Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Dual-branch Cross-modal Fusion Network with Progressive Feature Enhancement for RGB-T Crowd Counting

Liang Li
Shandong Jiaotong University
Qiqi Chang
Shandong Normal University
Lei Lyu
lvlei@sdnu.edu.cn

Shandong Normal University

Research Article

Keywords: Crowd management, Cross-Modal, RGB-T crowd counting, Dual-branch framework

Posted Date: November 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11068356/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Dual-branch Cross-modal Fusion Network with Progressive Feature Enhancement for RGB-T Crowd Counting

Liang Li[1], Qiqi Chang[2], Lei Lyu[2*]

[1]School of Transportation and Logistics Engineering, Shandong Jiaotong University, Haitang Road, Jinan, 264209, Shandong, China.
[2]School of Information Science and Engineering, Shandong Normal University, Wendong Road, Jinan, 250014, Shandong, China.

*Corresponding author(s). E-mail(s): lvlei@sdnu.edu.cn;
Contributing authors: ll61787@126.com; qqcsdnu@sina.com;

Abstract

To address the reliability imbalance between RGB and thermal modalities under varying illumination conditions, we propose a Dual-branch Cross-modal Fusion Network (DCF-Net). A Cross-modal Feature Fusion Module (CFFM) is introduced to perform reciprocal interaction and adaptive gating between the two modalities, enabling complementary information to be selectively enhanced while preserving modality-specific cues. In addition, a progressive Feature Enhancement Module (FEM) aggregates hierarchical features in a top-down manner after spatial and channel alignment, thereby improving robustness to large pedestrian scale variations. Experiments on RGBT-CC, DroneRGBT, and ShanghaiTechRGBD demonstrate the effectiveness and generalization ability of the proposed approach.

Keywords: Crowd management, Cross-Modal, RGB-T crowd counting, Dual-branch framework

1 Introduction

In recent years, the management of dense crowds has become a critical issue in public safety and emergency management. Against this backdrop, researchers widely recognize that leveraging intelligent technologies to achieve precise perception and analysis of crowd distributions is a key breakthrough for enhancing management effectiveness. Currently, vision-based crowd counting methods, as an important component of intelligent video surveillance, play a significant role in crowd behavior analysis and public safety governance, providing a technical foundation for refined management in high-density scenarios. The core of this approach lies in generating high-precision density maps to accurately estimate the number of individuals in images, thereby providing reliable data support for real-time crowd monitoring, risk warning, and safety management decision-making.

Existing research methods primarily adopt RGB-T network architectures based on convolutional neural networks (CNNs), aiming to improve counting accuracy through cross-modal fusion. However, analysis reveals that current approaches still exhibit notable deficiencies in effectively extracting multimodal information, particularly in the full optimization of cross-modal image detail fusion and multi-scale feature integration. This directly impacts the accurate perception of crowd density and spatial distributions in real dense scenarios, thereby limiting the early warning capability of management and control systems. To address the reliability imbalance between RGB and thermal modalities under varying illumination conditions, as well as the substantial scale variation in dense crowds, we propose a Dual-branch Cross-modal Fusion Network (DCF-Net) for RGB-T crowd counting. DCF-Net aims to adaptively exploit complementary RGB and thermal cues while progressively aggregating multi-level features to improve crowd density estimation under challenging illumination and scale variations.

The main contributions of this work are summarized as follows. First, we propose a Cross-modal Feature Fusion Module (CFFM) that performs reciprocal interaction between RGB and thermal features and adaptively emphasizes complementary cues under varying illumination conditions. Second, we introduce a progressive Feature Enhancement Module (FEM), which aligns and aggregates adjacent multi-level features in a top-down manner to improve multi-scale crowd representation. Third, extensive experiments on RGBT-CC, DroneRGBT, and ShanghaiTechRGBD demonstrate the effectiveness and generalization ability of the proposed framework, while ablation studies verify the individual contributions of cross-modal fusion and hierarchical feature enhancement.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed method. Section 4 describes the experiments and analyzes the results. Finally, Section 5 concludes the paper and discusses future work.

2 Related work

Recent surveys generally categorize crowd counting approaches into detection-based, regression-based, density-estimation-based, and deep-learning-based methods

[1]. Detection-based methods estimate crowd numbers by detecting individual pedestrians or heads, but their performance tends to degrade in highly congested scenes due to severe occlusion. Regression-based methods directly learn a mapping from handcrafted image features to crowd counts, while density-estimation methods further model the spatial distribution of crowds. Lempitsky et al. [2] introduced density maps as regression targets, laying an important foundation for subsequent density-based crowd counting methods.

With the rise of deep learning and the powerful feature extraction capabilities of Convolutional Neural Networks (CNNs), crowd counting methods based on classic CNN architectures have continuously emerged. SKT [3] introduces a structured knowledge transfer framework that distills both intra-layer feature patterns and inter-layer relations from a teacher network into a lightweight student model, improving inference efficiency while maintaining counting accuracy. Despite their strong local representation ability, conventional convolution-based models are relatively limited in capturing long-range contextual dependencies. Attention-based and Transformer-style architectures have therefore been introduced into crowd counting to improve global context modeling. MAN [4], for example, exploits multifaceted attention to improve feature representation. For multimodal crowd counting, recent studies have shifted their focus from simple feature fusion toward explicit cross-modal alignment and complementary representation learning. Meng et al. [5] introduced cross-modal alignment and regional density supervision without adding inference-time parameters, achieving substantial improvements on multiple multimodal crowd-counting benchmarks.

