

Late better than early: A decision-level information fusion approach for RGB-Thermal crowd counting with illumination awareness

Jian Cheng, Chen Feng, Yang Xiao, Zhiguo Cao *

Key Laboratory of Image Processing and Intelligent Control, Ministry of Education; School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Communicated by D. Liu

Keywords:

Crowd counting
RGB-T image
Decision-level fusion

ABSTRACT

In this paper, we make the first research effort to address the RGB-Thermal (RGB-T) crowd counting problem with the decision-level late fusion manner. Being different from the existing pixel-level or feature-level fusion methods, our proposition chooses to fuse the density maps yielded by RGB and thermal counterparts via spatially adaptive weighting with RGB illumination-aware attention. Our key intuition to conduct RGB-T density map fusion lies in 2 main folders. First, compared with the raw RGB-T images or convolutional feature maps, RGB-T density maps contain stronger counting-wise semantic meanings. Secondly, they are also of high spatial resolution for revealing fine local details. To fuse them adaptively, a spatial weighting map for each modality, together with an illumination-related RGB weight is generated. In this way, the issues of RGB illumination awareness and local counting pattern characterization ability are concerned jointly. To the best of our knowledge, we are the first to leverage RGB-T crowd counting concerning these 2 issues in a unified way. Meanwhile, cross-modality feature interaction is conducted between RGB and thermal modalities to facilitate spatial weighting map generation. The experiments on 2 well-established RGB-T crowd counting datasets (i.e., RGBT-CC and DroneRGBT) verify the superiority of our proposition. The source code and pretrained models will be released upon acceptance at <https://github.com/hustaia/DLF-IA>.

1. Introduction

Crowd counting aims to estimate the number of persons in unconstrained scenes. As an important subtask of visual reasoning [1–4], it has a wide range of applications in public safety [5–8]. In recent years, with the rapid growth of representation capabilities of deep learning technology, much progress has been made on the research topic of RGB crowd counting for its low cost and wide application. However, under poor light conditions, as the RGB images become obscure even invisible, the performance of RGB crowd counting degrades significantly [9,10]. To tackle this issue, researchers resort to thermal (T) data to complement the RGB images [11,12]. The thermal images are good at distinguishing the crowd from the cluttered backgrounds, even when the RGB ones are uninformative. The RGB images can help eliminate the false positives in the thermal ones caused by other heating objects, such as walls and lamps. Therefore, RGB-T crowd counting may provide more robust solutions in practical scenarios. One key issue of RGB-T crowd counting is how to effectively fuse the multi-modality information [13–15].

The fusion strategies of existing works are conducted on either pixel level or feature level. Pixel-level fusion methods [16,17] (Fig. 1(a)) first fuse RGB and thermal images into one input, and then conduct crowd counting via single-modality methods. Albeit simple and efficient, they

are susceptible to noisy patterns, such as other heating objects in thermal images, and objects with similar color in RGB images. The noisy patterns are indistinguishable from the crowd at the image level, and they hinder the method from effectively capturing the descriptive patterns present in both RGB and thermal modalities, which are crucial for crowd counting. Consequently, the complementarity and descriptive clues between the two modalities are lost prematurely. Feature-level fusion approaches [9,11,18–21] (Fig. 1(b)) first extract features of RGB and thermal inputs, and then perform feature fusion to obtain multi-modality feature, which is finally fed to the decoder to predict the final result. With improved semantic capacity and enhanced resistance to noisy patterns, these methods alleviate the problem of pixel-level fusion to some extent. However, they conduct information fusion on the convolutional feature maps that are generally of low spatial resolution. For crowd counting, since both big and small heads may appear in unconstrained scenarios, both high-level semantic meaning and fine local details are required. Therefore, feature fusion-based methods are likely to lose fine local details and may not predict high-quality results in dense regions.

To address the above problems, we propose to fuse the multi-modality information at the decision level (Fig. 1(c)) in RGB-T crowd

* Corresponding author.

E-mail addresses: jian_cheng@hust.edu.cn (J. Cheng), chen_feng@hust.edu.cn (C. Feng), Yang.Xiao@hust.edu.cn (Y. Xiao), zgcao@hust.edu.cn (Z. Cao).

<https://doi.org/10.1016/j.neucom.2024.127888>

Received 14 February 2024; Received in revised form 20 April 2024; Accepted 13 May 2024

Available online 17 May 2024

0925-2312/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

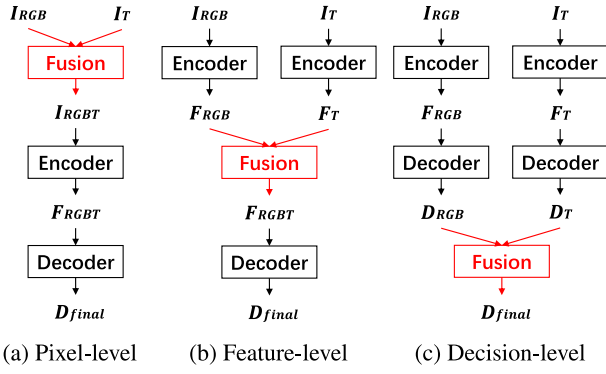


Fig. 1. Three fusion strategies for RGB-T crowd counting. “I”, “F” and “D” denote image, feature map and density map, respectively. Our proposed decision-level fusion approach leverages the predicted RGB-T density maps that retain high semantic clue and fine local details.

counting. We find that the predicted density maps yielded by the RGB and the thermal counterparts are suitable for information fusion for the following 2 reasons. First, compared with images, density maps contain more counting-wise semantic clues. They are less vulnerable to the noisy patterns at the image level, such as other heating objects resembling crowds in thermal images, and objects with similar color to the crowd in RGB images. This helps the model capture the descriptive information from the two modalities. Secondly, compared with feature maps, due to the upsample operation in the decoding process, density maps are naturally of higher spatial resolution, so they reveal finer local details of small objects. The local details are important for crowd counting, since crowd counting models tend to predict more erroneous results on regions containing small heads than on the sparse regions.

Specifically, to characterize local counting patterns, each modality is assigned a spatial weighting map, and the fusion is done via weighted summation of the two predicted density maps. Besides, since the illumination condition has a large effect on the performance of the RGB modality [11,12,19,21], in the process of spatial weighting map generation, RGB illumination awareness is concerned via an illumination-aware RGB weighting mechanism. To make the illumination-related weight more flexible, we design a learnable Sigmoid function to map the illumination value of the RGB input to illumination-aware attention. Besides, a cross-modality feature interaction module is conducted between the RGB and thermal features to further leverage their complementarity and enhance spatial weighting.

Experimental results on 2 RGB-T crowd counting benchmarks (*i.e.*, RGBT-CC [11] and DroneRGBT [12]) show that our approach achieves very competitive results compared with state-of-the-art methods. Further ablation studies verify the effectiveness of the proposed components, including decision-level fusion strategy, illumination-aware RGB weighting mechanism, and cross-modality feature interaction. In summary, the main contributions of this paper can be summarized as follows:

- We propose the first decision-level fusion method via density map fusion towards RGB-T crowd counting;
- An illumination-aware RGB weighting mechanism is introduced, so that the issues of RGB illumination awareness and local counting pattern characterization ability are concerned jointly;
- Cross-modality feature interaction is conducted between RGB-T features to facilitate the generation of the spatial weighting map.

The remainder of this paper is organized as follows. Section 2 discusses the related works. The proposed approach is detailed in Section 3. Section 4 presents the experimental configurations. Experiments and analyses are reported in Section 5. Finally, we draw the conclusion in Section 6.

2. Related works

In this section, the related works will be introduced in terms of RGB crowd counting, RGB-T crowd counting and decision-level fusion methods.

2.1. RGB crowd counting

According to how the problem is formulated, existing crowd counting methods can be divided into two categories: localization-based methods and density map regression-based approaches. Localization-based approaches perform counting by predicting the location of each person. The localization can be obtained by direct object detection [22, 23], peak point detection [24,25], foreground segmentation [26,27] and query-based regression [28,29]. Density map regression-based methods formulate counting as a dense prediction task and learn to regress a density map [30]. The count can be obtained by summing over the predicted density map. In this paradigm, network architectures [31,32], loss functions [33,34], multi-scale strategies [35,36], attention mechanisms [37,38], data augmentations [39,40] and the usage of reinforcement learning [41] and diffusion models [42] are explored. Readers can refer to [43,44] for more comprehensive surveys.

Although RGB crowd counting can achieve good results in ideal circumstances, the visual features extracted from the RGB modality are sensitive to illumination variation and are easily affected by the cluttered background. The performance of RGB crowd counting methods will degrade when the quality of the RGB inputs becomes worse [9,10]. To address this problem, the thermal images are introduced to complement the RGB images to enhance crowd counting in unconstrained scenarios [11,12].

2.2. RGB-T crowd counting

With the publicity of RGBT-CC [11] and DroneRGBT [12] benchmarks, RGB-T crowd counting has become an active research topic. The major concern is how to fuse the complementary information from RGB and thermal modalities.

