 June 2020, Long Beach, CA, USA, pp. 7509–7518. IEEE (2020)
21. Piao, Y., Rong, Z., Zhang, M., Ren, W., Lu, H.: A2dele: adaptive and attentive depth distiller for efficient RGB-D salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, 14–19 June 2020, Seattle, WA, USA, pp. 9060-9069. IEEE (2020)
22. Zhao, J., Cao, Y., Fan, D., Cheng, M., Li, X., Zhang, L.: Contrast prior and fluid pyramid integration for rgbd salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, 16-20 June 2019, Long Beach, CA, USA, pp. 3927–3936. IEEE (2019)
23. Ye, M., Lan, X., Li, J., Yuen, P.C.: Hierarchical discriminative learning for visible thermal person re-identifification. In: Thirty-Second AAAI Conference on Artificial Intelligence, 2–7 Feb 2018, Hilton New Orleans Riverside, New Orleans, LO, USA. IEEE (2018)
24. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: International Joint Conference on Artificial Intelligence, 13–19 July 2018, Stockholm, Sweden, pp. 677–683. IJCAI (2018)
25. Li, Y., Zhang, X., Chen, D.: Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In: IEEE Conference on Computer Vision and Pattern Recognition, 18–22 June (2018), Salt Lake City, UT, USA, pp. 1091–1100. IEEE (2018)
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. (2014). arXiv preprint arXiv:1409.1556
27. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L.: Pytorch: an imperative style, high-performance deep learning library. In: Neural Information Processing Systems, 3-6 Dec 2018, Montréal, Canada, pp. 8026–8037. NIPS (2018)
28. Kingma, D., Ba, J.: Adam: a method for stochastic optimization (2014). arXiv:1412.6980
29. Guerrero-Gómez-Olmedo, R., Torre-Jiménez, B., López-Sastre, R., Maldonado-Bascón, S., Oñoro-Rubio, D.: Extremely overlapping vehicle counting. In: Iberian Conference on Pattern Recognition and Image Analysis, 17–19 June 2015, Santiago de Compostela, Spain, pp. 423–431 (2015)
30. Pang, Y., Zhang, L., Zhao, X., Lu, H.: Hierarchical dynamic filtering network for RGB-D salient object detection. In: European Conference on Computer Vision, 23–28 Aug 2020 (2020)
31. Zhang, Q., Chan, A.: Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In: IEEE Conference on Computer Vision and Pattern Recognition, 16–20 June 2019, Long Beach, CA, pp. 8297–8306. IEEE (2019)