Most existing crowd counting methods primarily focus on RGB images and have achieved remarkable progress. However, they are inherently limited by their reliance on optical data and consequently exhibit subpar performance under challenging illumination conditions, such as low-light or nighttime scenarios. Researchers have sought to address this issue by leveraging cross-modal feature representation, a pursuit further aided by the advancement of multi-sensor systems. In this context, thermal imaging has emerged as a particularly complementary modality to RGB [6, 7]. Thermal imaging captures the heat signatures of pedestrians, providing clear crowd outlines that are independent of visible light conditions. This allows thermal images to effectively distinguish pedestrians from complex backgrounds, making them invaluable for 24/7 crowd monitoring applications. Consequently, RGB-T crowd counting approaches typically center around establishing effective cross-modal feature representations, with a core emphasis on fusing the complementary information from these two distinct modalities. Compared to uni-modal RGB crowd counting, cross-modal problems (such as RGB-T) have been less explored, primarily due to the historical absence of large-scale benchmarks. Specifically, CSCA [8] devises a cross-modal information aggregation module for blending complementary features from paired RGB and thermal images, seamlessly embeddable into existing CNN-based crowd counting frameworks. MAT [9] employs cross-modal mutual attention to establish long-range dependencies and enrich semantic features in crowd counting tasks.

More recently, Shao et al. further demonstrated that multimodal crowd-counting models may suffer substantial performance degradation when one modality is corrupted or unavailable, and introduced a modality-reconfigurable framework to improve robustness under degraded modality conditions [10].

Motivated by the aforementioned discussions, this paper explores effective scale feature representation and enhanced feature extraction during the cross-modal feature fusion process.

3 Method

3.1 Overall architecture

As shown in Fig. 1, DCF-Net adopts a dual-branch architecture for RGB-T crowd counting. RGB and thermal images are separately processed by two parallel PVTv2 backbones to extract hierarchical features. PVTv2 [11] is adopted because its overlapping patch embedding and hierarchical Transformer structure provide multi-level representations suitable for dense prediction tasks.

Given an aligned RGB-thermal image pair, the four stages of the two branches produce hierarchical RGB features R_i and thermal features T_i, where $i = 1, \ldots, 4$. At each stage, CFFM performs reciprocal cross-modal interaction to obtain the fused representation F_i. FEM then progressively aggregates $F_1, \ldots, F_4$ in a top-down manner to obtain the enhanced representation D_1, which is finally fed into the regression head to generate the crowd density map.

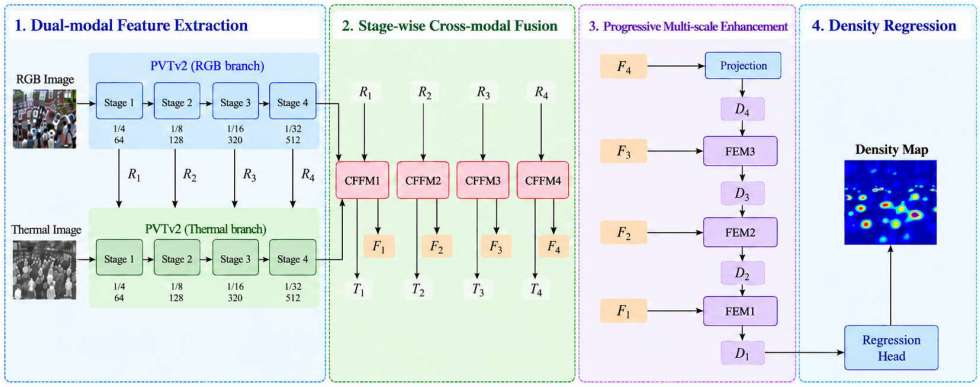

Fig. 1 Overall architecture of the proposed DCF-Net. The network consists of four parts: (1) dual-modal feature extraction, where RGB and thermal images are processed by two parallel PVTv2 branches to obtain multi-level features; (2) stage-wise cross-modal fusion, where the features from the two modalities at each stage are fused by the Cross-modal Feature Fusion Module (CFFM); (3) progressive multi-scale enhancement, where the fused features are progressively refined by the Feature Enhancement Module (FEM); and (4) density regression, where the final enhanced feature is fed into the regression head to generate the crowd density map.

3.2 Cross-modal Feature Fusion Module

Let R_i and T_i denote the RGB and thermal feature maps produced by the i-th stage of the two PVTv2 branches, respectively, where $i = 1, \ldots, 4$. As shown in Fig. 2, the proposed CFFM consists of four stages: cross-modal interaction, reciprocal cross-modal gating, residual refinement, and feature fusion. Its objective is to exploit complementary information between the two modalities while preserving modality-specific representations.