Most existing methods specifically designed for RGB-T crowd counting adopt a feature-level fusion strategy. In this paradigm, the multi-modal feature is obtained via feature interaction and fusion during the encoding stage. After that, the decoder converts the multi-modal feature to the final density map. For example, CMCR [11], TAFNet [18] and MC³Net [45] adopt a three-stream framework to learn the RGB feature, the thermal feature, and the modality-shared or concatenated RGB-T feature. MAT [20], CSCA [19] and Liu et al. [9] use attention or Transformer blocks to build long-range dependencies between multi-modal inputs. MMCCN [12] proposes a network whose pipeline contains multi-scale feature learning, multi-modal feature alignment, and adaptive feature fusion. DEFNet [21] uses multi-modal fusion, receptive field enhancement, and multi-layer fusion to highlight the crowd position and suppress the background noise. Apart from the above methods, to make better use of the complementarity of RGB and thermal inputs, some works adopt asymmetric architectures for the two modalities. For example, R2T [46] first learns the cross-modal feature representation through feature mapping, and then incorporates the learned representation to yield the final counting result. CGINet [47] uses RGB features to supplement details for the thermal branch and transfers thermal features to enrich information for the RGB branch. DHRNet [10] adopts a thermal-main RGB-auxiliary strategy and aggregates the multi-modal feature through a cross-modal fusion module. Being effective, the feature-level fusion methods become the mainstream solution for RGB-T crowd counting. However, information fusion is generally conducted in a low spatial resolution, and these methods are likely to lose fine local details which is important for crowd counting.

Another solution is to fuse the RGB-T data at pixel level [16,17]. The task-agnostic image fusion is performed on the RGB and the thermal inputs. Then a single-modality crowd counter is adopted to convert the fused image to the final result. Albeit simple and efficient, these methods are susceptible to noisy patterns in the input data.

Being different from previous works, our decision-level fusion strategy fuses the predicted RGB-T density maps that retain high resolution, semantic clue, and complementarity. In this way, the complementary information of the RGB-T data is utilized more effectively. What is more, no existing methods have considered RGB illumination. By contrast, our illumination-aware attention concerns the issues of RGB illumination variation and local counting pattern characterization ability simultaneously.

2.3. Decision-level fusion methods

Apart from the most commonly used feature-level fusion paradigm, the decision-level fusion strategy is also adopted in other multi-modal tasks. For example, for RGB-T visual object tracking, based on the observation that the response scores of RGB and TIR data are biased, DFAT [48] proposes a decision-level fusion strategy that focuses on suppressing the bias between two modalities. For multispectral pedestrian detection, in LG-FAPF [49], two spatial locality maps are predicted by visible and thermal branches, and they are then utilized to provide prediction confidence scores in both channels. For cross-modal 3D object detection, CLOCs [50] operates on the combined output candidates of camera- and LiDAR-based detectors, and the geometric and semantic consistencies in bi-modal predictions are leveraged to produce final detection results. For RGB-T salient object detection, Xu et al. [51] obtain visual and thermal saliency maps via a two-stream encoder-decoder network and achieve the final prediction by adding the reconstructed thermal saliency map to the visual one.

Our approach extends the idea of decision-level fusion to the task of RGB-T crowd counting, in which pixel-wise crowd density value needs to be estimated in unconstrained scenarios. Besides, the above methods overlook the effect of illumination, while RGB illumination awareness and local counting pattern characterization ability are concerned jointly in our method.

3. Method

In this section, we first present the motivation behind fusing density maps. Then we sketch the proposed density-level fusion approach with illumination awareness (DLF-IA). Next, we introduce each component: decision-level fusion strategy, illumination-aware RGB weighting mechanism, and cross-modality feature interaction. Finally, loss function is described.

3.1. Motivation behind fusing density maps

In this subsection, we validate the two key advantages of density maps. First, density maps contain stronger counting-wise semantic meanings than images, so they tend to be less susceptible to the noisy patterns of the input images. Secondly, due to the upsample operation at the decoding stage, density maps are naturally of higher resolution than feature maps, so they tend to reveal finer local details.

To verify the above advantages of density maps, in Fig. 2, we build three models based on pixel, feature map, and density map fusion. (1) For the pixel-level fusion model, RGB and thermal inputs are first concatenated by channel and then fed to a single-modality crowd counter to obtain the final result. (2) Towards the decision-level fusion model, we use two parallel backbones to extract features from RGB and thermal inputs. Two density decoders are adopted to convert the extracted features to the predicted density maps of RGB and thermal counterparts. At the same time, the extracted features are fed to two weight decoders to obtain two spatial weighting maps

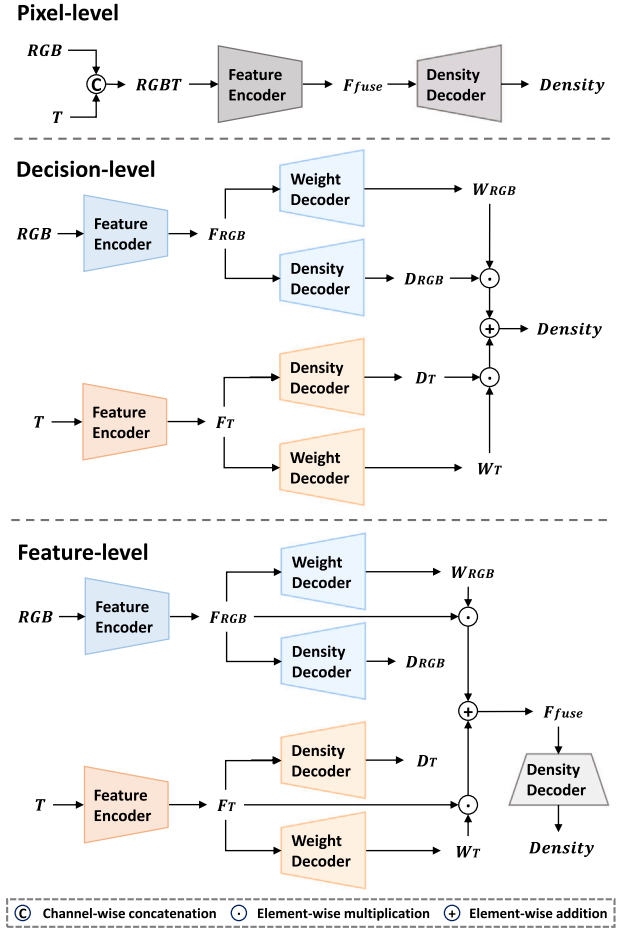


Fig. 2. The technical pipelines of three fusion models.

of RGB and thermal modalities. The final output is the element-wise weighted summation of RGB-T density maps. The ground truth annotations are employed to supervise the learning of RGB, thermal, and final predicted density maps. (3) Regarding the feature-level fusion model, after extracting the RGB and thermal features, two weight decoders are adopted to obtain the spatial weighting maps of RGB and thermal modalities, respectively. The RGB-T features are element-wise weighted and summed to predict the fused feature. A density decoder is employed to convert the fused feature to the final output. To ensure a fair comparison with the decision-level one, two density maps of RGB and thermal counterparts are also predicted. Similarly, the learning of three predicted density maps is supervised by the ground truth annotations. We train and evaluate the above models on the RGBT-CC [11] dataset. Mean absolute error (MAE) and root mean square error (RMSE) are adopted as the performance indicators (Section 4.2). For both metrics, a lower value means better performance.

First, to illustrate that density maps tend to be less susceptible to the noisy patterns than images, we degrade the test RGB and thermal images by adding Gaussian noise with a mean value of 0 and a variation of σ , as shown in Fig. 3. Then the decision- and pixel-level fusion models are evaluated on the degraded test samples. As in Table 1, even when the inputs are severely degraded, the decision-level fusion model can still achieve satisfactory results. It still holds the ability to count people. Contrastively, the pixel-level fusion model fails to tackle severely degraded inputs. In practical applications, noise is very common in the process of image acquisition and transmission. The results demonstrate that the decision-level fusion method is more robust to noisy data than the pixel-level one.

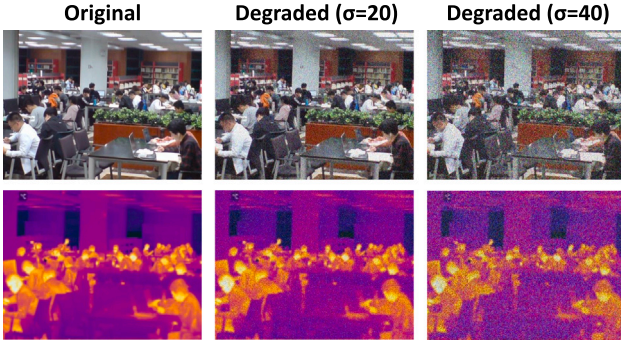


Fig. 3. A sample of RGB-T test images before and after Gaussian degradation with different σ .

Table 1

Performance comparison of decision- and pixel-level fusion models on samples of different degrees of degradation. “None” denotes that the model is evaluated on the original test samples.

Method	σ	MAE	RMSE
Decision-level	None	11.24	19.49
	10	17.71	31.61
	20	19.80	34.90
	40	21.73	36.73
	None	12.51	23.97
Pixel-level	10	25.20	35.62
	20	122.78	200.93
	40	393.18	546.16

Table 2

Performance comparison of decision- and feature-level fusion models on different test (sub)sets.

Test (sub)set	Method	MAE	RMSE
Full	Decision-level	11.24	19.49
	Feature-level	11.68	22.88
Large	Decision-level	5.30	7.51
	Feature-level	4.53	6.86
Small	Decision-level	22.81	41.32
	Feature-level	27.57	48.34

Secondly, to verify that density maps tend to reveal finer local details than feature maps, we manually select test samples that only contain large/small heads and build two subsets. Then the decision- and feature-level fusion models are evaluated on the test set and these two subsets. As in Table 2, compared to the feature-level fusion model, the decision-level one performs much better on the small subset. Meanwhile, the two models achieve comparable results on the large subset. The results reveal that the decision-level fusion method is more skilled at dealing with small heads than the feature-level one.

Based on the above discussion, we choose to fuse density maps at the decision level for RGB-T crowd counting.

3.2. Overview

The technical pipeline of DLF-IA is shown in Fig. 4. Similar to the decision-level fusion model in Fig. 2, RGB and thermal density maps are predicted in two parallel branches. At the same time, two spatial attention maps of RGB and thermal modalities, and an illumination-related RGB weight are generated to weight the RGB-T density maps. The final output is the element-wise weighted summation of RGB-T density maps, which is detailed in Section 3.3.

