MIANet: Bridging the Gap in Crowd Density Estimation With Thermal and RGB Interaction

Shuyu Wang, Weiwei Wu, *Member, IEEE*, Yinglin Li, Yuhang Xu, and Yan Lyu

***Abstract*— Video surveillance and crowd analysis are essential for urban public safety, particularly for accurate crowd counting and density estimation. Existing methods primarily use RGB modality, which limits their effectiveness in complex environments. With advancements in thermal sensors, some studies have combined RGB and thermal images to improve crowd counting accuracy. However, the existing studies face the risk of *introducing redundant information during modality interaction* and *exacerbating the influence of non-uniform crowd density for modality fusion*. Therefore, further advancements are needed to bridge the gap in crowd density estimation by RGB and thermal image interaction. To adequately capture information from both RGB and thermal crowd images and alleviate the above difficulties, we propose a Modality Interaction Attention Network (MIANet). Specifically, Modality Interaction Attention (MIA) module consists of two Multi-Scale Attention (MSA) and a Channel Direction Attention (CDA), which serve to remove redundant information and amplify modality attributes. The MSA incorporates multi-scale kernel factors, enabling its application to still images to solve non-uniform crowd density in one image. To combine modality-specific attributes, the Tri-level MIA modules are connected to the front-end network in a stacked manner. Polished fusion features are further extracted using the Grid Block that combine level-by-level features. On two real-world datasets, we conducted in-depth experiments. Results of the evaluation reveal that our MIANet works better than cutting-edge baseline methodologies and MIANet variants in relation to a variety of prediction inaccuracies, highlighting the efficiency of MIANet and each of its essential modules in crowd density estimation. Code is available at Github.**

***Index Terms*— Crowd counting, crowd density estimation, modality interaction attention, channel direction attention, thermal images.**

I. Introduction

CROWD counting and density estimation have garnered considerable interest in recent times owing to their vital applications in video surveillance [1], [2], [3], [4], crowd analysis [5], [6], [7], [8], and public safety [9], [10], [11], [12].

In the recent past, there has been an influx of models [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] for crowd counting by using RGB videos or images. Nonetheless, these modalities are vulnerable to suboptimal illumination conditions, rendering this category of approaches inapplicable in unrestricted scenes. To mitigate the impact of illumination conditions on RGB images, some research endeavors have explored the fusion of depth or thermal with RGB imagery, given that the former is relatively immune to ambient lighting. Due to the limitations of depth cameras in perceiving distance, their application in large scenes is challenging [27], [28], [29], [30]. This hinders research on combining depth and RGB images for crowd counting. More researchers are combining thermal and RGB images for crowd counting because thermal images complement RGB images and are not affected by perceived distance. Fig. 1 indicates that thermal images display resilience in the face of illumination variations, which enhances the detection of crowds in complex environments. Conversely, integrating RGB imaging can alleviate the incidence of erroneous positive detections in thermal imagery.

Numerous works [31], [32], [33], [34], [35], [36], [37], [38], [39] have explored RGB and thermal images to realize crowd counting and density estimation. Several studies [32], [34], [37], [38] have shown that CNN-based approaches without attention mechanisms struggle to integrate feature representations from different modalities effectively and introduce redundant information. Although a few models [33], [35], [39] incorporate vanilla attention mechanisms to integrate features from different modalities, non-uniform crowd density in image pairs poses greater challenges in modality interactions for fusion. A couple of methods [31], [36] have adopted transformer concepts to merge feature representations across modalities. Although transformers excel at capturing contextual information over long sequences with self-attention, these works create sequence features through patches. The lack of patch location embedding limits their ability to extract

Received 30 January 2024; revised 2 September 2024; accepted 20 September 2024. Date of publication 23 October 2024; date of current version 9 January 2025. This work was supported in part by the Natural Science Foundation of China under Grant 62102082, Grant 62472093, Grant 62232004, Grant 62263028, and Grant 62262062; in part by Jiangsu Natural Science Foundation of China under Grant BK20210203 and Grant BK20230024; in part by the Fundamental Research Funds for the Central Universities; in part by the 2024 Research Program of the Education Department of Shaanxi Province (Research on Key Technologies of Crowd Density Estimation Algorithms Based on Multi-Source Images in Complex Environments); in part by the General Project of Xizang Minzu University under Grant 24MDY06; in part by the Key Laboratory of Computer Network and Information Integration (Southeast University); and in part by the Xizang Key Laboratory of Optical Information Processing and Visualization Technology (Xizang Minzu University). The Associate Editor for this article was J. Hemanth. *(Shuyu Wang and Weiwei Wu contributed equally to this work.) (Corresponding authors: Yinglin Li; Yuhang Xu; Yan Lyu.)*

Shuyu Wang is with the School of Cyber Science and Engineering, Southeast University, Nanjing 210018, China, and also with the School of Information Engineering, Xizang Minzu University, Xianyang 712000, China (e-mail: shywang@seu.edu.cn).

Weiwei Wu, Yinglin Li, and Yan Lyu are with the School of Computer Science and Engineering, Southeast University, Nanjing 210018, China (e-mail: weiweiwu@seu.edu.cn; yinglinli@seu.edu.cn; lvyanly@seu.edu.cn).

Yuhang Xu is with Northern Information Control Research Academy Group Company Ltd., Nanjing 211153, China (e-mail: yuhang_xu@seu.edu.cn).

Digital Object Identifier 10.1109/TITS.2024.3478292

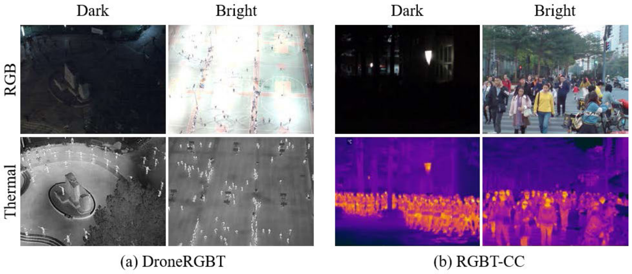


Fig. 1. Images presented in the 1st and 2nd rows correspond to RGB and thermal modality, respectively. (a) Samples from DroneRGBT[1] dataset. The left two images in dark illumination and the right in bright illumination. (b) Samples from RGBT-CC[2] dataset. The left two in dark illumination and the right in bright illumination.

crowd features after modality interaction. Therefore, further advancements are needed to bridge the gap in crowd density estimation by RGB and thermal image interaction.

Based on the above findings, we are motivated to explore effective approaches for merging multimodal information without introducing redundancy and fully leveraging complementary information across modalities. The two challenges we have identified are as follows: 1) *Reconciling the complementarity inherent in multimodal image pairs runs the risk of introducing redundant information.* 2) *Addressing non-uniform crowd density in multimodal image pairs poses greater challenges in modality interactions for feature fusion.*

To overcome the above technical obstacles, we have proposed a Modality Interaction Attention Network (MIANet) to bridge the gap in crowd density estimation by integrating RGB and thermal modalities. The MIANet is presented in Fig. 2. Our approach tackles *challenge 1*) by incorporating attention mechanisms into feature interaction to reduce redundant information. To achieve this, we propose Modality Interaction Attention (MIA) module and Grid Block. Multi-Scale Attention (MSA) and Channel Direction Attention (CDA) are utilized to construct the Modality Interaction Attention (MIA) module, which is subsequently applied to form a Tri-level MIA to merge the features obtained from RGB and thermal crowd images. Furthermore, the Grid Block is leveraged to facilitate level-by-level feature fusion, thereby promoting improved fusion feature extraction and reducing redundant information. To tackle *challenge 2*), MSA module integrates multi-scale kernel factors to improve crowd features from both RGB and thermal branch, respectively. This integration optimizes feature extraction in the presence of non-uniform crowd densities within individual images. Our contributions are:

- introduced a Modality Interaction Attention (MIA) module, which is designed to promote improved fusion feature extraction and reduce redundant information. MIA is constructed using Multi-Scale Attention (MSA) and Channel Direction Attention (CDA). MSA improves feature extraction in non-uniform crowd densities within individual images, while CDA reduces redundancy of complementary information between modalities.
- proposed Modality Interaction Attention Network (MIANet). MIA modules are arranged in a tri-level stacking manner. Additionally, a Grid Block is introduced to consolidate diverse features from various MIA levels, enhancing the accuracy of fusion features for the backend regression process.
- conducted several in-depth experiments on two RGBT datasets to evaluate the effectiveness of MIANet and its variants. The findings demonstrate the superiority of MIANet, which not only surpasses previous cutting-edge methods but also outperforms its own variants, exhibiting remarkable performance.