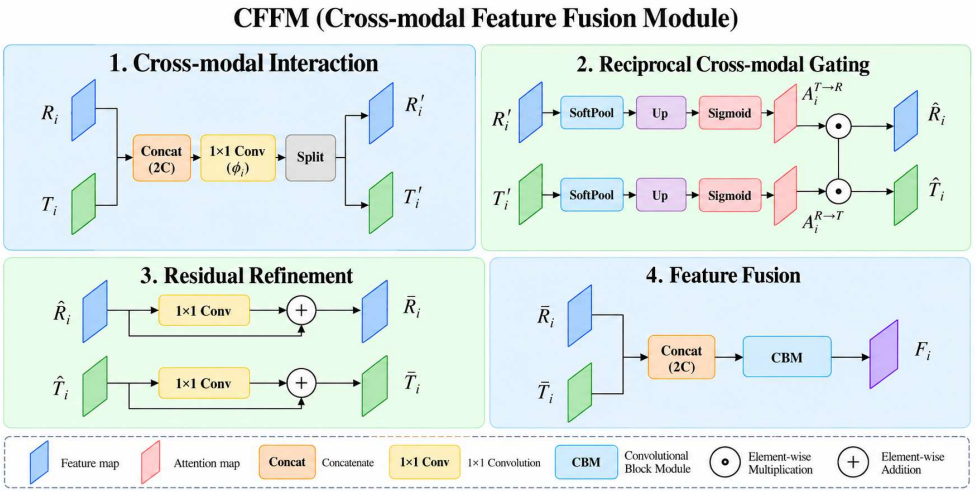

Fig. 2 Architecture of the proposed Cross-modal Feature Fusion Module (CFFM).

In addition to aligning the original features with RGB features, we further investigate the influence of the RGB modality on enhancing thermal features. Following the procedure outlined in [8], the CFFM processes each modality across four stages, and its computational process can be formally expressed as follows:

First, feature interaction is performed: the features of the two modalities are concatenated along the channel dimension and then split, generating hybrid features containing cross-modal information:

$$T_i = \mathrm{PVT}_{v2}\,\mathrm{thermal_stage}(i); \quad R_i = \mathrm{PVT}_{v2}\,\mathrm{RGB_stage}(i) \tag{1}$$

$$\begin{aligned} X_i &= \mathrm{Concat}(R_i, T_i), \\ R_i', T_i' &= \mathrm{Split}\left(\phi_i(X_i)\right). \end{aligned} \tag{2}$$

The RGB and thermal features are first concatenated along the channel dimension. A lightweight 1×1 convolution $\phi_i(\cdot)$ is then used to exchange cross-modal information and align the channel representation. The resulting tensor is subsequently split into two modality-specific hybrid features R_i' and T_i'. Split() is the Split function. Subsequently, a reciprocal cross-modal gating mechanism is employed to estimate

modality-dependent responses. Attention weights are generated using SoftPool layers [12] and Sigmoid functions. Unlike common max pooling and average pooling layers, the SoftPool layer maintains feature expressiveness while reducing the feature map size and preserving information exponentially. The weight coefficients computed by the Sigmoid function are element-wise multiplied with the original modal features to produce enhanced features:

$$\hat{R}_i = \text{Sigmoid}\left(\text{SoftPool}\left(T_i'\right)\right) \odot R_i'; \quad \hat{T}_i = \text{Sigmoid}\left(\text{SoftPool}\left(R_i'\right)\right) \odot T_i \tag{3}$$

Sigmoid() and SoftPool() represent the sigmoid function and SoftPool layer. The final fusion process is expressed as:

$$\bar{R}_i = \text{Conv}_{1\times1}\left(\hat{R}_i\right) + R_i \tag{4}$$

$$\bar{T}_i = \text{Conv}_{1\times1}\left(\hat{T}_i\right) + T_i \tag{5}$$

$$F_i = \text{CBM}\left(\text{Concat}\left(\bar{R}_i, \bar{T}_i\right)\right) \tag{6}$$

CBM() and Conv() represent the block of CBM (concatenation layer, 3×3 convolutional layer and ReLU activation function) and 1×1 convolutional layer. The two tensors undergo 1×1 convolution followed by residual connection with the original features, are then fused through element-wise summation, and finally fed into the CBM module. This module comprises a concatenation layer, a 3×3 convolutional layer, a batch normalization layer, and a ReLU activation function.

Overall, CFFM consists of four stages: cross-modal interaction, reciprocal cross-modal gating, residual refinement, and feature fusion. Cross-modal interaction first introduces information from the complementary modality into each branch. Reciprocal gating then adaptively recalibrates the two hybrid features. Residual refinement preserves modality-specific information from the original backbone features, while the final fusion stage integrates the two refined representations into F_i.

3.3 Feature Enhancement Module

Here, i denotes the feature stage index. CBR denotes a 3×3 convolution followed by Batch Normalization and ReLU activation. The CBM module's content is delineated within the CFFM module, and our proposed module formulation is presented as follows:

$$D_4 = \text{Conv}_{1\times1}(F_4) \tag{7}$$

$$\tilde{D}_{i+1} = \text{Conv}_{1\times1}\left(\text{Up}(D_{i+1})\right), \qquad i = 3, 2, 1 \tag{8}$$

$$D_i = \text{CBR}\left[\text{CBM}\left(F_i, \tilde{D}_{i+1}\right) + F_i + \tilde{D}_{i+1}\right], \qquad i = 3, 2, 1 \tag{9}$$

The CFFM outputs F_i contain modality-fused representations at four semantic levels. Since features from different stages differ substantially in spatial resolution

and semantic abstraction, directly combining them may lead to feature inconsistency. To improve multi-scale representation, FEM adopts a progressive top-down aggregation strategy that transfers high-level semantic information to finer-resolution features while preserving local spatial details. Inspired by [6], we adapt its structural approach to integrate scale information and features within the FEM module, enhancing adaptability to changes in crowd density.