Besides, the illumination condition is concerned through the generation of illumination-related RGB weight. The image is first transformed from RGB space to the HSV color space. An illumination weight

function is then designed to convert the V channel of the HSV image to the illumination-related RGB weight. Section 3.4 details the illumination-aware RGB weighting mechanism.

In addition, to facilitate the generation of spatial weighting maps, a cross-modality feature interaction module is designed. The RGB and thermal features are interacted based on the attention mechanism to highlight the interested area, so that the complementary information of RGB-T data is leveraged. Section 3.5 presents the details of the cross-modality feature interaction module.

3.3. Decision-level fusion strategy

Through encoding and decoding in separate branches, two density maps of RGB and thermal modalities, D_{RGB} and D_T are obtained. Towards how to yield the mixing weights of two modalities, a simple way is to assign each modality with a confidence score. However, the reliability of prediction in different spatial locations may be variational. Hence, it is not enough to apply a single score. In this spirit, we assign the density map of each modality with a spatial weighting map. Specifically, the RGB and thermal features (after cross-modality feature interaction, Section 3.5) are input into two weight decoders to generate spatial attention maps W_{RGB}^{spa} and W_T^{spa} , respectively. D_{RGB} and W_{RGB}^{spa} are of the same size, so that they can be processed by the Hadamard product. Meanwhile, an illumination-related RGB weight u_{RGB}^{illu} (Section 3.4) is leveraged to further weight D_{RGB} . Note that W_{RGB}^{spa} and W_T^{spa} are attention maps, and u_{RGB}^{illu} is a scalar. The final density map D_{final} is obtained by

$$D_{final} = u_{RGB}^{illu} \cdot W_{RGB}^{spa} \odot D_{RGB} + W_T^{spa} \odot D_T, \quad (1)$$

where \odot is the Hadamard product.

3.4. Illumination-aware RGB weighting mechanism

Illumination condition has a large effect on the reliability of RGB modality. We first investigate the impact of illumination conditions on the performance of RGB and thermal modalities. Since the V channel of the HSV image is mostly related to the illumination, we convert the image from RGB space to HSV color space and split the V channel as the illumination map. The illumination map is then averaged to a value to reflect the overall brightness (denoted as “illumination value” for simplicity). When overexposure does not happen, a higher illumination value means better illumination conditions. Fig. 5 exhibits the mean relative error (MRE) of RGB and thermal predictions *w.r.t.* illumination condition of RGB image on RGBT-CC [11] dataset. It can be observed that, as the illumination condition becomes worse (illumination value declines), the MRE of RGB predictions increases greatly, while the MRE of thermal predictions remains stable. *I.e.*, the RGB density maps are unstable to illumination variation and unreliable under poor light conditions. Therefore, it may be beneficial to assign different importance to RGB density maps in various light conditions. Inspired by this, the illumination-aware RGB weighting mechanism is proposed.

Specifically, we first obtain the illumination value as in Fig. 5. Then, a Sigmoid-like illumination weight function (IWF) is adopted to convert the illumination value to the illumination-related RGB weight. Finally, the illumination-related RGB weight and the spatial weighting map of RGB modality are multiplied to obtain the illumination-aware RGB weighting map.

Here we describe the design of the Sigmoid-like IWF in details. A naive implementation of IWF is to map the illumination value to illumination-related weight linearly. However, the linear mapping function is not strong enough to adjust the importance of RGB modality. First, think about dark RGB images in which pedestrians are obscure, *e.g.*, the case in the first column of Fig. 6. Since thermal images provide more salient information about the crowd than RGB ones, a reasonable solution is to depend on the thermal modality to distinguish the crowd

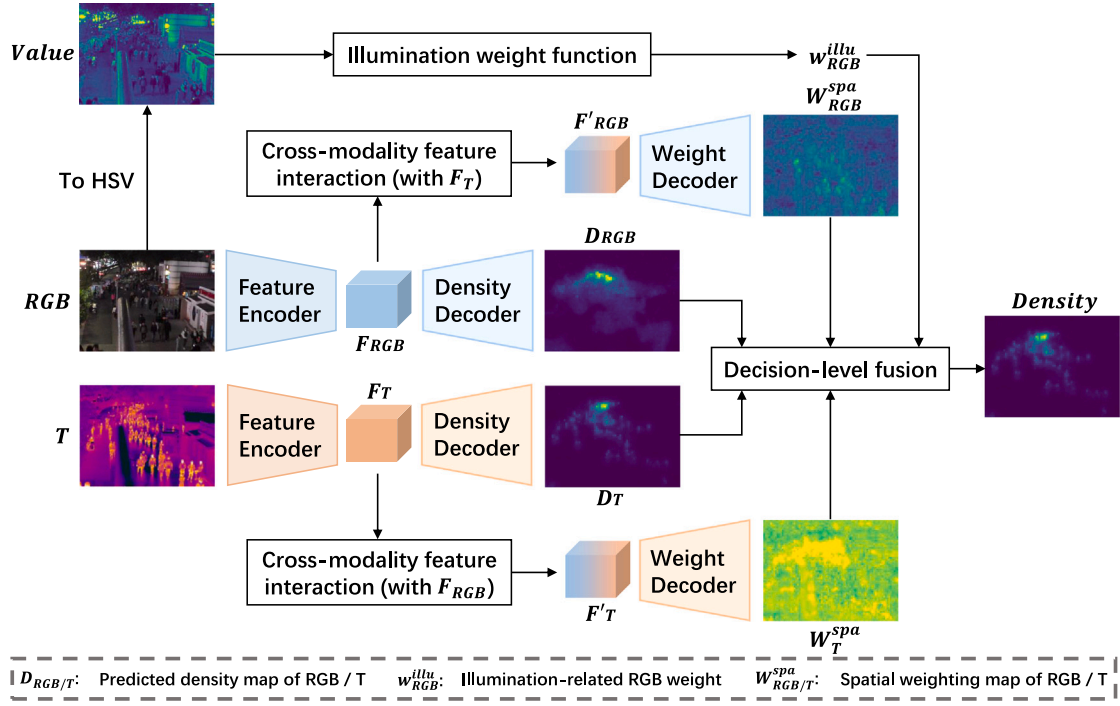


Fig. 4. The main technical pipeline. RGB and thermal inputs are fed into two separate branches to obtain two density maps. Then the density maps are fused with illumination-related RGB weight and RGB-T spatial weighting maps. Meanwhile, cross-modality feature interaction is applied to leverage the RGB-T complementarity and facilitate the generation of spatial weighting maps.

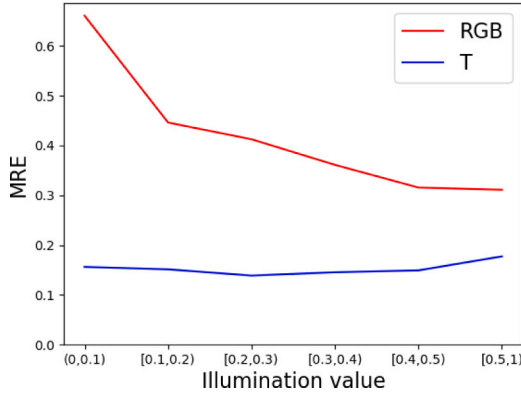


Fig. 5. Mean relative error (MRE) of the samples in different illumination conditions on RGBT-CC dataset.

and assign low importance to the RGB modality. Secondly, the illumination values of bright RGB images may not be so high, e.g., the cases in the second and third columns of Fig. 6. Since they provide rich color and texture information for distinguishing the pedestrians, it would be better not to let the weights weaken the impact of RGB modality. Thirdly, for RGB images whose illumination values are large enough (e.g., above 0.7), it is rational to assign them similar illumination-related weights, and vice versa. Since a modified Sigmoid function can meet the above requirements, to make the illumination-related weight more flexible, we design a learnable Sigmoid-like function to map the illumination value to the weight. Specifically, the illumination-related weight is calculated as

$$w_{RGB}^{illu} = \frac{1}{1 + \exp(-k \cdot (value - b))}, \quad (2)$$

where *value* is the illumination value; *k* and *b* are two learnable parameters; *k* controls the rate at which the weight changes with illumination value, and *b* is the illumination value at the center of symmetric of the

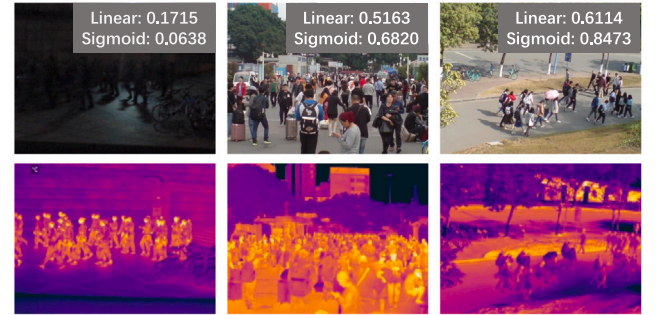


Fig. 6. Samples of RGB (top) and thermal (bottom) image pairs in RGBT-CC dataset. The numbers in the upper right corner of each RGB image are the illumination-related RGB weights calculated by linear and Sigmoid-like IWF, respectively.

function curve. Fig. 7 shows the illustration of linear and Sigmoid-like IWF.

3.5. Cross-modality feature interaction

As shown in Fig. 4, we reuse RGB and thermal features extracted by backbone encoders to generate the spatial weighting map. To leverage the complementary information of RGB and thermal features, a cross-modality feature interaction module is proposed.