The subsequent paragraphs of this paper are structured as follows: Section II introduces the literature review of related research. Section III formally proposes the MIANet and its critical components in detail. Section IV delineates performance evaluation, while Section V is conclusion.

II. Related Work

We discuss pertinent prior research, encompassing crowd density estimation using only RGB images, as well as using multimodal images.

A. RGB-Only Image Crowd Counting

Crowd counting estimates the number of people, whereas density estimation creates a relative density map for a crowd image. Recent studies have aimed to tackle both tasks simultaneously using deep learning. Deep learning models for crowd estimation with only RGB images can be classified into three types: vanilla CNN, multi-branch, and single-branch architectures.

1) Vanilla CNN: Fu et al. [17] first adopted CNN to implement crowd counting tasks. Their model meets engineering applications due to the real-time advantage of execution speed. To tackle the issue of manual feature extraction in crowd counting tasks within dense environments, Wang et al. [16] employ a CNN for estimating crowd density representation using imagery captured by cameras. To eliminate the influence of non-crowd areas in the image, they introduced crowd-free images marked as 0, which enhanced the robustness of the model.

2) Multi-Branch: MCNN [18] is a multi-branch architecture comprising three convolutional neural network columns with filters of varying sizes. MCNN represents a seminal work that explicitly addresses the multi-scale issue. CrowdNet [19] is utilized for addressing counting tasks within highly populated crowds. This approach leverages a fusion of semantic and low-order features, alongside data augmentation to boost counting accuracy. Facing the disparity in camera view across different images, and the irregular crowd density within a single image, Switch-CNN [20] initially grids the original image, which decreases the density and viewing angle differences within the grid, augmenting the image for subsequent analysis. Following this preprocessing, multiple regression heads are tailored to accommodate distinctive image blocks.

3) Single-Branch: CSRNet [21] comprises two crucial components. The first component is responsible for feature extraction from the crowd images. The second is engineered

[1] https://github.com/VisDrone/DroneRGBT

[2] https://github.com/chen-judge/RGBTCrowdCounting

to counteract the computational performance degradation attributable to a limited receptive field, achieved through the dilated convolution operators. ADCrowdNet [22] exhibits a two-stage network architecture. The primary phase involves the generation of an attention map, precisely aimed at identifying crowd locations. The subsequent phase comprises the generation of a density map with prior congested knowledge to generate a density map. TEDnet [23] adopts a lattice-style encoder-decoder network architecture, incorporating myriad decoding pathways, allowing for the extraction of multi-scale features. This is further enhanced by the integration of dense skip connections, which facilitate the employment of supervised data within the model. Earlier crowd estimation methods mainly used RGB images, which can be inadequate in uncontrolled settings.

B. Multi-Modality Image Crowd Counting

To improve crowd-counting performance, research has focused on two main areas in multimodal image-based crowd counting: RGB-depth and RGB-thermal image pairs.

1) RGB-Depth: In contrast to RGB images, depth maps offer supplementary details about head localization. Bondi et al. [27] devise a meeting real-time requirements approach that leverages depth imagery to detect and count crowds. A segmentation-task and density estimation-task paradigm is proposed by Arteta et al. [28], integrating the partitioning of foreground-background elements and the unambiguous evaluation of localized indeterminacy. Song et al. [29] develop a deep detection network for people counting that relies on depth information gathered from an overhead vertical Kinect sensor. Lian et al. [30], [40] put forth a double-route detection framework (DPDNet) capable of enumerating and identifying the spatial positions of congregated individuals. Yang et al. [41] introduce DECCNet which capitalizes on estimated depth information employing a creative BCA module. Li et al. [42] present an innovative approach to crowd counting using RGB-D, comprising a CmCaF module coupled with an innovative Fine-Coarse guidance technique.

2) RGB-Thermal: MMCCN [32] adopts an architecture supplemented with modality alignment and adaptable fusion modules, merging RGB and thermal information for crowd counting. I-MMCCN [33] incorporates a new loss considering integral for density map and hard example mining head to refine the MMCCN. IADM [34] aims to facilitate cross-modality feature extracting, with gating units to fuse RGB and thermal information. Tang et al. [35] elucidate the implementation of TAFNet, designed towards the extraction of combination features of RGB-T via utilizing a shared mainstream. MAFNet introduced by Chen et al. [31] has been designed with the aim of effectively capturing and considering long-term contextual features acquired from both RGB and thermal modalities. MAT presented in [36] aims to adequately harness the diverse information available from multi-modality by leveraging their complementarity. The study presented by Li et al. [37] recommends the utilization of a discriminative feature representation method that spans across modality domains to evaluate crowd distribution. Purely CNN-based techniques [32], [34], [37], [38] have limited effectiveness in appropriately integrating cross-modality features. A small number of models [33], [35] integrate vanilla attention mechanisms to fuse cross-modal features, but these models do not completely extract intermodality interaction information. Some other models [31], [36] utilize transformer-based ideas to fuse cross-modality feature representation. While transformers are highly effective at capturing contextual information in long sequences using self-attention, these approaches generate sequence features by patches. However, the absence of location embeddings hinders their capability to extract crowd features following modality interaction. Therefore, further advancements are needed to bridge the gap in crowd density estimation using RGB and thermal image interaction.

III. Proposed Method

This research introduces MIANet, a novel RGB-thermal crowd image density estimation and counting model. MIANet incorporates two modality-specific features using Tri-level Modality Interaction Attention (MIA) modules arranged in a stacked manner. To achieve superior feature fusion outcomes, Grid units amalgamate level-by-level features to form Grid Block. Specifically, MIANet includes the RGB-specific underlying network, the thermal-specific underlying network, three MIA modules, the Grid Block, and the backend regressor network.

A. Overview of Network Architecture

As depicted in Fig. 2, MIANet composes two backbones of RGB and thermal modality followed by Tri-level Modality Interaction Attention (MIA) in a stacked manner to extract multi-level features.

Within the MIA module, Multi-Scale Attention (MSA) improves feature extraction in non-uniform crowd densities within individual images and Channel Direction Attention (CDA) serve to reduces redundancy of complementary information between modalities. To sufficiently interact information of the two modalities, shared features from distinct levels of MIA modules should combine. We proposed a Grid Block to combine level-by-level features to extract polished fusion features. An element-wise addition operation is adopted, with the outputs feeding into a regressor called “Reg”, as illustrated in Fig. 2, to produce optimal-quality density maps and facilitate crowd counting. Two front-end backbones are developed using the initial 16 CNN layers of VGG19 [43]. The regressor (named Reg in Fig. 2) is developed using the CSRNet [21].

Fig. 2 displays that MIANet takes two inputs of RGB and thermal image pairs denoted as V and T, respectively, and lastly outputs a density map $\hat{D}$. The dual backbones are accountable for the extraction of attributes from RGB and thermal images, correspondingly. To simplify the notation, the initial features extracted are denoted as F_{v} and F_{t} (also written as F_{v}^{1} and F_{t}^{1}), respectively, obtained from the two front-end segments. To enable the effective fusion of features extracted and leverage the complementarity between RGB and thermal, three consecutive MIA modules are connected in a stacked manner after the two backbones. Additionally, the Grid Block