The CFFM outputs F_i contain modality-fused information at four semantic levels, as shown in Fig3. However, these features differ substantially in spatial resolution and semantic abstraction. FEM therefore performs progressive top-down aggregation. The deepest feature F_4 is first projected as D_4. At each subsequent level, the enhanced higher-level feature D_{i+1} is upsampled and channel-aligned before being fused with the current feature F_i. This process transfers high-level semantic information to finer-resolution features while preserving local spatial details. The residual addition of F_i and the aligned higher-level feature prevents information loss during multi-scale aggregation, whereas CBM and CBR further refine the fused representation. Specifically, the deepest feature F_4 is first projected to obtain D_4. For each subsequent level $i = 3, 2, 1$, the higher-level refined feature D_{i+1} is upsampled and channel-aligned to obtain $\tilde{D}_{i+1}$. The aligned higher-level feature and the current-stage feature F_i are concatenated and processed by CBM. Meanwhile, both F_i and $\tilde{D}_{i+1}$ are retained through residual connections. Their combination is finally refined by CBR to generate D_i. This process progressively produces $D_4 \rightarrow D_3 \rightarrow D_2 \rightarrow D_1$.

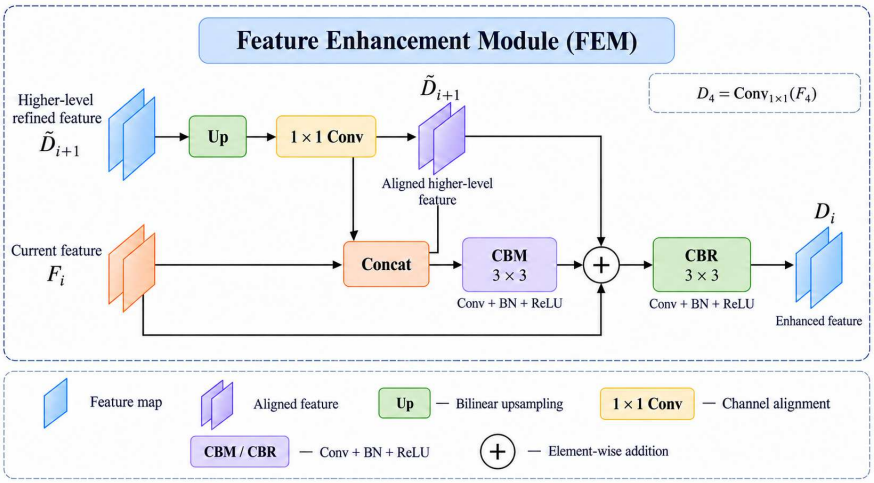

Fig. 3 Architecture of the proposed Feature Enhancement Module (FEM). The higher-level refined feature D_{i+1} is progressively aligned and fused with the current-stage feature F_i in a top-down manner to generate D_i .

3.4 Regression Layer

Inspired by previous work [13], we adopt dilated convolution as a regression layer for generating density maps, which demonstrates superior performance in crowd counting. Integrating convolutional layer fusion across multiple scales enhances our network's ability to extract a more comprehensive and richer set of feature information. Accordingly, our custom-designed regression head layer comprises four dilated convolution layers and one pure convolution layer. The dilated convolution layers utilize kernel sizes of 7×7, 5×5, 3×3, and 1×1, with corresponding dilation rates of 4, 3, 2, and 1. The pure convolution layer employs a 1×1 kernel size.

4 Experiments

4.1 Datasets and evaluation metrics

Experiments are conducted on three public multimodal crowd-counting datasets: RGBT-CC, DroneRGBT, and ShanghaiTechRGBD. Their main characteristics and evaluation protocols are described below.

RGBT-CC [7]. The RGBT-CC dataset is an existing large-scale and challenging dataset comprises 2030 RGB images paired with thermal image, annotated with a total of 138,389 pedestrians. We further process the RGBT-CC dataset by cropping into corresponding RGB regions and resizing them to 640×480. And it is randomly partitioned into three subsets: 1030 image pairs for training, 200 images for validation and 800 images for testing. Within the dataset, 1,013 pairs are captured in well-lit conditions, while 1,017 pairs are captured in low-light and dark conditions.

DroneRGBT [14]. The DroneRGBT dataset is an RGB-T crowd counting dataset for drone view, which contains 3600 pairs of RGB-T images with the size of 640×512 consisting of 175,698 pedestrians. Among them, about 1500 images were captured under dusk conditions, 1200 images under light conditions, and 900 images under dark conditions. We randomly choose 1800 images for training set and the other images for testing set.

ShanghaiTechRGBD [15]. The ShanghaiTechRGBD is a large-scale RGB-D dataset that contains 2,193 images with 144,512 annotated head counts. And it includes depth images ranging from 0 to 20 meters under different lighting conditions. 1,193 images are captured as the training set and the remaining images are utilized as the test set. For experiments on the ShanghaiTechRGBD dataset, the thermal branch in DCF-Net is replaced by the depth branch, while the overall dual-branch architecture and fusion strategy remain unchanged.

Following previous RGB-T crowd-counting studies [7], Grid Average Mean Absolute Error (GAME) and Root Mean Square Error (RMSE) are adopted as evaluation metrics. GAME(l) evaluates both global counting accuracy and spatial distribution accuracy. For each test image, the image is divided into 4^l non-overlapping regions, where l denotes the subdivision level. A larger l corresponds to a finer spatial partition and therefore imposes a stricter requirement on the spatial accuracy of the estimated density distribution. When $l = 0$, GAME(0) is equivalent to MAE. Lower values indicate better performance.

$$\mathrm{GAME}(l) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{4^l}\left|\hat{p}_i^j - p_i^j\right| \tag{10}$$

Here, $\hat{p}_i^j$ and p_i^j denote the predicted and ground-truth counts in the j-th region of the i-th image, respectively, and N is the number of test images.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - p_i)^2} \tag{11}$$

where $\hat{p}_i$ and p_i denote the predicted and ground-truth counts of the i-th image, respectively.