Fig. 8 exhibits the detail of the cross-modality feature interaction module. Inspired by the Universal Representation Transformer (URT) layer of [52], the attention mechanism is applied during the feature interaction process. Specifically, RGB and T features are concatenated by channel, and they are then flattened and processed by a linear transformation to obtain the query vector:

$$Q = W^q \cdot Flatten(Concat([F_{RGB}, F_T])) + b^q, \quad (3)$$

where W^q and b^q denote the weight and bias of the query transformation function. Similarly, RGB or T feature is flattened and processed

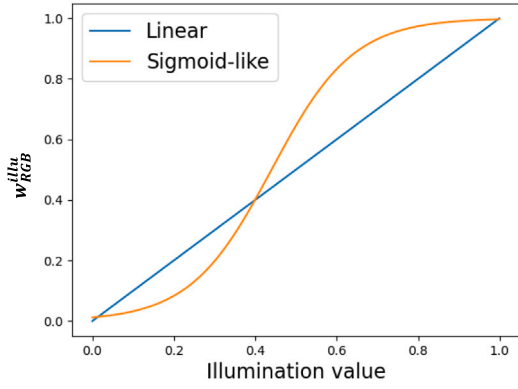


Fig. 7. The illustration of linear and Sigmoid-like IWF. Linear IWF assigns low weights to samples under poor illumination conditions. Compared with linear IWF, Sigmoid-like IWF further weakens the influence of dark samples and pays more attention to bright ones. This helps mitigate the effect of unreliable prediction on dark RGB images, and emphasize the importance of bright RGB inputs.

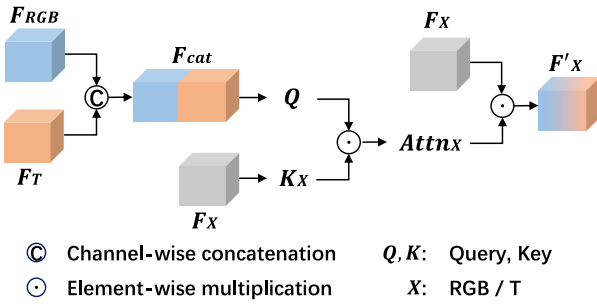


Fig. 8. The cross-modality feature interaction module.

with linear transformation to obtain the key vector:

$$K_X = W_X^k \cdot \text{Flatten}(F_X) + b_X^k, X \in \{RGB, T\}, \quad (4)$$

where W_X^k and b_X^k denote the weight and bias of the key transformation function for RGB or T modality. The attention matrix is then calculated by

$$\text{Attn}_X = \sigma(\text{Unflatten}(Q \odot K_X)), X \in \{RGB, T\}, \quad (5)$$

where σ is the Sigmoid function; Unflatten is the reverse operation of Flatten ; \odot is the Hadamard product. Attn_X can be regarded as the affinity matrix between the features from the fusion and each single modality. Finally, the interacted feature is obtained through element-wise multiplication between the original feature and the attention matrix:

$$F'_X = \text{Attn}_X \odot F_X, X \in \{RGB, T\}. \quad (6)$$

The visualization of W_T^{spa} without and with cross-modality feature interaction is shown in Fig. 9. When cross-modality feature interaction is removed, the foreground region is not salient in W_T^{spa} . By contrast, with the aid of the RGB feature, the crowded areas are highlighted in W_T^{spa} , and the response of the background region is suppressed to some degree. I.e., during the fusion process, in terms of the foreground areas, the model will pay more attention to the thermal density map that is more accurate than the RGB one (quantified comparison will be shown in Table 10).

3.6. Loss function

The density loss, L^{den} , is the summation of the losses between the ground truth and the predicted density maps that consist of the final

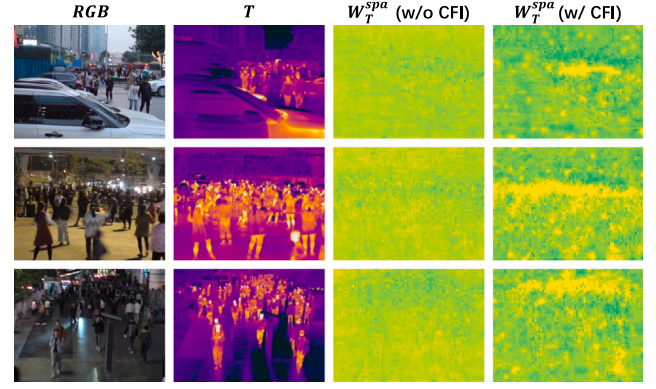


Fig. 9. Visualization of RGB and thermal images and spatial weighting map of thermal modality (W_T^{spa}) without (w/o) and with (w/) cross-modality feature interaction (CFI).

density map and the intermediate results of RGB and thermal branches:

$$L^{den} = L_{final}^{bays} + L_{RGB}^{bays} + L_T^{bays}. \quad (7)$$

The Bayesian loss [33] is adopted in each term of L^{den} for fast convergence and good performance. It is given by

$$L_X^{bays} = \sum_k |P_0(k) \cdot D_k| + \sum_{i=1}^N |1 - \sum_k P_i(k) \cdot D_k|, \quad (8)$$

$$X \in \{final, RGB, T\},$$

where i and N denote the index and the number of the annotated points; D is the density map; $P_0(k)$ denotes the background likelihood at position k , and $P_i(k)$ is the posterior of the occurrence of the i th annotation at position k .

A regularization term, L^{weight} , is added to ensure that the weighted summation of the spatial and illumination weights is close to 1 at each position:

$$L^{weight} = L_1(w_{RGB}^{illu} \cdot W_{RGB}^{spa} + W_T^{spa}, \mathbf{1}), \quad (9)$$

where $L_1(\cdot)$ is the ℓ_1 norm, and $\mathbf{1}$ is the all-one matrix. This term prevents the final density map from being far from the estimated RGB or thermal density map.

The overall loss function is the summation of the above two terms weighted by λ :

$$L = L^{den} + \lambda L^{weight}. \quad (10)$$

4. Experimental configurations

In this section, the experimental configurations will be introduced in terms of datasets, evaluation metrics, and implementation details.

4.1. Datasets

All concerned methods are evaluated on 2 currently available RGB-T crowd counting benchmarks: RGBT-CC [11] and DroneRGBT [12].

RGBT-CC consists of 2030 free-view RGB-thermal image pairs with a total of 138,389 annotated people. The images are officially divided into three parts. I.e., 1030, 200, and 800 image pairs are used for training, validation, and testing, respectively. 1013 image pairs are captured with illumination, and the remaining 1017 pairs are taken in darkness. The images are captured from various scenes, such as malls, streets, playgrounds, stations, etc.

DroneRGBT contains 3600 drone-view RGB-thermal image pairs with a total of 175,6984 annotated pedestrians. According to the official division, 1800 pairs are training samples, and the rest are used for testing. This dataset covers a wide range of scenarios, e.g., campus, street, park, parking lot, playground, and plaza. The images are with large variations in scale, viewpoint, and background clutter.

4.2. Evaluation metrics

As in previous works [9,11,18–21], we adopt the Grid Average Mean Absolute Error (GAME) [53] and the Root Mean Square Error (RMSE) as the evaluation metrics:

$$GAME(L) = \frac{1}{N} \sum_{n=1}^N \left(\sum_{l=1}^{4^L} |e_n^l - g_n^l| \right), \quad (11)$$

where N is the number of images, e_n^l is the estimate count in the region l of image n , and g_n^l is the ground truth count of the same region. This metric divides an image into 4^L non-overlapping regions and calculates the counting error in each region, and L is typically set to 0, 1, 2, and 3. Note that GAME(0) is equivalent to Mean Absolute Error (MAE);

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

where N is the number of images; C_i is the predicted count; C_i^{GT} is the ground truth count. GAME(0) focuses on counting accuracy, and GAME(3) concerns majorly on localization accuracy. GAME(1) and GAME(2) reflect both counting and localization accuracy, and RMSE reveals robustness. For both GAME and RMSE, a lower value means better performance.

4.3. Implementation details

We implement DLF-IA based on PyTorch. Multi-scale feature extraction is applied as in [54–56]. Features of 3 scales are extracted from the second, third, and fourth stages of backbone encoders. During the decoding stage, the interpolation operator is adopted to modify the resolution of multi-scale features to $\frac{1}{8}$ of the original image. The modified features are then concatenated by channel, and finally fed to a 1×1 convolution layer to adjust the number of channels to 1. During training, we randomly crop image patches of size 256×256 from the original image pairs. The ground-truth density map is generated with Gaussian kernels with a fixed bandwidth 8. The encoders are initialized with the encoding layers of the Swin-B [57] model pretrained on ImageNet-22k. Other layers are initialized from a random Gaussian distribution with zero mean and a standard deviation of 0.01. For k and b in illumination weight function, we empirically set them to 10 and the mean illumination value of the dataset (0.440 on RGBT-CC and 0.364 on DroneRGBT), respectively. The hyper-parameter λ in the loss function is set to 0.1 in default. The learning rate is set to $1e-5$ and Adam is used to optimize our network. DLF-IA is trained in an end-to-end manner, and the batch size is set to 1.

5. Results and discussions

In this section, we present the experimental results, together with an extensive analysis. First, the proposed decision-level information fusion approach with illumination awareness (DLF-IA) is compared with state-of-the-art methods. Then, ablation studies are conducted to analyze the effect of each component. Finally, more analysis experiments are presented.

5.1. Comparison with state-of-the-arts

The performance of the proposed DLF-IA compared with state-of-the-art methods on RGBT-CC [11] and Drone-RGBT [12] benchmarks are shown in Tables 3 and 4, respectively. First focus on the two most important evaluation metrics of counting accuracy: GAME(0) and RMSE. Compared with the previous best-performed method, DLF-IA brings a relative improvement of 8.3% and 3.6% respectively on RGBT-CC under the two metrics. Similar improvements can also be observed on the DroneRGBT, with a relative improvement of 6.1% and 3.3%, respectively. Besides, DLF-IA achieves the best GAME(1) on both

Table 3

Performance comparison of DLF-IA with state-of-the-art RGB-T crowd counting methods on RGBT-CC. The best performance is in **boldface**. The same hereinafter.