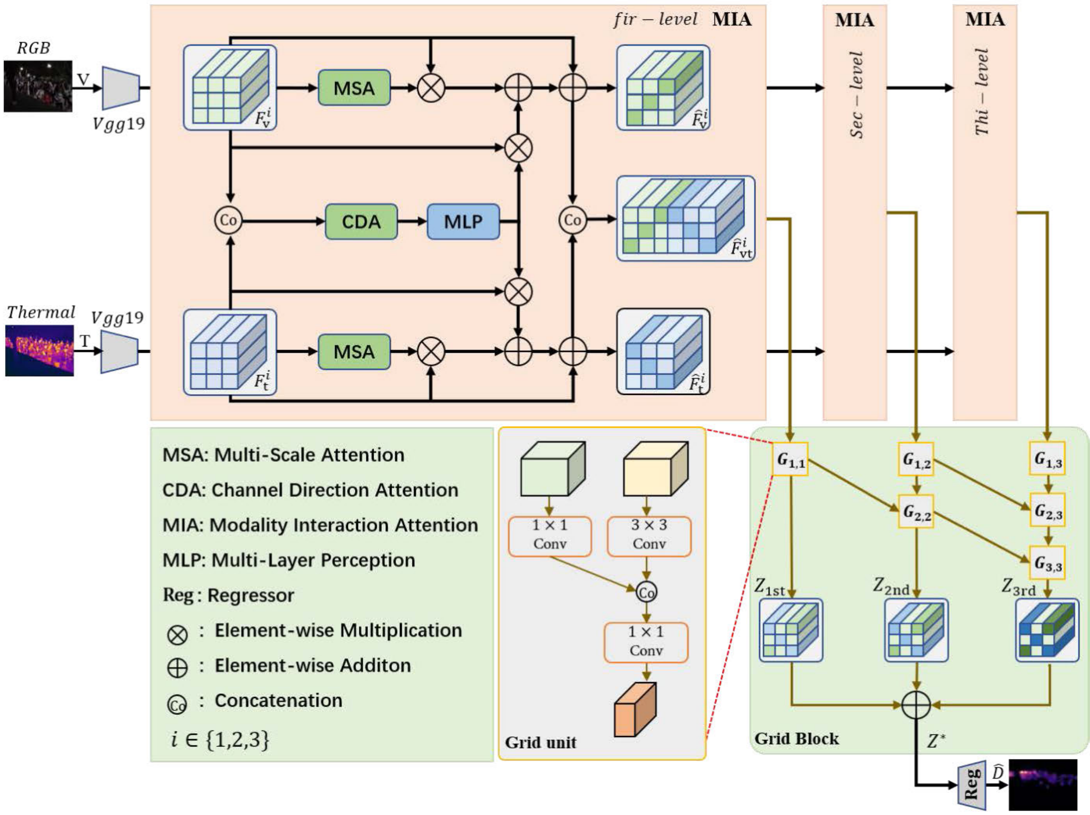


Fig. 2. MIANet for RGBT crowd counting. With the proposed MIA connected to the front-end network in a tri-level stacking method, and grid block combining level-by-level features, our method can adequately leverage modality-wise complementarity. "G" represents the Grid unit, which forms a grid block. Note that "Reg" stands for regressor.

facilitates level-by-level feature combination, which further enhances the quality of combined features. The operation of MIA can be formulated as follows,

$$\hat{F}_v^i, \hat{F}_t^i, \hat{F}_{vt}^i = \mathrm{MIA}(F_v^i, F_t^i), \tag{1}$$

where $i \in \{1, 2, 3\}$, $\hat{F}_v^i$, $\hat{F}_t^i$ are the enhanced features of F_v^i and F_t^i, respectively, and $\hat{F}_{vt}^i$ is refined and shared features of two modality features. And then $\hat{F}_v^i$ and $\hat{F}_t^i$ are fed into the following MIA modules to further refine and augment complementarity.

The Grid unit combines different level share features of MIA modules to get three level features $Z_{1\mathrm{st}}$, $Z_{2\mathrm{nd}}$, and $Z_{3\mathrm{rd}}$, which are fed into element-wise addition operation to get a better fusion feature Z^*. At last, Z^* is forwarded to a regressor for superior-quality density maps.

B. Modality Interaction Attention

1) Multi-Scale Attention known as MSA is intended to produce a spatial influence map to improves feature extraction in non-uniform crowd densities within individual images. Fig. 3 (a) illustrates a more concrete computing process. To extract multi-scale features and focus on an informative location, intermediate feature F is first fed to four encoding branches with kernels of different sizes. Then pass through a 2D convolutional layer to yield an influence map by using multi-scale connections between the diverse elements present. The 2D spatial attention map $A_s \in \mathbb{R}^{1\times H\times W}$ encodes where to accentuate or dampen the importance of specific elements.

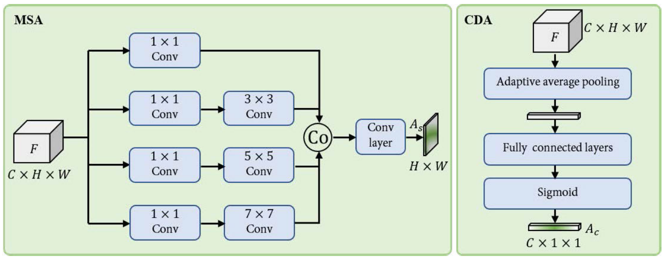


Fig. 3. The computational procedure of (a) Multi-scale attention and (b) Channel direction attention. Circled Co in the figure denotes the process of concatenating in a channel-specific manner.

The operation of MSA can be expressed using the following formulation,

$$A_s = \mathrm{MSA}(F), \tag{2}$$

2) Channel Direction Attention named CDA is used to consider the influence of different modalities from channel direction on crowd features. Fig. 3 (b) illustrates a more detailed computing process. Average pooling has been the predominant method used for information aggregation [44], [45]. For gleaning valuable information from crowd images, the mid-level feature map $F \in \mathbb{R}^{C\times H\times W}$ undergoes an adaptive average pooling layer along the channel direction. This modified map is then input to the fully connected layer to yield the attention map $A_c \in \mathbb{R}^{C\times 1\times 1}$. The function of CDA is expressed as follows,

$$A_c = \mathrm{CDA}(F), \tag{3}$$

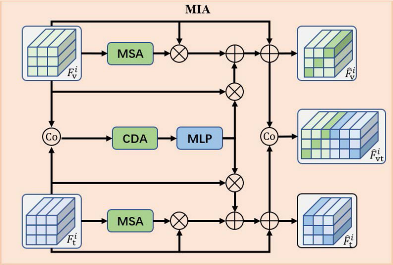


Fig. 4. MIA consists of MSA and CDA. MIA outputs enhanced and polished features.

To effectively utilize the complementarity among different modalities, we suggest the implementation of Modality Interaction Attention (MIA). As described in Fig. 4, MIA takes F_v^i and F_t^i as inputs and produces improved features accordingly, $\hat{F}_v^i$, $\hat{F}_t^i$, and a shared feature $\hat{F}_{vt}^i$. To simplify the presentation in this section, we will denote F_v^i, F_t^i, $\hat{F}_v^i$ and $\hat{F}_t^i$ as F_v, F_t, $\hat{F}_v$ and $\hat{F}_t$. In general, two input feature vectors of MIA module are first fed into the MSA to obtain spatial attention maps, respectively, and the concatenated feature vectors along the channel direction are passed through CDA, and Multi-Layer-Perception (MLP is utilized for diminishing the channel dimensions.) to produce a channel attention map. Ultimately, enhanced and polished features are received by computing the element-wise multiplication and addition of attention maps and input features.

3) Modality Interaction Attention, denoted as MIA has been formulated for the purpose of filtering and weighting features from multiple perspectives, capturing useful information and suppressing redundancy. We first explicitly describe the size of the input feature vectors $F_v, F_t \in \mathbb{R}^{C\times H\times W}$. The spatial resolution is represented by (H, W), channel number is denoted by C. The feature vectors are F_v and F_t from RGB and thermal modality, separately. And then F_v and F_t respectively are fed in MSA modules to produce a spatial map A_s^v and A_s^t according to Eq. 2, where $A_s^v, A_s^t \in \mathbb{R}^{1\times H\times W}$. To obtain cross-modality attention map, feature vectors are concatenated along the channel directions, and then connected feature $(F_v||F_t) \in \mathbb{R}^{2C\times H\times W}$ pass through CDA to produce channel attention map $A_c' \in \mathbb{R}^{2C\times 1\times 1}$ according to Eq. 3. MLP performs a 2D convolution operation with a kernel size of 1 to maintain the same channel size, resulting in $A_c \in \mathbb{R}^{C\times 1\times 1}$. The enhanced and refined features through the MIA module can be calculated by,

$$\begin{aligned}\hat{F}_v &= F_v + F_v \times A_s^v + F_v \times A_c,\\ \hat{F}_t &= F_t + F_t \times A_s^t + F_t \times A_c,\\ \hat{F}_{vt} &= (\hat{F}_v||\hat{F}_t),\end{aligned} \tag{4}$$

where $\hat{F}_v, \hat{F}_t \in \mathbb{R}^{C\times H\times W}$, and || denotes a channel-wise concatenation. Finally, all enhanced and refined features $\hat{F}_v$ and $\hat{F}_t$ are propagated forward into the subsequent MIA modules within the respective branches for further processing. Shared feature $\hat{F}_{vt}$ is fed into the Grid unit to extract polished features.

C. Grid Block

Grid Block consists of several Grid units between different levels of MIA. Grid unit combines level-by-level feature maps of MIA modules to extract a better fusion feature. The Grid Unit has two inputs from different levels of MIA, as shown in Fig. 2. In our work, we adopt Tri-level MIA, to get a better fusion feature, we utilize a Grid unit to aggregate different level features in a progressive way due to different levels of features often containing different information. The Grid unit consists of a 1×1 kernel size for the front-level convolutional layer and a 3×3 kernel size for the back-level convolutional layer and outputs different level features.