4.2 Implementation Details

The proposed network is implemented in PyTorch and trained on an NVIDIA RTX 3080 GPU. PVTv2-B3 is adopted as the backbone for both RGB and thermal branches. The two branches are initialized with ImageNet-pretrained weights and optimized independently. The stage-wise fused features are projected to 128 channels using 1×1 convolutions before hierarchical aggregation. Bilinear interpolation is employed for feature alignment in FEM.

During training, paired RGB and thermal images are synchronously cropped to 256×256, with random horizontal flipping for data augmentation. The model is trained using Adam with a batch size of 16, an initial learning rate of 1×10^{-5}, and a weight decay of 1×10^{-4}. The total loss is defined as:

$$L_{\text{total}} = L_{\text{count}} + \lambda_v L_{\text{variation}} \tag{12}$$

where λ_v is set to 0.1 to balance counting accuracy and spatial distribution consistency.

4.3 Performance

RGBT-CC [7]: Table 1 presents a comparative analysis between our innovative approach and several contemporary state-of-the-art crowd counting methodologies on the RGBT-CC benchmark. Each method in the table aims to comprehensively capture scene details in diverse ways to effectively fulfill the task of recognition and counting. Notably, MAN [4], and IADM [7] seamlessly integrate uni-modal images into crowd counting, leveraging backbones such as VGG19. Conversely, DEFNet [16] exploits RGB-T complementary information for crowd counting, while CaCmF [17] introduces cross-modal cycle-attention fusion for RGB-D crowd counting. Other methods, namely TAFNet [13], CGINET [18], MJPNet-T [19] and MCN [20], exhibit superior performance by amalgamating RGB and thermal information. These approaches utilize a cross-modal collaborative representation learning framework to systematically capture complementary information from different modalities for precise density map estimation. Building upon insights from prior research, we introduce a dual-branch cross-modal fusion module to merge thermal and RGB images while emphasizing

modality information extraction. The results show that DCF-Net achieves competitive performance compared with several representative RGB-T crowd-counting methods and clearly improves upon a number of earlier fusion-based approaches. However, more recent methods such as CAGNet [21] and Free Lunch [5] obtain stronger overall performance, indicating that further improvements in cross-modal alignment and representation learning remain possible.

Table 1 The performance comparison with different methods on the RGBT-CC benchmark

Method	G(0)	G(1)	G(2)	G(3)	RMSE
CMCRL [7]	17.94	21.44	26.17	33.33	30.91
MAN [4]	17.16	21.78	28.74	41.59	33.84
TAFNet [13]	12.38	16.98	21.86	30.19	22.45
DEFNet [16]	11.90	16.08	20.19	27.27	21.09
CaCmF [17]	15.87	19.92	24.65	28.01	29.31
CGINet [18]	12.07	15.98	20.06	27.73	20.54
MJPNNet-T [19]	11.56	16.36	20.95	28.91	17.83
CAGNet [21]	11.06	14.73	18.94	25.76	17.83
MCN [20]	11.56	15.92	20.16	28.06	19.02
Free Lunch [5]	9.57	12.62	16.27	21.85	17.05
DCF-Net	**11.81**	**15.95**	**19.96**	**26.42**	**19.85**

As shown in Fig. 4, compared to CMCRL [7], our method generates density maps with clearer contours and more accurate counting capabilities. RGB images provide rich optical information of the scene, while thermal images offer thermodynamic information. By fully leveraging contextual feature information, our approach enhances image counting performance. The clear density maps and precise counting capabilities provide a reliable data foundation for real-time monitoring systems, enabling accurate identification of crowd gathering hotspots and distribution changes, thereby supporting timely safety warnings and evacuation decisions.

To further evaluate the robustness of DCF-Net under varying illumination conditions, the 800 RGBT-CC test images are divided into two subsets: 406 bright-condition images and 394 low-light images. The results in Table 2 demonstrate that our method outperforms CMCRL [7] and TAFNet [13] by effectively utilizing the complementary RGB and thermal information.

DroneRGBT [14]: Table 3 lists the performance comparison between DCF-Net and seven state-of-the-art methods on DroneRGBT dataset. On DroneRGBT, DCF-Net achieves an MAE of 6.57 and an RMSE of 10.45. Compared with CrowdAlign [22], these values correspond to improvements of approximately 6.5% and 4.7%, respectively. DCF-Net also achieves performance close to CAGNet [21], although the recent Free Lunch method [5] obtains lower errors. These results indicate that DCF-Net maintains competitive generalization performance on aerial RGB-T crowd scenes.