Method	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
CMCRL [11]	15.61	19.95	24.69	32.89	28.18
CSCA [19]	14.32	18.91	23.81	32.47	26.01
TAFNet [18]	12.38	16.98	21.86	30.19	22.45
MAT [20]	12.35	16.29	20.81	29.09	22.53
CSA-Net [58]	12.45	16.46	21.48	30.62	21.64
R2T [46]	11.63	16.70	22.12	32.32	21.28
DEFNet [21]	11.90	16.08	20.19	27.27	21.09
MC ³ Net [45]	11.47	15.06	19.40	27.95	20.59
CGINet [47]	12.07	15.98	20.06	27.73	20.54
EAEFNet [59]	11.19	14.99	19.20	27.13	19.39
Liu et al. [9]	10.90	14.81	19.02	26.14	18.79
DLF-IA (Ours)	10.00	14.44	19.51	28.63	18.12

Table 4

Performance comparison of DLF-IA with state-of-the-art RGB-T crowd counting methods on DroneRGBT.

Method	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
CMCRL [11]	9.94	12.64	16.37	21.75	15.73
MMCCN [12]	7.27	–	–	–	11.45
I-MMCCN [60]	6.91	–	–	–	11.26
MC ³ Net [45]	7.33	–	–	–	11.17
CGINet [47]	8.37	9.97	12.34	15.51	13.45
DHRNet [10]	6.59	–	–	–	10.66
IAWT [17]	6.66	8.80	11.49	14.51	10.16
DLF-IA (Ours)	6.19	8.28	11.80	18.94	9.82

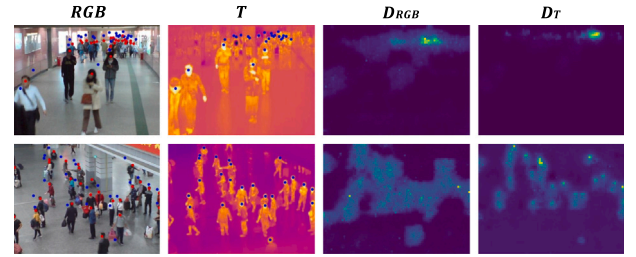


Fig. 10. Spatially misaligned samples of RGBT-CC dataset. As we can see, the annotations provided by the dataset (blue dots in the first two columns) are labeled based on the thermal images. They may be spatially misaligned with the exact positions of the pedestrians in the RGB images (red dots in the first column). As a result, the predicted RGB density maps are blurry and noisy. Best viewed in color and zoomed in.

datasets. These results verify the advantages of the proposed method, *i.e.*, fusing density maps with decision-level late fusion manner and taking RGB illumination into consideration. In this way, counting-wise semantic clues and fine local details can be better utilized, and the importance of different modalities can be flexibly adjusted.

We also note that DLF-IA achieves inferior results in terms of GAME(2) and GAME(3), which indicates that the localization ability of DLF-IA is not so strong. A possible reason lies in that RGB and thermal inputs may be spatially misaligned in RGB-T crowd counting datasets [61]. As shown in Fig. 10, the spatial misalignment problem may decline the localization accuracy of predicted RGB density maps, which may further impair the localization accuracy of our method. A more detailed analysis on the relationship between spatially misaligned RGB-T data and localization performance is presented in Section 5.3.1. To sum up, our method achieves the best performance on most of the evaluation indicators on both datasets, and the good performance of DLF-IA verifies the effectiveness of our propositions.

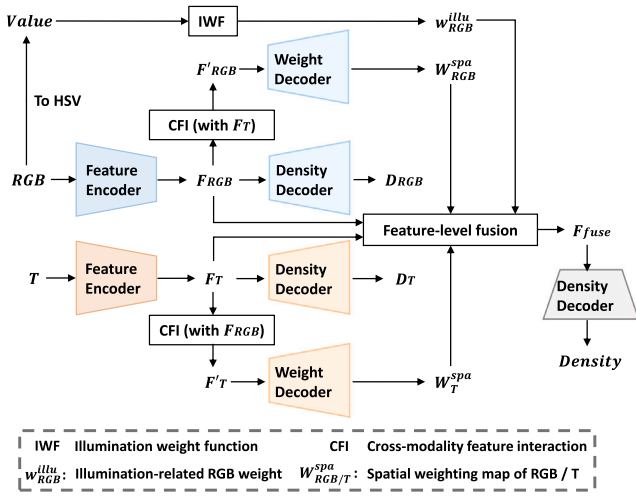


Fig. 11. The technical pipeline of the feature-level fusion baseline. The architecture of encoders and decoders, and the process of illumination-aware weight generation and cross-modality feature interaction are the same as in DLF-IA.

Table 5

Ablation study on fusion strategies.

Strategy	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
Pixel-level	12.51	16.82	22.79	32.13	23.97
Feature-level	10.92	14.99	20.00	28.89	20.54
Decision-level	10.00	14.44	19.51	28.63	18.12

Table 6

Ablation study on cross-modality feature interaction.

Method	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
Without	11.01	15.31	21.14	30.93	19.80
With	10.00	14.44	19.51	28.63	18.12

5.2. Ablation study on RGBT-CC dataset

This section presents ablation study concerned with each component in DLF-IA, including decision-level fusion strategy, illumination-aware RGB weighting mechanism, and cross-modality feature interaction.

5.2.1. Fusion strategies

Here, we compare three different fusion strategies: pixel-level, feature-level, and decision-level (DLF-IA). For the pixel-level fusion baseline, RGB and thermal inputs are first concatenated by channel and then fed to a single-modality crowd counter to obtain the final result. As in Fig. 11, for the feature-level fusion baseline, two spatial weighting maps of RGB and thermal modalities and an illumination-related RGB weight are first obtained. Then fusion is conducted by

$$F_{fuse} = w_{RGB}^{illu} \cdot W_{RGB}^{spa} \odot F_{RGB} + W_T^{spa} \odot F_T. \quad (13)$$

Finally, a density decoder is adopted to convert the fused feature to the final density map. Similar to DLF-IA, two density maps of RGB and thermal counterparts are also predicted, and the Bayesian loss is applied to the three predicted density maps.

As in Table 5, our decision-level fusion approach performs best. Specifically, in terms of GAME(0), our method achieves a relative improvement of 18.9% and 8.4% compared with the pixel-level and feature-level fusion baselines. This phenomenon validates our motivation to conduct decision-level fusion. *I.e.*, it is beneficial to ensure the inputs of the fusion should contain high semantic clues and reveal fine local details.

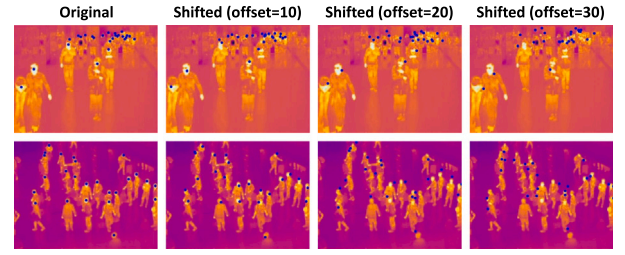


Fig. 12. Two samples of thermal images with (shifted) ground truth annotation points.

5.2.2. Cross-modality feature interaction

Here the effect of cross-modality feature interaction is investigated in Table 6. We can observe a 10.1% performance drop in terms of GAME(0) when the cross-modality feature interaction module is removed. A possible reason lies in that cross-modality feature interaction highlights the importance of thermal input in the foreground region, mitigating the impact of noisy prediction of RGB branch, as shown in Fig. 9. Therefore, cross-modality feature interaction enhances the representation capacities of RGB and thermal features and further facilitates the generation of spatial weighting maps.

5.2.3. Illumination-aware RGB weighting mechanism

Here we investigate the impact of different weighting mechanisms under various illumination conditions: Sigmoid-like, linear, and none (weight set as constant 1). As in Table 7, the illumination-aware RGB weighting mechanism is beneficial mainly on dark nights, with a relative improvement of 6.5% (linear) or 15.0% (Sigmoid-like) in terms of GAME(0). This accords with our motivation that less attention should be paid to RGB inputs under poor light conditions. As for the specific weighting methods, the Sigmoid-like weighting function performs better, with a relative improvement of 15.0% over the linear one in terms of GAME(0). The result verifies our motivation to dynamically adjust the RGB weight according to various light conditions.

5.3. Analysis experiments

5.3.1. Effect of spatially misaligned RGB-T data

In this subsection, we analyze the relationship between spatially misaligned RGB-T data and localization performance. Note that for spatially misaligned RGB and thermal inputs, the localization performance is different. Here we only discuss the localization performance of thermal modality since it aligns with the annotation.

As in Fig. 10, The RGB-T images in current RGB-T crowd counting datasets are spatially misaligned [61], and the ground truth annotation points (GT) are labeled based on the thermal input. So the misalignment between RGB-T data also behaves as the misalignment between RGB input and GT. Now we compare the localization performance sensitivity of decision- and feature-level fusion methods to the degree of RGB-T spatial misalignment. To ensure a fair comparison, the performance of DLF-IA and the feature-level fusion baseline (Section 5.2.1) is reported on the RGBT-CC [11] dataset. Since we cannot guarantee to exacerbate the misalignment between RGB input and GT, we switch to make thermal input and GT misaligned by adding random disturbances to GT, as in Fig. 12. Each annotation point is offset by a fixed pixel in a random direction. The shifted annotation points are then used to supervise the learning of two models. Results are shown in Table 8. We can observe that as the offset increases, the GAME(3) value of DLF-IA rises more rapidly than the feature-level fusion baseline. Therefore, the localization performance of the decision-level fusion method tends to be more susceptible to the spatial misalignment between input and GT than that of the feature-level fusion method.