The different level features are computed by,

$$\begin{aligned}Z_{1,1} &= \mathrm{G}_{1,1}((F_v^1||F_t^1), \hat{F}_{vt}^1),\\ Z_{1,2} &= \mathrm{G}_{1,2}((\hat{F}_v^1||\hat{F}_t^1), \hat{F}_{vt}^2),\\ Z_{1,3} &= \mathrm{G}_{1,3}((\hat{F}_v^2||\hat{F}_t^2), \hat{F}_{vt}^3),\\ Z_{2,2} &= \mathrm{G}_{2,2}(Z_{1,1}, Z_{1,2}),\\ Z_{2,3} &= \mathrm{G}_{2,3}(Z_{1,2}, Z_{1,3}),\\ Z_{3,3} &= \mathrm{G}_{3,3}(Z_{2,2}, Z_{2,3}),\end{aligned} \tag{5}$$

where $\mathrm{G}_{i,j}$ represents the grid unit in row i, column j, whose outcome is represented by $Z_{i,j}$, and || denotes a channel-wise concatenation. Different level feature $Z_{1st} = Z_{1,1}$, $Z_{2nd} = Z_{2,2}$, and $Z_{3rd} = Z_{3,3}$ are passed into weighted addition getting a polished feature Z^*.

D. Optional Fusion Operations

In our work, we propose a Tri-level MIA with Grid Block to fuse two modality features $F_v \in \mathbb{R}^{C\times H\times W}$ and $F_t \in \mathbb{R}^{C\times H\times W}$ from front-end networks. There are also some classic feature fusion operations as follows.

1) Gating is employed to fuse the attribute vectors. A gating function is utilized to automatically adjust reliance on the F_v and F_t attribute vectors obtained from the RGB and thermal modalities, respectively. Gating feature $Z_{gat} \in \mathbb{R}^{C\times H\times W}$ involves a balance between two distinct attribute vectors, utilizing an adjustable gate weighting parameter β, which can be adapted throughout the learning process, i.e.,

$$Z_{gat} = \beta \times F_v + (1-\beta) \times F_t. \tag{6}$$

2) Addition means element-wise addition to merging two modality-specify feature vectors F_v and F_t to get $Z_{add} \in \mathbb{R}^{C\times H\times W}$, i.e.,

$$Z_{add} = F_v + F_t. \tag{7}$$

3) Multiplication represents the product of the corresponding elements at the same position for feature vectors with the same shape to get $Z_{mul} \in \mathbb{R}^{C\times H\times W}$. The expression is presented below,

$$Z_{mul} = F_v \times F_t. \tag{8}$$

4) Concatenation indicates that the feature vectors with the same size are concatenated along the channel direction. F_v and F_t have same spatial size (H, W). The expression is presented below,

$$Z_{con} = F_v||F_t, \tag{9}$$

where $Z_{con} \in \mathbb{R}^{2C\times H\times W}$, $\oplus$ denotes concatenation along channel direction. To maintain the same size of the attribute

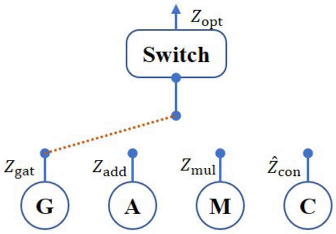


Fig. 5. Switch operation. Circled G, A, M and C indicate different fusion methods.

vector after the optional operation. We need to perform a convolution operation on Z_{con} with kernel size $= 1$ to obtain $\hat{Z}_{\text{con}} \in \mathbb{R}^{C\times H\times W}$.

5) Switch allows selecting an optimal way of fusing feature vectors from Gating, Addition, Multiplication, and Concatenation. Switch operation achieves $Z_{\text{opt}} \in \mathbb{R}^{C\times H\times W}$. Fig. 5 illustrates the process of switch operation. The output of optional operation is denoted as X, which can be $Z_{\text{gat}}, Z_{\text{add}}, Z_{\text{mul}}, \hat{Z}_{\text{con}}$, and Z_{opt}. X is fed to Reg to produce density map $\hat{D}$.

E. Loss Function

In numerous prior investigations, models were trained with Mean Squared Error (MSE) loss, wherein the Euclidean distance between the truth and estimated density map served as a consistency gauge. To facilitate subsequent experimental interpretation, we designate this as consistency loss $\mathcal{L}_{\text{c}}$. The formula is delineated below,

$$\mathcal{L}_{\text{c}} = \frac{1}{K}\sum_{k=1}^{K}\left\| D_k - \hat{D}_k \right\|_2^2, \tag{10}$$

where K symbolizes the sample numbers within the training sets. D_k and $\hat{D}_k$ denote the truth and estimated density maps, respectively.

The aforementioned consistency loss presumes that neighboring pixels of density maps operate independently, neglecting the interconnectedness within the local vicinity. To avoid the above problem caused by Euclidean distance, we adopt Bayesian loss [46] to measure local pattern consistencies, the loss function for one crowd image is denoted as,

$$\mathcal{L}_{\text{bayes}} = \sum_{n=1}^{N}\mathcal{F}(1 - \sum_{m=1}^{M} p(y_n|x_m)\cdot \mathbf{D}^{\text{est}}(x_m)), \tag{11}$$

where N is crowd counts in estimated density map $\mathbf{D}^{\text{est}}(x_m)$ and x_m represents the m-th pixel among a total of M pixels. y_n means the label corresponding to every pixel. $p(y_n|x_m)$ denotes posterior label probability. The function $\mathcal{F}$ corresponds to the L_1 distance employed throughout our experiment.

For the training of MIANet to converge, the correlation of local patterns within crowd images is considered. The loss function is served by $\mathcal{L}_{\text{bayes}}$, namely,

$$\arg\min_{\boldsymbol{\theta}} \mathcal{L} = \mathcal{L}_{\text{bayes}}, \tag{12}$$

where $\boldsymbol{\theta}$ represents learnable parameters within MIANet. The training process of MIANet is summarised in Algorithm 1.

Algorithm 1 Training Algorithm of MIANet

```
**Input** RGB images $\{V\}_1^K$, thermal images $\{T\}_1^K$, crowd counts $\{N\}_1^K$ and learning rate $\alpha$
**Output** Density estimation model MIANet denoted $\boldsymbol{\theta}$
Randomly initialize $\boldsymbol{\theta}$.
while loss not converged do
    Get $F_{\text{v}}$ and $F_{\text{t}}$ by frontend networks as intial features.
    Get $A_{\text{s}}^{\text{v}} \leftarrow \text{MSA}(F_{\text{v}})$, $A_{\text{s}}^{\text{t}} \leftarrow \text{MSA}(F_{\text{t}})$ based on Eq. 2
    Get $A_{\text{c}} \leftarrow \text{MLP}(\text{CDA}(F_{\text{v}}||F_{\text{t}}))$ based on Eq. 3
    Get $\hat{F}_{\text{v}}^i, \hat{F}_{\text{t}}^i, \hat{F}_{\text{vt}}^i \leftarrow \text{MIA}(F_{\text{v}}^i, F_{\text{t}}^i)$ where $i \in \{1, 2, 3\}$ by MIA based on Eq. 1 and Eq. 4
    Get $Z_{1,1}, Z_{2,2}, Z_{3,3} \leftarrow$ Grid Block$(\hat{F}_{\text{v}}^i, \hat{F}_{\text{t}}^i, \hat{F}_{\text{vt}}^i, F_{\text{v}}, F_{\text{t}})$ by Grid Block based on Eq. 5
    Get $Z_{\text{1st}} \leftarrow Z_{1,1}$, $Z_{\text{2nd}} \leftarrow Z_{2,2}$, and $Z_{\text{3rd}} \leftarrow Z_{3,3}$
    Get $Z^* \leftarrow sum(Z_{\text{1st}}, Z_{\text{2nd}}, Z_{\text{3rd}})$
    Get $\hat{D} \leftarrow Reg(Z^*)$
    Calculate Bayesian loss $\mathcal{L}_{\text{baye}}$ based on Eq.11
    Update $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha\nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{baye}}$
Return $\boldsymbol{\theta}$
```

TABLE I
STATISTICS OF CROWD DATASETS

Dataset	DroneRGBT	RGBT-CC
Year	2020	2021
Number of Images	3600	2030
Average Resolution	512 × 640	640 × 480
Total number of annotated persons	175,698	138,389
Min number of per image	1	8
Average number of per image	48	68
Max number of per image	403	701

IV. EXPERIMENTS

A. Datasets

1) **DroneRGBT** [32] is the initial crowd counting dataset from drones perspective, featuring both RGB and thermal infrared images. It encompasses image pairs acquired from diverse locations, exhibiting substantial diversity in perspective, scale, and background complexities.