ShanghaiTechRGBD [15]: To further evaluate the generalization ability of the proposed DCF-Net and its capability in utilizing cross-modal information, we conduct experiments on the ShanghaiTechRGBD dataset [15]. Since ShanghaiTechRGBD is an

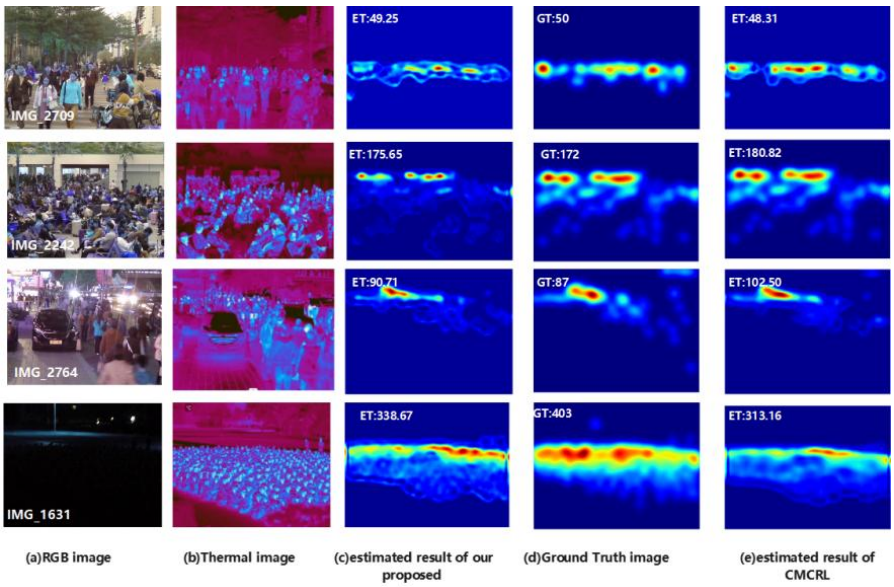

Fig. 4 Qualitative comparison on the RGBT-CC dataset. From left to right: RGB image, thermal image, DCF-Net prediction, ground-truth density map, and CMCRL prediction. Predicted and ground-truth counts are displayed in the upper-left corner of the corresponding density maps.

Table 2 Results comparison of the proposed method under different illumination.

Illumination	Method	G(0)	G(1)	G(2)	G(3)	RMSE
Brightness	CMCRL [7]	20.36	23.57	28.49	36.29	32.57
	TAFNet [13]	15.57	20.65	26.67	36.17	24.25
	DCF-Net	12.31	16.37	20.69	27.22	20.74
Darkness	CMCRL [7]	15.44	19.23	23.79	30.28	29.11
	TAFNet [13]	14.20	19.20	24.00	31.63	27.50
	DCF-Net	**11.52**	**15.66**	**19.73**	**26.09**	**19.03**

RGB-D crowd counting dataset, the thermal branch in DCF-Net is replaced by the depth branch, while the overall dual-branch architecture and fusion strategy remain unchanged. Following existing works, we compare the proposed method with two categories of methods, including multimodal methods and specifically designed RGB-D crowd counting models. The performance of all comparison methods is reported in Table 4. The results on ShanghaiTechRGBD demonstrate that DCF-Net can be directly extended from RGB-T to RGB-D crowd counting by replacing the thermal branch with a depth branch. Although specialized RGB-D approaches such as CCANet [26] achieve lower overall errors, DCF-Net remains competitive with several representative RGB-D baselines. This result indicates that the proposed cross-modal fusion strategy exhibits reasonable generalization across different auxiliary modalities.

Table 3 The performance comparison with different methods on DroneRGBT dataset benchmark.

Method	MAE	RMSE
CSRNet [23]	8.91	13.80
MC^{3}Net [24]	7.33	11.17
CAGNet [21]	6.48	10.30
CFAF-Net [25]	6.72	11.06
CrowdAlign [22]	7.03	10.96
Free Lunch [5]	4.94	7.81
DCF-Net	**6.57**	**10.45**

Table 4 The performance comparison with different methods on ShanghaiTechRGBD dataset benchmark.

Method	G(0)	G(1)	G(2)	G(3)	RMSE
CCANet [26]	3.78	5.34	7.63	10.17	5.56
RDNet [15]	4.96	-	-	-	7.22
MCNN+IADM [7]	9.61	11.89	15.44	20.69	14.52
BL+IADM [7]	7.13	9.28	13.00	19.53	10.27
CSRNet+IADM [7]	4.38	5.95	8.02	11.02	7.06
MCN [20]	4.68	-	-	-	6.87
VPMFNet [27]	5.24	6.85	9.14	13.81	8.19
Free Lunch [5]	3.69	5.88	8.60	11.85	5.38
DCF-Net	**4.58**	**6.81**	**9.16**	**12.09**	**6.64**

4.4 Ablation experiments

In accordance with the methodology outlined in Section 3.1, the envisioned comprehensive architecture comprises the PVTv2 backbone, the CFFM module, and the FEM module. This section delves into the effectiveness of these components in the proposed approach, exploring their impact on various typical scenes in detail.

Table 5 Comparison of different backbone configurations on the RGBT-CC dataset.

Method	PVTv2	G(0)	G(1)	G(2)	G(3)	RMSE
DCF-Net	B1	15.31	19.58	24.43	32.56	22.33
	B2	13.86	19.25	23.51	32.18	21.12
	B2-li	12.21	17.04	21.68	29.23	20.23
	B3	11.81	15.95	19.96	26.42	19.85
	B4	13.78	18.97	23.64	31.24	20.74
	B5	15.61	19.95	24.69	32.85	22.46
	VGG-16	20.23	22.58	28.03	35.55	35.21
	VGG-19	19.07	22.66	27.22	35.87	33.59

As depicted in Table 5, we conducted extensive experiments to assess the efficacy and validate the sustained stability of our proposed methodology. Our focus was on leveraging its inherent advantages to extract feature information from both RGB and thermal images within the framework. Firstly, we subjected six variants of PVTv2 to rigorous testing on identical datasets without any alterations to conditions or modules. Notably, the results revealed that among these variants, PVTv2 B3 exhibited superior performance, yielding a more accurate estimation density map. Additionally, numerous experiments demonstrated that PVTv2 B3 outperformed other variants, producing better results in terms of accuracy. To further substantiate the effectiveness of PVTv2 as the primary data processing tool, we employed pre-trained VGG-16 and VGG-19 networks as the foundational structure for the dual-branch framework, rather than using PVTv2 as the module backbone. The comparison presented in Table 5 illustrates that our proposed method outperforms the baseline, showcasing improved performance across various metrics.