Now we try to explain this phenomenon. We guess that the localization accuracy may be related to the stage of RGB-T fusion when the

Table 7

Ablation study on illumination-aware RGB weighting mechanisms under various illumination conditions. "None" denotes that the illumination-related RGB weight is set to 1.

Illumination	Function type	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
Brightness	None	10.79	15.89	26.12	39.43	19.83
	Linear	10.80	14.83	19.53	27.76	19.70
	Sigmoid-like	10.82	15.04	21.15	31.35	19.75
Darkness	None	10.79	16.66	26.71	37.87	21.47
	Linear	10.09	14.48	20.87	31.45	17.08
	Sigmoid-like	9.17	13.82	17.83	25.84	16.28
Overall	None	10.79	16.27	26.41	38.66	20.66
	Linear	10.44	14.65	20.21	29.63	18.41
	Sigmoid-like	10.00	14.44	19.51	28.63	18.12

Table 8

Performance comparison of decision- and feature-level fusion methods supervised by various shifted annotation points. "None" denotes that ground truth annotation points are used.

Method	Offset	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
Decision-level	None	10.00	14.44	19.51	28.63	18.12
	10	10.65	16.24	24.32	36.76	19.79
	20	11.41	18.73	30.23	46.34	20.17
	30	12.42	21.76	39.11	60.33	24.50
Feature-level	None	10.92	14.99	20.00	28.89	20.54
	10	11.28	17.05	25.28	37.03	20.41
	20	11.89	18.74	29.72	45.02	22.56
	30	12.50	22.01	34.24	54.73	24.26

RGB-T inputs are spatially misaligned. As mentioned above, in current RGB-T crowd counting datasets, the ground truth annotation points are aligned with the thermal input but misaligned with the RGB input, so the supervision signal is correct for thermal modality but erroneous for RGB one. When the network capacity is powerful enough, the thermal branch may predict accurate results, while the prediction of the RGB branch is prone to errors. The erroneous RGB feature or density map affects the final localization accuracy via RGB-T fusion. When fusion is conducted at an early stage, the subsequent encoding and decoding process may provide the opportunity to reduce the impact of the RGB feature disturbance by attaching importance to the thermal feature. Contrastively, for the decision-level fusion model, the disturbed RGB density map influences the final localization accuracy explicitly. This may be why the localization ability of our model is not so strong on spatially misaligned RGB-T data. Though the issue of RGB-T spatial misalignment affects our model, spatially aligned RGB-T data can be acquired when camera parameters are known. Our model is a good choice under such application scenarios.

5.3.2. Parameter sensitivity of λ in loss function

The loss function (Eq. (10)) is the summation of density loss and a regularization term weighted by λ . Larger λ indicates that the spatial and illumination weights are more strictly regularized. To validate the effect of the regularization term and investigate the parameter sensitivity of λ , DLF-IA is trained with different λ s.

Results are shown in Table 9. The best performance is achieved when $\lambda = 0.1$. Besides, setting $\lambda = 0$, i.e., removing the regularization term, leads to a performance drop. What is more, when λ grows from 0.1 to 10, the performance declines rapidly. The results indicate that proper restrictions should be placed on the weights.

5.3.3. Performance on RGB-D data

To investigate the generalization ability of our method on more diverse multi-modality data, we also evaluate the performance of DLF-IA on an RGB-D crowd counting dataset, ShanghaiTechRGBD [62]. This

Table 9

Quantitative results on different λ s cf. Eq. (10).

λ	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
0	10.44	14.64	20.18	29.46	19.59
0.01	10.31	14.55	20.62	29.98	18.42
0.1	10.00	14.44	19.51	28.63	18.12
1	10.13	15.05	22.62	33.02	17.80
10	10.78	18.00	31.32	46.41	18.89

Table 10

Performance comparison of DLF-IA with state-of-the-art RGB-T crowd counting methods on ShanghaiTechRGBD.

Method	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
CMCRL [11]	4.38	5.95	8.02	11.02	7.06
CSCA [19]	5.78	7.70	10.45	15.88	8.66
CSA-Net [58]	4.09	8.00	13.05	20.25	5.87
DEFNet [21]	6.93	7.32	7.65	7.88	9.01
DLF-IA	3.91	5.36	7.58	11.05	5.64

dataset consists of 2193 RGB-depth image pairs captured by surveillance cameras. The average number of people in each image is 65.9. Since most of the RGB images in this dataset are taken at daytime, we remove the illumination-aware RGB mechanism.

Performance comparison results are shown in Table 10. DLF-IA achieves the best performance on most of the performance indicators among all methods. The result verifies the generalization ability of DLF-IA to various multi-modality crowd counting scenarios.

5.3.4. Multi-modality inputs

To explore whether using multi-modality inputs is effective for crowd counting, we compare DLF-IA with single-modality counters under various illumination conditions. These counters adopt the same architecture as the single-modality branch of DLF-IA.

Results are shown in Table 11. We can see that the multi-modality model predicts more accurate results compared with the single-modality counters, especially on samples at nighttime. Specifically, in terms of GAME(0), it brings a relative improvement of 49.6% and 8.7% over RGB and thermal single-modality baselines. Therefore, RGB and thermal images complement each other, and the multi-modality method can capture the complementary information in both RGB and thermal modalities. Besides, the performance of the thermal-based counter is much superior to that of the RGB-based one, no matter what the illumination condition is. Hence thermal images provide features that are more stable and descriptive for crowd counting than RGB ones.

Table 11
Quantitative results of different input modalities under various illumination conditions.

Illumination	Input modality	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
Brightness	RGB	15.27	20.86	31.39	47.11	30.75
	T	10.99	15.12	20.74	30.23	17.79
	RGB-T	10.82	15.04	21.15	31.35	19.75
Darkness	RGB	25.15	34.29	44.22	58.32	53.45
	T	10.91	14.26	18.43	25.51	19.96
	RGB-T	9.17	13.82	17.83	25.84	16.28
Overall	RGB	19.84	26.88	36.07	50.81	42.32
	T	10.95	14.69	19.60	27.90	18.89
	RGB-T	10.00	14.44	19.51	28.63	18.12

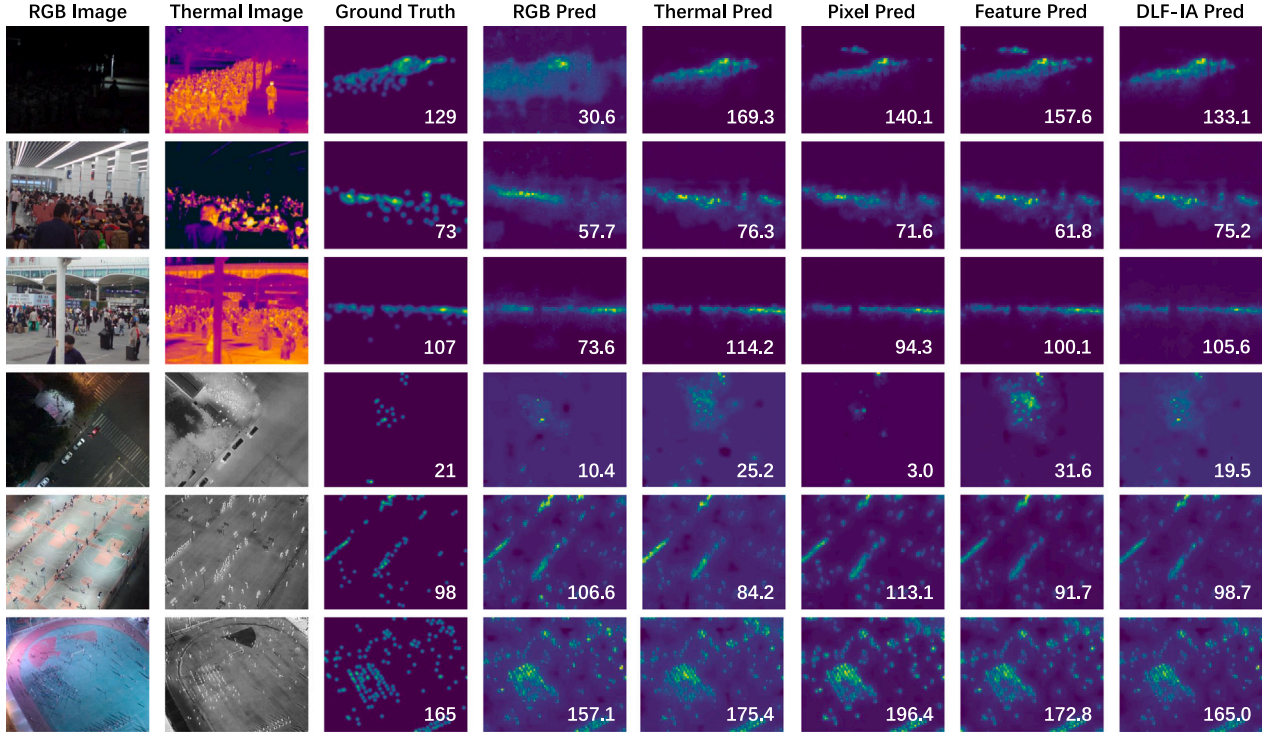


Fig. 13. Successful cases. Samples in the top three rows and bottom three rows are from RGBT-CC and DroneRGBT datasets. The number within the density maps is the total count. “Pred” denotes prediction. “RGB” and “thermal” denote two single-modality branches. “Pixel” and “feature” denote pixel-level and feature-level fusion strategies, respectively.