2) **RGBT-CC** [34] is recorded for analyzing urban crowd dynamics with diverse density levels. Comprehensive information regarding datasets is presented in Table I. As shown in Fig. 6, the training set and the test set of the two datasets exhibit similar distributions. This ensures that the model's performance on the test set will more accurately reflect its performance on previously unseen data. Additionally, this similarity allows for a more precise evaluation of each model's strengths and weaknesses, aiding in the selection of the most suitable model.

B. Baseline Methods

- **MMCCN** (2020) [32] adopts an architecture supplemented with modality alignment and adaptable fusion modules, merging RGB and thermal images.

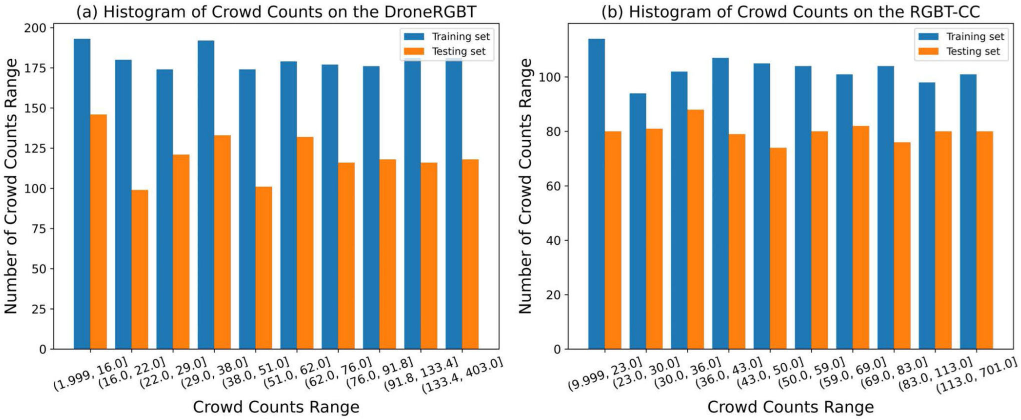


Fig. 6. Histogram of the crowd counts range.

- **IADM** (2021) [34] aims to facilitate cross-modality feature extracting, with gating units to fuse RGB and thermal information.
- **I-MMCCN** (2021) [33] incorporates a new loss considering integral for density map and hard example mining head to refine the MMCCN [32].
- **TAFNet** (2022) [35] is introduced as a three-branch network, designed towards the extraction of combination features of RGB-T via utilizing a shared mainstream.
- **MAT** (2022) [36] is devised to optimally utilize the complementary information present within distinct modalities.
- **CSCA** (2022) [47] is designed to facilitate the convenient integration of any modality-specific architecture.
- **R2T** (2022) [37] denotes model of [37]. The model comprises two stages; the first is dedicated to acquiring the intrinsic discriminative feature, which is then fused with the second stage to obtain the count results.
- **CSA-Net** (2022) [48] is designed to aggregate multi-modality features to realize cross-modal feature representation by considering multi-scale context.
- **CNCTrans** (2022) [49] combines Convolutional Neural Networks with a newly proposed cross-modal transformer for effectively processing cross-modal data.
- **DEFNet** (2022) [38] aims to accomplish RGBT crowd counting, encompassing a robust data amplification mechanism for the integration of synergetic attributes across dual modalities.
- **UCCF** (2023) [39] is a multi-task framework that encompasses an image integration network, coupled with a network dedicated to crowd counting.

C. Evaluation Metrics

Root Mean Square Error (RMSE) and the Grid Average Mean absolute Error (GAME) are employed to evaluate both the MIANet and the baseline models. The advantage of RMSE is its sensitivity to large errors, while GAME is more intuitive and less affected by large errors. To comprehensively evaluate the model's performance, both RMSE and GAME are used. The corresponding formulas are provided below,

$$\mathrm{RMSE} = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(D_s - \hat{D}_s\right)^2}, \tag{13}$$

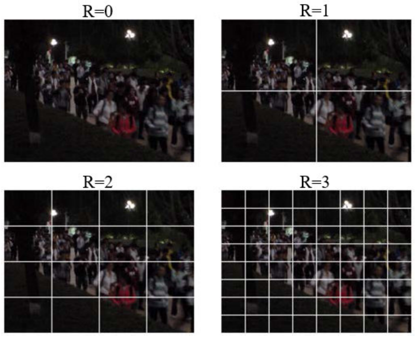


Fig. 7. An example of evenly splitting crowd image for GAME(R).

$$\mathrm{GAME}(R) = \frac{1}{S}\sum_{s=1}^{S}\sum_{r=1}^{4^R}\left|D_s^r - \hat{D}_s^r\right|, \tag{14}$$

where S signifies the amount of images present in the test set, while D_s and $\hat{D}_s$ correspond to the truth and predicted crowd counts in s-th image, respectively. To calculate the GAME(R), the image is divided into 4^R discrete patches, evenly split both vertically and horizontally as shown in Fig. 7. In our study, $R \in \{0, 1, 2, 3\}$. In GAME(R), D_s^r and $\hat{D}_s^r$ correspond to the truth and predicted crowd counts in s-th image's r-th patch, respectively. GAME(R) involves dividing the image into 4^R patches, calculating the absolute difference between ground truth and estimated crowd counts in each patch, and then averaging these differences. If $R = 0$, the GAME formula degenerates into the calculation for Mean Absolute Error (MAE).

D. Implementation and Hyperparameter Settings

We executed experiments on a Linux-based server featuring a single Intel(R) Core(TM) i9-10940X processor operating at 3.30GHz, complemented by a 24GB NVIDIA GeForce RTX 3090 GPU. The MIANet is implemented using the PyTorch [50]. The optimization of the trainable parameters during the network training is carried out using Adam [51]. The training process is governed with a learning rate set at 1×10^{-5} and limited to a maximum of 200 iterations. To prevent over-fitting during the training, we utilize a validation-based early stopping mechanism [52]. By monitoring the Mean

TABLE II
COMPARISON BETWEEN MIANET AND BASELINE METHODS ON DRONERGBT

Method	Journal/Venue &Year	GAME(0)↓[3]	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
MMCCN	ACCV 2020	7.27	-	-	-	11.45
UCCF	IVC 2023	7.96	-	-	-	12.50
IADM	CVPR 2021	10.40	12.10	14.97	19.55	17.30
I-MMCCN	IC-NIDC 2021	9.41	10.85	13.23	16.52	15.10
DEFNet	TITS 2022	8.27	10.20	13.06	16.92	13.59
MIANet	ours	**6.74**	**8.64**	**11.49**	**16.31**	**10.58**

TABLE III
COMPARISON BETWEEN MIANET AND BASELINE METHODS ON RGBT-CC

Method	Journal/Venue &Year	GAME(0)↓[3]	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
IADM	CVPR 2021	15.61	19.95	24.69	32.89	28.18
I-MMCCN	IC-NIDC 2021	16.15	20.04	23.61	30.00	26.51
TAFNet	ISCAS 2022	12.38	16.98	21.86	30.19	22.45
MAT	ICME 2022	13.65	18.03	22.94	28.65	22.53
CSCA	ACCV 2022	14.32	18.91	23.81	32.47	26.01
R2T	KBS 2022	16.92	20.72	25.90	32.10	32.91
CSA-Net	ESWA 2022	15.77	19.40	24.14	30.14	29.17
CNCTrans	IVC 2022	13.96	17.98	23.03	31.15	24.55
DEFNet	TITS 2022	15.31	19.41	23.47	29.96	27.94
MIANet	ours	**11.97**	**15.65**	**19.93**	**27.54**	**22.17**

Absolute Error of the validation set, the training process is halted if no improvement is observed for ten consecutive epochs. The DroneRGBT contains 1807 couples of training images, 600 couples of validating images, and 800 couples of testing images. In RGBT-CC, the training set is 1030 RGB-thermal pairs (Brightness:510, Darkness:520), the validation set is 200 pairs (Brightness:97, Darkness:103), and the test set is 800 pairs (Brightness:406, Darkness:394). The parameters of the baseline methods are determined according to the specifications stipulated in the original papers.