To evaluate the contribution of each component in DCF-Net, ablation experiments are conducted on the RGBT-CC dataset. As shown in Table 6, the PVTv2 baseline yields relatively large counting errors, indicating that independently extracted backbone features are insufficient for effective RGB-T crowd counting. Introducing CFFM substantially reduces G(0) from 33.96 to 13.82 and RMSE from 55.22 to 22.39. This improvement demonstrates that cross-modal interaction enables the network to better exploit the complementary information between RGB and thermal features, thereby improving feature robustness under varying imaging conditions.

Table 6 Ablation studies on RGBT-CC Dataset

Component	G(0)	G(1)	G(2)	G(3)	RMSE
PVTv2	33.96	42.45	53.82	65.96	55.22
RGB+T+PVTv2+CFFM	13.82	19.28	23.41	32.78	22.39
RGB+T+PVTv2+FEM	15.23	22.76	27.93	37.24	27.62
RGB+RGB+PVTv2+CFFM+FEM	14.21	20.64	25.32	34.93	23.25
T+T+PVTv2+CFFM+FEM	12.52	17.72	20.96	28.89	21.34
DCF-Net	**11.81**	**15.95**	**19.96**	**26.42**	**19.85**

Incorporating FEM alone also improves the baseline, reducing G(0) to 15.23 and RMSE to 27.62. This result indicates that hierarchical feature enhancement is beneficial for combining semantic information from different feature levels and improving multi-scale crowd representation. Compared with either individual module, the complete DCF-Net achieves the best performance across all metrics, suggesting that CFFM and FEM play complementary roles: CFFM mainly enhances cross-modal feature interaction, whereas FEM further refines the fused representation through multi-level feature aggregation.

The RGB+RGB and T+T settings are further introduced to investigate the contribution of modality complementarity. The T+T configuration performs better than RGB+RGB, indicating that thermal information is generally more robust to illumination degradation. However, neither single-modality duplication achieves

the performance of the RGB+T configuration. In particular, the complete DCF-Net achieves G(0) of 11.81 and RMSE of 19.85, outperforming both RGB+RGB and T+T settings. These results suggest that the improvement does not merely arise from the dual-branch architecture itself, but from the complementary characteristics of RGB texture information and thermal saliency. This observation further supports the necessity of cross-modal fusion in DCF-Net.

Table 7 Comparison of different cross-modal fusion strategies on the RGBT-CC dataset.

Fusion Strategy	G(0)	G(1)	G(2)	G(3)	RMSE
RGB-T Add	16.87	22.34	27.86	37.18	27.15
RGB-T Concatenation	15.46	20.71	25.68	34.52	24.63
RGB-T + CFFM	13.82	19.28	23.41	32.78	22.39
DCF-Net	11.81	15.95	19.96	26.42	19.85

To evaluate the effectiveness of CFFM, we compare it with two common fusion strategies, element-wise addition and feature concatenation, while keeping the backbone and training settings unchanged. As shown in Table 7, CFFM achieves lower G(0) and RMSE than both Add and Concatenation, indicating that explicit cross-modal interaction is more effective than direct feature aggregation. With FEM further introduced, the complete DCF-Net achieves the best overall performance.

Table 8 Performance comparison under different crowd-scale conditions on the RGBT-CC dataset.

Method	Small		Medium		Large	
	G(0)	RMSE	G(0)	RMSE	G(0)	RMSE
PVTv2 baseline	40.82	64.35	34.11	55.40	26.95	45.91
PVTv2 + FEM	18.74	33.84	15.08	27.41	11.87	21.61
DCF-Net	13.92	24.31	11.74	19.74	9.77	15.50

To further evaluate the effectiveness of FEM under scale variation, the RGBT-CC test set is divided into small-, medium-, and large-scale subsets using average local annotation spacing as an approximate scale indicator. As shown in Table 8, FEM consistently improves the baseline across all scale groups, with a more evident improvement in the small-scale subset. The complete DCF-Net achieves the best performance under all three conditions, demonstrating improved robustness to crowd scale variation.

5 Conclusion

This paper presents DCF-Net for RGB-T crowd counting under challenging illumination and scale variations. The proposed CFFM performs reciprocal cross-modal interaction to exploit complementary RGB and thermal information, improving the robustness of feature representation when one modality becomes less reliable. The proposed FEM progressively aggregates hierarchical features after spatial and channel alignment, enhancing multi-scale crowd representation. Experimental results on three public datasets and corresponding ablation studies demonstrate the effectiveness of the two modules. Future work will investigate more lightweight cross-modal architectures and temporal modeling for video-based crowd analysis.

Acknowledgements. This work is financially supported in part by the National Natural Science Foundation of China (62302555), the Shandong Provincial Natural Science Foundation (ZR2022MG033, ZR2021QG004, ZR2024QF263), and Key R&D Program of Shandong Province (No.2024TSGC0071, 2025TSGCCZZB0115).