5.3.5. Qualitative results

First, qualitative results of successful cases are shown in Fig. 13. Compared with different baselines, our method performs favorably for both daytime and nighttime images, both indoor and outdoor scenes, and both sparse and dense crowds. Specifically, for the case shown in the 4th row, due to the poor illumination condition and the interference of stray lights, both RGB and thermal branches fail to predict accurately. The proposed DLF-IA can fuse the RGB-T information effectively and predict favorable results.

Secondly, Fig. 14 presents four failed cases. The model breaks down mainly when the prediction of the thermal branch is erroneous. This may be due to the low-quality imagery of small objects (the 1st row), or the background distraction (the 2nd ~ 4th rows) in the thermal data. In these cases, both our method and feature-level fusion baseline perform unfavorably. However, note that the RGB branch can still work even if the thermal branch fails (the 1st and 4th rows). Therefore, how to effectively fuse RGB and thermal information remains to be an important research issue.

6. Conclusion

In this paper, we propose the first decision-level fusion approach for RGB-T crowd counting. Our intuition to conduct information fusion based on density maps lies in their characteristics of high semantic meaning, high resolution, and high complementarity. Through the illumination-aware RGB weighting mechanism, the illumination condition is concerned, so that the model predicts more accurate results, especially at nighttime. Cross-modality feature interaction is conducted between RGB and thermal features before spatial weighting map generation, therefore the representation capacities of features are enhanced. The proposed DLF-IA achieves state-of-the-art performance on 2 RGB-T counting benchmarks. The results of ablation studies verify the effectiveness of each component.

In the future, we plan to establish a large-scale open-world RGB-T crowd counting benchmark that contains more scene conditions and covers diverse crowd distributions to facilitate the research of RGB-T crowd counting.

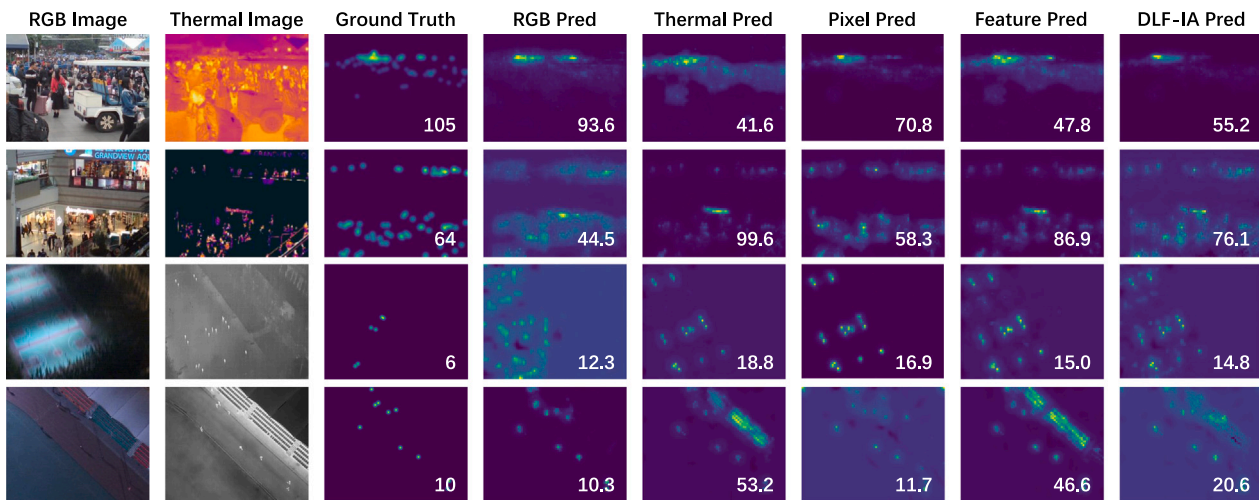


Fig. 14. Failed cases. Samples in the top two and bottom two rows are from RGBT-CC and DroneRGBT datasets.

CRedit authorship contribution statement

Jian Cheng: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Conceptualization. **Chen Feng:** Writing – review & editing, Methodology, Conceptualization. **Yang Xiao:** Writing – review & editing, Methodology, Conceptualization. **Zhiguo Cao:** Writing – review & editing, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This work was supported by the Natural Science Foundation of China under Grant 61876211.

References

- [1] Liqi Yan, Qifan Wang, Siqi Ma, Jingang Wang, Changbin Yu, Solve the puzzle of instance segmentation in videos: A weakly supervised framework with spatio-temporal collaboration, *IEEE Trans. Circuits Syst. Video Technol.* 33 (1) (2022) 393–406.
- [2] Yiming Cui, Liqi Yan, Zhiwen Cao, Dongfang Liu, Tf-blender: Temporal feature blender for video object detection, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, 2021, pp. 8138–8147.
- [3] Wenguan Wang, Cheng Han, Tianfei Zhou, Dongfang Liu, Visual recognition with deep nearest centroids, in: *Proceedings of International Conference on Learning Representations, ICLR*, 2022.
- [4] Dongfang Liu, Yiming Cui, Wenbo Tan, Yingjie Chen, Sg-net: Spatial granularity network for one-stage video instance segmentation, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2021, pp. 9816–9825.
- [5] Mounir Bendali-Braham, Jonathan Weber, Germain Forestier, Lhassane Idoumghar, Pierre-Alain Muller, Recent trends in crowd analysis: A review, *Mach. Learn. Appl.* 4 (2021) 100023.
- [6] Feng Xiong, Xingjian Shi, Dit-Yan Yeung, Spatiotemporal modeling for crowd counting in videos, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV*, 2017, pp. 5151–5159.
- [7] Lingbo Liu, Jiajie Zhen, Guanbin Li, Geng Zhan, Zhaocheng He, Bowen Du, Liang Lin, Dynamic spatial-temporal representation learning for traffic flow prediction, *IEEE Trans. Intell. Transp. Syst.* 22 (11) (2020) 7169–7183.
- [8] Cem Direkoglu, Melike Sah, Noel E. O'Connor, Abnormal crowd behavior detection using novel optical flow-based features, in: *Proceedings of the IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS*, 2017, pp. 1–6.
- [9] Zhengyi Liu, Wei Wu, Yacheng Tan, Guanghui Zhang, RGB-T Multi-Modal Crowd Counting Based on Transformer, in: *Proceedings of British Machine Vision Conference, BMVC*, 2022, pp. 1–14.
- [10] Zhengyi Liu, Yacheng Tan, Wei Wu, Bin Tang, Dilated high-resolution network driven RGB-T multi-modal crowd counting, *Signal Process., Image Commun.* 112 (2023) 116915.
- [11] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, Liang Lin, 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, CVPR*, 2021, pp. 4823–4833.
- [12] Tao Peng, Qing Li, Pengfei Zhu, RGB-T crowd counting from drone: A benchmark and MMCCN network, in: *Proceedings of the Asian Conference on Computer Vision, ACCV*, 2020, pp. 497–513.
- [13] Tiancheng Zhi, Bernardo R. Pires, Martial Hebert, Srinivasa G. Narasimhan, Deep material-aware cross-spectral stereo matching, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2018, pp. 1916–1925.
- [14] Yaoxian Song, Jun Wen, Dongfang Liu, Changbin Yu, Deep robotic grasping prediction with hierarchical rgb-d fusion, *Int. J. Control Autom. Syst.* 20 (1) (2022) 243–254.
- [15] Dongfang Liu, Yiming Cui, Zhiwen Cao, Yingjie Chen, Indoor navigation for mobile agents: A multimodal vision fusion model, in: *International Joint Conference on Neural Networks, IJCNN*, 2020, pp. 1–8.
- [16] Xiaoling Li, Houjin Chen, Yanfeng Li, Yahui Peng, MAFusion: Multiscale attention network for infrared and visible image fusion, *IEEE Trans. Instrum. Meas.* 71 (2022) 1–16.
- [17] Lin Zhou, Zhenzhong Chen, Illumination-aware window transformer for RGBT modality fusion, *J. Vis. Commun. Image Represent.* 90 (2023) 103725.
- [18] Haihan Tang, Yi Wang, Lap-Pui Chau, Tafnet: A three-stream adaptive fusion network for rgb-t crowd counting, in: *Proceedings of the IEEE International Symposium on Circuits and Systems, ISCAS*, 2022, pp. 3299–3303.
- [19] Youjia Zhang, Soyun Choi, Sungeun Hong, Spatio-channel attention blocks for cross-modal crowd counting, in: *Proceedings of the Asian Conference on Computer Vision, ACCV*, 2022, pp. 90–107.
- [20] Zhengtao Wu, Lingbo Liu, Yang Zhang, Mingzhi Mao, Liang Lin, Guanbin Li, Multimodal crowd counting with mutual attention transformers, in: *Proceedings of the IEEE International Conference on Multimedia and Expo, ICME*, 2022, pp. 1–6.
- [21] Wujie Zhou, Yi Pan, Jingsheng Lei, Lv Ye, Lu Yu, DEFNet: Dual-branch enhanced feature fusion network for RGB-T crowd counting, *IEEE Trans. Intell. Transp. Syst.* 23 (12) (2022) 24540–24549.
- [22] Yuting Liu, Miaoqing Shi, Qijun Zhao, Xiaofang Wang, Point in, box out: Beyond counting persons in crowds, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2019, pp. 6469–6478.
- [23] Deepak Babu Sam, Skand Vishwanath Peri, Mukuntha Narayanan Sundararaman, Amogh Kamath, R. Venkatesh Babu, Locate, size, and count: accurately resolving people in dense crowds via detection, *IEEE Trans. Pattern Anal. Mach. Intell.* 43 (8) (2020) 2739–2751.
- [24] Chenchen Liu, Xinyu Weng, Yadong Mu, Recurrent attentive zooming for joint crowd counting and precise localization, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2019, pp. 1217–1226.