E. Comparison With Baseline Methods

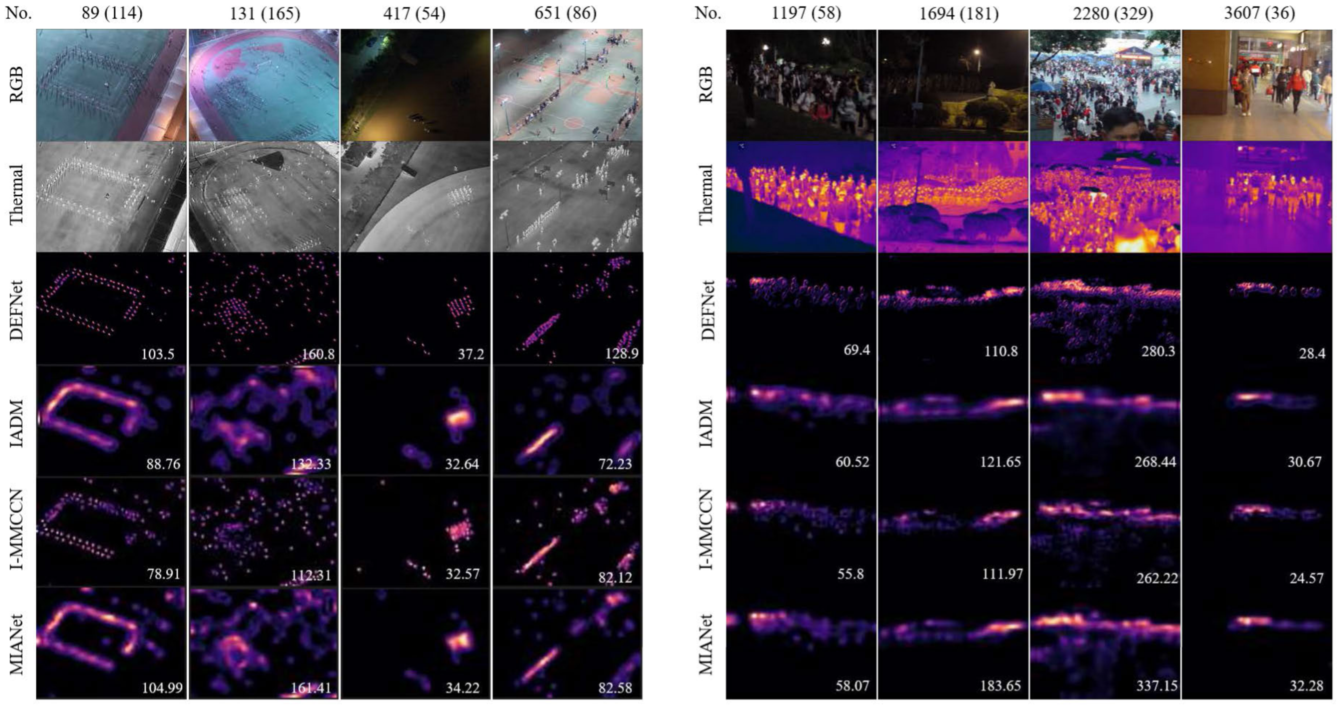
Our study presents a thorough evaluation of the efficacy of MIANet in contrast to baseline approaches through a quantitative analysis. In addition, we qualitatively assess the accuracy of the estimation by visualizing estimation maps.

The performance comparison between MIANet and the latest models on both datasets is presented in Table II and Table III, respectively. The results indicate MIANet exhibits superior performance surpassing the latest existing methods.

In Table II, the results show our MIANet performs great accuracy on DroneRGBT. Compared to the second-best method, MMCCN [32], MIANet reduces the GAME(0) value from 7.27 to 6.74, achieving a reduction of approximately 7.29%. Similarly, the RMSE value decreases from 11.45 to 10.58, representing a reduction of about 7.60%. Earlier cutting-edge models like UCCF (RMSE: 12.50) and I-MMCCN (RMSE: 15.10), which utilize vanilla attention mechanisms, outperform IADM (RMSE: 17.30) without attention mechanism. These improvements show that the attention mechanism effectively avoids redundancy during modality interaction and fusion, leading to better complementary features between RGB and thermal images. Compared to UCCF (RMSE: 12.50) and I-MMCCN (RMSE: 15.10), MIANet (RMSE: 10.58) demonstrates superior efficiency, achieving reductions of approximately 15.36% and 29.17% respectively in RMSE. This performance improvement is attributed to MIANet's MSA module, leveraging multi-scale attention to effectively address challenges posed by uneven crowd density during feature extraction and fusion.

The table III assesses the performance of 10 diverse models for crowd counting in terms of GAME(0-3) and RMSE on RGBTCC. From the table, MIANet outperforms all baseline models due to MIA module decreases redundant feature during feature interaction and Grid Block consolidates diverse features from various MIA levels. Additionally, we can observe that TAFNet and CNCTrans also achieve relatively good performance in terms of GAME(0) and RMSE scores due to TAFNet with vanilla attention mechanism and CNC-Trans adopt transformer concepts. Meanwhile, R2T without attention mechanism proves to be less effective in crowd counting, obtaining the highest GAME(0) and RMSE value among all evaluation methods. These findings demonstrate that attention mechanisms can effectively avoid redundant information, improve feature complementarity, and enhance feature extraction during the interaction between RGB and thermal images. Compared to the second-best method, TAFNet [35], our MIANet reduces the GAME(0) value from 12.38 to 11.97, a decrease of approximately 3.31%, and the RMSE value from 22.45 to 22.17, a decrease of about 1.25%. This finding suggests that MIANet's use of multi-scale attention (MSA) effectively addresses the challenges posed by uneven crowd density during feature extraction and fusion. Our MIANet reduces the GAME(0) from 13.96 to 11.97, which is a reduction of approximately 14.26%, and the RMSE from 24.55 to 22.17, which is a reduction of approximately 9.70% compared to CNCTrans [49]. This finding suggests that while transformers are effective at capturing contextual information

[3] ↓ indicates that the lower the indicator, the better the method.



(a) DroneRGBT

(b) RGBT-CC

Fig. 8. Visualized examples of density maps on (a) DroneRGBT and (b) RGBT-CC. The 1st and 2nd rows show RGB and thermal inputs, respectively. The other rows show estimation density maps by different methods. The white number at the bottom right corner of images indicates crowd counts. The numbers within the parentheses represent corresponding ground truth crowd counts.

over long sequences by self-attention, CNCTrans [49] relies on forming sequence features through patches. This approach limits its ability to extract crowd features after modality interaction due to the absence of location embedding.

Overall, on both datasets, MIANet showed the best performance in RGB-thermal crowd counting. MIANet reconciles the complementarity inherent without introducing redundant information and addresses non-uniform crowd density in image pairs in modality interactions for feature fusion, tackling two challenges identified in existing approaches. Applying it to the task of analyzing and monitoring dense crowds can better ensure crowd safety.

A visual comparison of the partial cutting-edge models with MIANet on the two datasets is presented in Fig. 8 (a) and (b). The figures present the serial numbers of the input image pairs in the test datasets as the top numbers, while the numbers within the parentheses represent the corresponding ground truth crowd counts for each image. Furthermore, the estimated crowd counts for each respective method are indicated by the white number located in the bottom right of individual image. Evidently, the comparison of these results indicates that the crowd counts predicted by MIANet closely correspond to truth values, further demonstrating the superior performance of MIANet.

Through a closer examination of the image pairs with serial number 651 in the test dataset in Fig. 8 (a), it is challenging to determine if the top left of current thermal crowd image is crowded. In contrast, the corresponding RGB image provides a more definitive indication that the object in the top left corner is a tall streetlamp. By combining the two modalities, different methods produce crowd density maps that resemble each other closely. Nonetheless, our proposed method exhibits the most prominent performance in accurately estimating crowd counts.

Observing the test image pairs of serial number 417 in Fig. 8 (a) and serial numbers 1197 and 1694 in Fig. 8 (b), we can find the poor lighting conditions make it challenging to clearly discern the distribution and quantity of the crowds in RGB modality. However, the corresponding thermal images allow us to approximate the crowd distribution with greater ease. The fusion of these two modalities for the task of crowd counting produces varying results across different methods. Although most methods can provide an approximate indication of crowd distribution, our proposed MIANet outperforms all other methods regarding crowd counting accuracy.