References

[1] Gao, G.Q., Gao, J.Y., Liu, Q., Wang, Q., Wang, Y.H.: A survey of deep learning methods for density estimation and crowd counting. Vicinagearth **2**, 2 (2025)

[2] Lempitsky, V., Zisserman, A.: Learning to count objects in images. In: Proceedings of Advances in Neural Information Processing Systems, pp. 1324–1332 (2010)

[3] Liu, L., Chen, J.Q., Wu, H.F., Chen, T.S., Li, G.B., Lin, L.: Efficient crowd counting via structured knowledge transfer. In: Proceedings of the 28th ACM International Conference on Multimedia, pp. 2645–2654 (2020)

[4] Lin, H., Ma, Z., Ji, R., Wang, Y., Hong, X.: Boosting crowd counting via multifaceted attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19628–19637 (2022)

[5] Meng, H., Hong, X., Lai, Z., Shang, M.: Free lunch enhancements for multi-modal crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14013–14023 (2025)

[6] Yang, S.D., Su, H.T., Hsu, W.H., Chen, W.C.: DECCNet: depth enhanced crowd counting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)

[7] Liu, L., Chen, J., Wu, H., Li, G., Li, C., Lin, L.: Cross-modal collaborative representation learning and a large-scale RGB-T benchmark for crowd counting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4823–4833 (2021)

[8] Zhang, Y., Choi, S., Hong, S.: Spatio-channel attention blocks for cross-modal crowd counting. In: Proceedings of the Asian Conference on Computer Vision, pp. 90–107 (2022)

[9] Wu, Z., Liu, L., Zhang, Y., Mao, M., Lin, L., Li, G.: Multimodal crowd counting with mutual attention transformers. In: 2022 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6 (2022)

[10] Shao, P., Ma, C., Zheng, Y.M., Neri, F., Meng, Y.X., Qing, A.Y., Wang, Y.: Robust multimodal crowd counting with modality reconfigurability. Integrated Computer-Aided Engineering **33**(3), 235–247 (2026)

[11] Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: PVT v2: Improved baselines with pyramid vision transformer. Comput. Vis. Media. **8**, 415–424 (2022)

[12] Stergiou, A., Poppe, R., Kalliatakis, G.: Refining activation downsampling with SoftPool. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10357–10366 (2021)

[13] Tang, H., Wang, Y., Chau, L.P.: TAFNet: a three-stream adaptive fusion network for RGB-T crowd counting. In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 3299–3303 (2022)

[14] Peng, T., Li, Q., Zhu, P.: RGB-T crowd counting from drone: a benchmark and MMCCN network. In: Proceedings of the Asian Conference on Computer Vision (2020)

[15] Lian, D., Li, J., Zheng, J., Luo, W., Gao, S.: Density map regression guided detection network for RGB-D crowd counting and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1821–1830 (2019)

[16] Zhou, W., Pan, Y., Lei, J., Ye, L., Yu, L.: DEFNet: dual-branch enhanced feature fusion network for RGB-T crowd counting. IEEE Trans. Intell. Transp. Syst. **23**(12), 24540–24549 (2022)

[17] Li, H., Zhang, S., Kong, W.: RGB-D crowd counting with cross-modal cycle-attention fusion and fine-coarse supervision. IEEE Trans. Ind. Informat. **19**(1), 306–316 (2023)

[18] Pan, Y., Zhou, W., Qian, X., Mao, S., Yang, R., Yu, L.: CGINet: cross-modality grade interaction network for RGB-T crowd counting. Eng. Appl. Artif. Intel. **126**, 106885 (2023)

[19] Zhou, W.J., Yang, X., Dong, X., Fang, M.X., Yan, W.Q., Luo, T.: MJPNet-S*:

multistyle joint-perception network with knowledge distillation for drone RGB-thermal crowd density estimation in smart cities. IEEE Internet Things J. **11**, 20327–20339 (2024)

[20] Kong, W., Liu, J., Hong, Y., Li, H., Shen, J.: Cross-modal collaborative feature representation via transformer-based multimodal mixers for RGB-T crowd counting. Expert Syst. Appl. **255**, 124483 (2024)

[21] Yang, X., Zhou, W., Yan, W., Qian, X.: CAGNet: Coordinated attention guidance network for RGB-T crowd counting. Expert Syst. Appl. **243**, 122753 (2024)

[22] Kong, W., Yu, Z., Li, H., Tong, L., Zhao, F., Li, Y.: CrowdAlign: shared-weight dual-level alignment fusion for RGB-T crowd counting. Image. Vision Comput. **148**, 105152 (2024)

[23] Li, Y., Zhang, X., Chen, D.: CSRNet: dilated convolutional neural networks for understanding the highly congested scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1091–1100 (2018)

[24] Zhou, W., Yang, X., Lei, J., Yan, W., Yu, L.: MC3Net: multimodality cross-guided compensation coordination network for RGB-T crowd counting. IEEE Trans. Intell. Transp. Syst. **25**(5), 4156–4165 (2024)

[25] Kong, W., Yu, Z., Li, H., Zhang, J.: Cross-modal misalignment-robust feature fusion for crowd counting. Eng. Appl. Artif. Intel. **136**, 108898 (2024)

[26] Liu, Y., Cao, G., Shi, B., Hu, Y.: CCANet: A collaborative cross-modal attention network for RGB-D crowd counting. IEEE Trans. Multimedia **26**, 154–165 (2024)

[27] Mu, B., Shao, F., Xie, Z., Chen, H., Jiang, Q., Ho, Y.S.: Visual prompt multi-branch fusion network for RGB-thermal crowd counting. IEEE Internet Things J. **11**(19), 31758–31775 (2024)