- [25] Jian Cheng, Haipeng Xiong, Zhiguo Cao, Hao Lu, Decoupled two-stage crowd counting and beyond, *IEEE Trans. Image Process.* 30 (2021) 2862–2875.
- [26] Shahira Aousamra, Minh Hoai, Dimitris Samaras, Chao Chen, Localization in the crowd with topological constraints, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35, No. 2, 2021, pp. 872–881.
- [27] Junyu Gao, Maoguo Gong, Xuelong Li, Congested crowd instance localization with dilated convolutional swin transformer, *Neurocomputing* 513 (2022) 94–103.
- [28] Qingyu Song, Changan Wang, Zhengkai Jiang, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, Yang Wu, Rethinking counting and localization in crowds: A purely point-based framework, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021*, pp. 3365–3374.
- [29] Chengxin Liu, Hao Lu, Zhiguo Cao, Tongliang Liu, Point-query quadtree for crowd counting, localization, and more, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2023*.
- [30] Victor Lempitsky, Andrew Zisserman, Learning to count objects in images, in: *Proceedings of Advances in Neural Information Processing Systems, NeurIPS, 2010*, pp. 1324–1332.
- [31] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, Yi Ma, Single-image crowd counting via multi-column convolutional neural network, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2016*, pp. 589–597.
- [32] Yuhong Li, Xiaofan Zhang, Deming Chen, Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2018*, pp. 1091–1100.
- [33] Zhiheng Ma, Xing Wei, Xiaopeng Hong, Yihong Gong, Bayesian loss for crowd count estimation with point supervision, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2019*, pp. 6142–6151.
- [34] Boyu Wang, Huidong Liu, Dimitris Samaras, Minh Hoai Nguyen, Distribution matching for crowd counting, in: *Proceedings of Advances in Neural Information Processing Systems, NeurIPS, 2020*, pp. 1595–1607.
- [35] Haipeng Xiong, Hao Lu, Chengxin Liu, Liang Liu, Zhiguo Cao, Chunhua Shen, From open set to closed set: Counting objects by spatial divide-and-conquer, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2019*, pp. 8362–8371.
- [36] Usman Sajid, Guanghui Wang, Towards more effective prn-based crowd counting via a multi-resolution fusion and attention network, *Neurocomputing* 474 (2022) 13–24.
- [37] Yuqiang He, Yinfeng Xia, Yizhen Wang, Baoqun Yin, Jointly attention network for crowd counting, *Neurocomputing* 487 (2022) 157–171.
- [38] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, Xiaopeng Hong, Boosting crowd counting via multifaceted attention, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2022*, pp. 19628–19637.
- [39] Joey Tianyi Zhou, Le Zhang, Jiawei Du, Xi Peng, Zhiwen Fang, Zhe Xiao, Hongyuan Zhu, Locality-aware crowd counting, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (7) (2021) 3602–3613.
- [40] Rui Wang, Reem Alotaibi, Bander Alzahrani, Arif Mahmood, Gaoxiang Wu, Han Xia, Abeer Alshehri, Sahar Aldaheri, AAC: Automatic augmentation for crowd counting, *Neurocomputing* 500 (2022) 90–98.
- [41] Liang Liu, Hao Lu, Hongwei Zou, Haipeng Xiong, Zhiguo Cao, Chunhua Shen, Weighing counts: Sequential crowd counting by reinforcement learning, in: *Proceedings of European Conference on Computer Vision, ECCV, 2020*, pp. 164–181.
- [42] Yasiru Ranasinghe, Nithin Gopalakrishnan Nair, Wele Gedara Chaminda Bandara, Vishal M. Patel, Diffuse-denoise-count: Accurate crowd-counting with diffusion models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023*.
- [43] Haoyue Bai, Jiageng Mao, S.-H. Gary Chan, A survey on deep learning-based single image crowd counting: Network design, loss function and supervisory signal, *Neurocomputing* 508 (2022) 1–18.
- [44] Zizhu Fan, Hong Zhang, Zheng Zhang, Guangming Lu, Yudong Zhang, Yaowei Wang, A survey of crowd counting and density estimation based on convolutional neural network, *Neurocomputing* 472 (2022) 224–251.
- [45] Wujie Zhou, Xun Yang, Jingsheng Lei, Weiqing Yan, Lu Yu, MC³Net: Multimodal-ity cross-guided compensation coordination network for RGB-T crowd counting, *IEEE Trans. Intell. Transp. Syst.* (2023) 1–10.
- [46] He Li, Shihui Zhang, Weihang Kong, Learning the cross-modal discriminative feature representation for RGB-T crowd counting, *Knowl.-Based Syst.* 257 (2022) 109944.
- [47] Yi Pan, Wujie Zhou, Xiaohong Qian, Shanshan Mao, Rongwang Yang, Lu Yu, CGINet: Cross-modality grade interaction network for RGB-T crowd counting, *Eng. Appl. Artif. Intell.* 126 (2023) 106885.
- [48] Zhangyong Tang, Tianyang Xu, Hui Li, Xiao-Jun Wu, Xuefeng Zhu, Josef Kittler, Exploring fusion strategies for accurate RGBT visual object tracking, *Inf. Fusion* (2023) 101881.
- [49] Yanpeng Cao, Xing Luo, Jiangxin Yang, Yanlong Cao, Michael Ying Yang, Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection, *Inf. Fusion* 88 (2022) 1–11.
- [50] Su Pang, Daniel Morris, Hayder Radha, CLOCs: Camera-LiDAR object candidates fusion for 3D object detection, in: *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2020*, pp. 10386–10393.
- [51] Chang Xu, Qingwu Li, Mingyu Zhou, Qingkai Zhou, Yaqin Zhou, Yunpeng Ma, RGB-T salient object detection via CNN feature and result saliency map fusion, *Appl. Intell.* 52 (10) (2022) 11343–11362.
- [52] Lu Liu, William L. Hamilton, Guodong Long, Jing Jiang, Hugo Larochelle, A universal representation transformer layer for few-shot image classification, in: *Proceedings of International Conference on Learning Representations, ICLR, 2020*.
- [53] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, Daniel Onoro-Rubio, Extremely overlapping vehicle counting, in: *Proceedings of Iberian Conference on Pattern Recognition and Image Analysis, IbPRIA, 2015*, pp. 423–431.
- [54] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, Rainer Stiefelshagen, CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers, *IEEE Trans. Intell. Transp. Syst.* (2023).
- [55] Liqi Yan, Yiming Cui, Yingjie Chen, Dongfang Liu, Hierarchical attention fusion for geo-localization, in: *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2021*, pp. 2220–2224.
- [56] Yawen Lu, Guoyu Lu, Superthermal: Matching thermal as visible through thermal feature exploration, *IEEE Robot. Autom. Lett.* 6 (2) (2021) 2690–2697.
- [57] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, Baining Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision, ICCV, 2021*, pp. 10012–10022.
- [58] He Li, Junge Zhang, Weihang Kong, Jienan Shen, Yuguang Shao, CSA-Net: Cross-modal scale-aware attention-aggregated network for RGB-T crowd counting, *Expert Syst. Appl.* 213 (2023) 119038.
- [59] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, Tin Lun Lam, Explicit attention-enhanced fusion for RGB-thermal perception tasks, *IEEE Robot. Autom. Lett.* (2023) 1–8.
- [60] Binyu Zhang, Yunhao Du, Yanyun Zhao, Junfeng Wan, Zhihang Tong, I-MMCCN: Improved MMCCN for RGB-T crowd counting of drone images, in: *Proceedings of the IEEE International Conference on Network Intelligence and Digital Content, IC-NIDC, 2021*, pp. 117–121.
- [61] Martin Thißen, Elke Hergenröther, Why existing multimodal crowd counting datasets can lead to unfulfilled expectations in real-world applications, 2023, *CoRR* abs/2304.06401.
- [62] Dongze Lian, Xianing Chen, Jing Li, Weixin Luo, Shenghua Gao, Locating and counting heads in crowds with a depth prior, *IEEE Trans. Pattern Anal. Mach. Intell.* 44 (12) (2021) 9056–9072.



Jian Cheng received the B.S. degree from Huazhong University of Science and Technology, China. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China.

His research interests include deep learning and computer vision, particularly focusing on multi-modal object counting.



Chen Feng received the B.S. degree from Central South University, China. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, China. His research interests include few-shot learning, transfer learning, and multi-modal fusion in computer vision field, particularly on object recognition and detection.



Yang Xiao received his B.S., M.S., and Ph.D. degrees from Huazhong University of Science and Technology, China. He is currently an associate professor in the School of Artificial Intelligence and Automation at Huazhong University of Science and Technology, China. Previously, he was ever the research fellow in the School of Computer Engineering and Institute of Media Innovation at Nanyang Technological University, Singapore. Dr. Xiao was a recipient of IEEE Innovation Spotlight Research Paper Award 2020, EurAgEng Outstanding Paper Award 2018, and the Best Paper Award at ICIRA 2018. His research interests involve computer vision, image processing and machine learning. He also serves as the Associate Editor of IET Image Processing.



Zhiguo Cao (Member, IEEE) received the B.S. and M.S. degrees in communication and information system from the University of Electronic Science and Technology of China and the Ph.D. degree in pattern recognition and intelligent system from Huazhong University of Science and Technology. He is currently a Professor with the School of Artificial Intelligence and Automation, Huazhong University of Science and Technology. His research interests spread across computational photography, monocular depth estimation, 3d video processing, motion detection, and human

action analysis. He has published dozens of papers in international journals and prominent conferences, which have been applied to image processing in mobile phone cameras, and automatic observation systems for crop growth in agriculture and for weather phenomena in meteorology.