F. Ablation Studies

We accomplish ablation studies using two distinct datasets: RGBTCC and DroneRGBT. RGB and thermal crowd images captured under monitoring lenses exhibit spatial stratified heterogeneity (SSH) [53], [54], [55], [56], [57] due to the complexity and diversity of crowd environments, non-uniform crowd density, and varying shooting angles. Consequently, the crowd image data processed by our network model inherently displays SSH properties, which may confounded when applying global modeling techniques. The ablation studies demonstrate the effectiveness of each component within the Modality Interaction Attention (MIA) module, showing that MIA can alleviate the confounding effects of global modeling. The baseline model employed in the ablation study utilizes a model featuring a dual VGG19 [43] architecture, coupled

TABLE IV

MIANet and Its Variants on DroneRGBT

MIA		Grid Block	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
MSA	CDA						
×	×	×	7.02	8.95	11.81	16.54	11.22
✓	×	×	6.92	8.93	11.91	16.88	10.93
×	✓	×	6.93	8.93	11.85	16.36	11.16
✓	✓	×	6.82	8.64	11.58	16.33	10.71
✓	✓	✓	**6.74**	**8.64**	**11.49**	**16.31**	**10.58**

TABLE V

MIANet and Its Variants on RGBT-CC

MIA		Grid Block	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
MSA	CDA						
×	×	×	14.70	18.05	21.80	28.87	28.36
✓	×	×	14.07	17.71	21.85	29.31	26.85
×	✓	×	13.93	17.39	21.43	29.20	26.27
✓	✓	×	12.83	16.66	20.80	28.44	23.68
✓	✓	✓	**11.97**	**15.65**	**19.93**	**27.54**	**22.17**

with the same regressor head as present in our proposed network. To ensure the credibility of the experimental results, we progressively integrated our proposed modules into the above baseline model to generate different variants. Evaluation of the performance of the variants is conducted using two metrics in comparison to the baseline. The results are shown in Tabel IV and V.

Compared to the baseline model, MIA substantially enhances counting accuracy. It achieves remarkable improvements of 2.8% and 4.5% in GAME(0) and RMSE on the DroneRGBT dataset, and even more substantial gains of 12.7% and 16.5% on the RGBT-CC dataset. Integrating all modules into the baseline model further elevates its counting accuracy, with significant improvements of 4.0% and 5.7% on DroneRGBT, and a substantial 18.6% and 21.8% on RGBT-CC. Initially devoid of added modules, the baseline model gradually incorporates proposed components (MSA, CDA, and Grid Block) in successive variants. The highest performance is achieved when the baseline model includes all three modules. These results underscore MIANet's superiority over the baseline and its variants, as evidenced by the marked reductions in GAME scores and RMSE values on the DroneRGBT and RGBT-CC datasets.

The findings demonstrate that the proposed MIANet enhances performance in comparison to the model variants, as reflected in the lower GAME scores and RMSE on DroneRGBT and RGBTCC in Table IV and Table V, respectively. Both tables provide evidence to suggest that each module integrated into the MIANet has a positive effect on crowd counting. Removal of any of the modules from the MIANet resulted in a decrease in the performance of RMSE or GAME(R), highlighting the effectiveness of these modules (MIA, MSA, CDA and Grid Block) in enhancing the ability of crowd feature extraction and fusion.

Overall, MIA module uses a fine-grained attention mechanism to filter and weight features from multiple perspectives, capturing useful information and suppressing redundancy. To be specific, MSA selectively emphasizes task-relevant features from multi-scale views while suppressing redundant or noisy ones, reducing the impact of irrelevant information. CDA dynamically adjusts the contributions of modality features by calculating the importance weight of each modality, highlighting significant modality features while reducing the influence of redundant ones.

G. Comparison of Fusion Operations

We also utilize a VGG19 [21] network with a single-modality image as input to further verify the optional operations. Optional operations are Gating, Addition, Multiplication, Concatenation and Switch.

We conduct a comparative analysis of the counting accuracy achieved under various fusion operations on the DroneRGBT and RGBTCC, which result is shown in Table VI and Table VII, respectively.

It can be found our Tri-level MIA with Grid Block performs better than Gating, Addition, Multiplication, Concatenation, and Switch operations. Additionally, Other operations (except switch operation) perform erratically in terms of the performance of the evaluation metrics. For example, the Gating operation performs the worst in GAME(0) and GAME(1) on both datasets, the Addition performs the worst in GAME(2) on DroneRGBT, the Multiplication performs the worst in GAME(2) and GAME(3) on RGBT-CC. The Switch operation selects the fusion operation that best matches the selection rule (The lowest GAME(0)) from the four fusion operations (Gating, Addition, Multiplication, and Concatenation) based on the selection rule. Compared to Gating operations, the improvements of Tri-level MIA are 5.9% and 5.8% for GAME(0) and RMSE on DroneRGBT, 22.1% and 17.5% for GAME(0) and RMSE on RGBT-CC, respectively. Within DroneRGBT, Concatenation gains the second-best, but Addition acts the second-best on the RGBTCC. Switch can utilize gating, addition, multiplication, and concatenation to reach the second-best on both datasets.

H. Effect of Illumination Conditions

We partition the RGBT-CC dataset into two subsets and evaluate the performance under different illumination conditions to furnish supplementary proof supporting the complementarity of multimodal images. A single CSRNet [21] network with single-modality as input is assumed to be the baseline to verify RGB modality and thermal modality are both beneficial to estimate crowd density. Table VIII highlights that the use of both RGB and thermal image pairs as network inputs led to a noticeable improvement in crowd density estimation. The proposed MIANet, along with IADM, TAFNet, and CSCA, generate more accurate results than the unimodal single CSRNet under different illumination settings. Additionally, the MIANet stands out as the best-performing method in comparison to IADM, TAFNet, and CSCA.

Compared to RGB-modality as network inputs in brightness conditions, the improvements of MIANet are 41.7% and 41.1% for GAME(0) and RMSE, the performance is improved by 13.3% and 28.3% in IADM [34] method, by 33.7% and 46.6% in TAFNet [35], by 38.7% and 45.5% in CSCA [47], respectively. Compared to thermal-modality in brightness conditions, the improvements of MIANet are 45.7% and 34.1%

TABLE VI

THE PERFORMANCE OF DIFFERENT INPUTS AND DIFFERENT OPTIONAL OPERATIONS ON DRONERGBT

Input Image	Optional Operations	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
RGB	-	7.20	9.12	11.99	16.69	11.44
Thermal	-	7.72	9.82	12.83	17.76	12.32
RGB-Thermal	Gating	7.16	9.09	11.79	16.16	11.23
	Addition	7.03	9.01	12.04	16.92	11.02
	Multiplication	7.07	8.90	11.58	16.16	11.23
	Concatenation	7.02	8.95	11.81	16.54	11.22
	Switch	7.02	8.95	11.81	16.54	11.22
	Tri-level MIA (ours)	**6.74**	**8.64**	**11.49**	**16.31**	**10.58**

TABLE VII

THE PERFORMANCE OF DIFFERENT INPUTS AND DIFFERENT OPTIONAL OPERATIONS ON RGBT-CC

Input Image	Optional Operations	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
RGB	-	29.06	33.26	38.42	46.78	65.42
Thermal	-	17.18	20.51	24.01	30.96	29.89
RGB-Thermal	Gating	15.36	18.35	21.91	29.03	26.86
	Addition	14.57	17.58	21.08	28.07	26.54
	Multiplication	14.74	18.26	22.50	29.68	30.69
	Concatenation	14.70	18.05	21.80	28.87	28.36
	Switch	14.57	17.58	21.08	28.07	26.54
	Tri-level MIA (ours)	**11.97**	**15.65**	**19.93**	**27.54**	**22.17**

TABLE VIII

THE COMPARISON WITH OTHER METHODS IN DIFFERENT ILLUMINATION CONDITIONS ON RGBT-CC

Illumination	Input Data	Method	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
Brightness	RGB	CSRNet	23.49	30.14	37.47	48.46	45.40
	Thermal	CSRNet	25.21	28.98	34.82	42.25	40.60
	RGB-Thermal	MIANet (ours)	**13.69**	**18.01**	**22.96**	**31.41**	26.76
		IADM	20.36	23.57	28.49	36.29	32.57
		TAFNet	15.57	20.65	26.67	36.17	**24.25**
		CSCA	14.41	18.85	24.71	34.20	24.74
Darkness	RGB	CSRNet	44.72	51.70	57.45	66.21	87.81
	Thermal	CSRNet	17.97	23.38	28.39	34.95	33.74
	RGB-Thermal	MIANet (ours)	**13.88**	**18.28**	23.46	32.17	**25.15**
		IADM	15.44	19.23	23.79	**30.28**	29.11
		TAFNet	14.20	19.20	24.00	31.63	27.50
		CSCA	14.22	18.97	**22.89**	30.69	27.25

for GAME(0) and RMSE, the performance is improved by 19.2% and 19.8% in IADM [34] method, by 38.2% and 40.3% in TAFNet [35], by 42.8% and 39.1% in CSCA [47], respectively. We can easily find that the RGB modality brings greater gains for the crowd counting task compared to the thermal in bright conditions. Considering both modalities can significantly boost the effectiveness of crowd estimation.

Under dark conditions, MIANet reduces GAME(0) and RMSE values of RGB-modality as inputs from 44.72 to 13.88 and from 87.81 to 25.15, respectively. MIANet also reduces GAME(0) and RMSE values of thermal-modality as inputs from 17.97 to 13.88 and from 33.74 to 25.15, respectively. It can be deduced that, in conditions of low luminance, the efficacy of utilizing the thermal-based approach surpasses that of the RGB methodology for estimating. The susceptibility of RGB-based data to fluctuations in light intensity diminishes the accuracy of its crowd estimation.

The findings presented reveal that thermal image is instrumental in identifying potential pedestrians in a disordered environment or in the presence of inadequate lighting. Moreover, the utilization of RGB image is valuable in the removal of thermally warm non-pedestrian items from thermal imagery.

TABLE IX

THE PERFORMANCE OF DIFFERENT LOSS FUNCTIONS IN MIANET

Dataset	Loss Function	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
DroneRGBT	$\mathcal{L}_{\text{bayes}}$	**6.74**	**8.64**	**11.49**	**16.31**	**10.58**
	$\mathcal{L}_{\text{c}}$	9.08	10.55	13.26	17.95	15.01
	$\mathcal{L}_{\text{bayes}} + \mathcal{L}_{\text{c}}$	7.36	9.24	12.13	16.88	12.88
RGBTCC	$\mathcal{L}_{\text{bayes}}$	**11.97**	**15.65**	**19.93**	**27.54**	**22.17**
	$\mathcal{L}_{\text{c}}$	17.18	22.05	25.47	33.90	31.55
	$\mathcal{L}_{\text{bayes}} + \mathcal{L}_{\text{c}}$	13.43	17.88	21.88	29.26	26.50

I. Effect of Loss Function

We have defined the loss function in Section III-E, to explore the impact of $\mathcal{L}_c$ or $\mathcal{L}_{\text{bayes}}$ for model training, we study $\mathcal{L}_c$ and $\mathcal{L}_{\text{bayes}}$ and their combination for MIANet training. Table IX shows the experimental results. Our results indicate MIANet achieves the best performance when trained by only using $\mathcal{L}_{\text{bayes}}$ reflected in its lowest GAME(0-3) and RMSE scores. In contrast, MIANet performs the worst using $\mathcal{L}_{\text{c}}$ training reflected in its highest GAME(0-3) and RMSE. The study underlines the phenomenon predicated on the assumption of Euclidean loss, which considers adjacent pixels as isolated entities and disregards local correlations in the density map. Additionally, the adoption of a hybrid of

TABLE X
COMPARISON OF RUNNING TIME AND FLOPs

Dataset	Method	Running time for each estimation (ms)	FLOPs (G)
DroneRGBT	IADM [34]	43.8	28.90
	DEFNet [38]	82.7	210.89
	I-MMCCN [33]	35.5	13.99
	MIANet (ours)	57.9	91.28
RGBTCC	IADM [34]	43.0	28.90
	DEFNet [38]	80.3	210.89
	I-MMCCN [33]	35.6	13.99
	MIANet (ours)	55.9	91.28

TABLE XI
GENERALIZATION ABILITY OF MIANet

Method	GAME(0)↓	GAME(1)↓	GAME(2)↓	GAME(3)↓	RMSE↓
RDNet [40]	4.96	-	-	-	7.22
DPDNet [30]	4.23	-	-	-	6.75
IADM [34]	4.38	5.95	8.02	11.02	7.06
MAT [36]	4.05	5.49	7.49	10.25	6.01
CmCaF [42]	4.03	-	-	-	5.81
MIANet (ours)	**3.63**	**5.26**	**7.26**	**10.03**	**5.43**

two loss functions can not achieve optimal results because $\mathcal{L}_c$ does not take into account the correlation between adjacent pixels while $\mathcal{L}_c$ affecting $\mathcal{L}_{\text{bayes}}$ makes it impossible to achieve optimal results.

J. Runtime Efficiency and FLOPs

In Table X, we compare the running time [7], [58] and floating point operations (FLOPs) [58], [59] of different deep learning models for crowd counting and density estimation on two RGBT datasets. Observing the experimental results, it is evident that the running time of the same method on the DroneRGBT dataset is longer than on the RGBTCC dataset. This may be due to the larger size and higher resolution of the images in the DroneRGBT dataset. I-MMCCN [33] exhibits the lowest FLOPs, leading to the shortest estimation time. In contrast, DEFNet [38] has the highest FLOPs, which results in the longest estimation time. Furthermore, IADM [34] requires the second shortest estimation time for crowd density, attributed to its higher FLOPs compared to I-MMCCN [33]. Despite the larger FLOPs of MIANet leading to longer estimation times, its execution time remains within the limits required for real-time applications.

K. Generalization Ability of MIANet

In order to demonstrate the generalization ability of MIANet, ShanghaiTechRGBD [40] dataset is used for training and testing, and the experimental results are shown in the Table XI. Our MIANet establishes a new satisfactory performance on the ShanghaiTechRGBD benchmark, achieving the lowest GAME(0) score of 3.63 and an RMSE of 5.43. This experiment demonstrates the universality and effectiveness of our MIANet for RGBD crowd counting. The likely reason is that MIANet suppresses redundant information and noisy ones while highlighting important crowd features during the interaction and fusion of the two modalities. Additionally, Grid Block of MIANet optimizes the representation of crowd features by aggregating the outputs of the MIA modules at different levels. These ensure the model has satisfactory generalization ability. SST [60] provides a safe framework to assess the generalizability of a study. By combining our work with the SST framework, the advantages of the generalization ability of our MIANet is also illustrated to a certain extent.

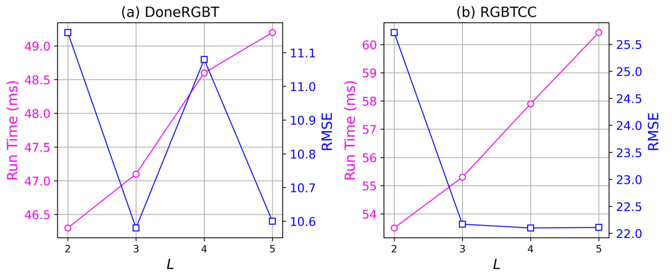


Fig. 9. Influence of MIA layer numbers L.

L. Influence of MIA Layer Numbers

To verify that a tri-level MIA optimally balances computational cost and crowd counting performance, we define the number of MIA as L and conduct several experiments on the DroneRGBT and RGBTCC datasets. The experimental results are shown in Fig. 9. As L values increase, the running time gradually escalates for both the DroneRGBT and RGBTCC datasets. Despite these variations, minimal changes in RMSE are observed across different L values, with notable results of 10.58 for $L = 3$ in the DroneRGBT dataset and 22.10 for $L = 4$ in the RGBTCC dataset.

Overall, $L = 3$ exhibits balanced performance across both datasets, offering a favorable compromise between computational efficiency, reflected in moderate running times, and model accuracy, supported by competitive RMSE values. Therefore, we recommend $L = 3$, representing tri-level MIA, as optimal for achieving an effective balance between computational efficiency and crowd counting performance of MIANet. This conclusion is consistent with the viewpoint of previous scholars [61] that excessive stacking of modules hampers the training of MIANet and diminishes counting performance. Additionally, progressive feature fusion refinement [31], [34], [35] enhances the overall quality of feature fusion.

V. CONCLUSION

This study introduces MIANet, a novel Modality Interaction Attention Network tailored for accurate crowd density and count estimation in challenging scenarios. MIANet innovatively integrates RGB and thermal camera image pairs to enhance precision in crowd counting and density estimation. The Modality Interaction Attention (MIA) module plays a pivotal role by facilitating efficient information extraction from each modality through mutual interaction. It effectively compensates for missing information while minimizing redundancy. Particularly noteworthy is the Multi-Scale Attention (MSA) module within MIA, adept at addressing non-uniform crowd density in modality interactions for feature extraction by considering multi-scale problems. Channel Direction

Attention (CDA) is used to assess the influence of different modalities on crowd features from the channel direction. MIANet leverages VGG19 as front-end networks for both modalities, employing a Tri-level MIA to merge and enhance modality-specific features. To optimize feature fusion further, a Grid Block composed of grid units is introduced, effectively combining multi-level features from MIA. Across two real-world benchmarks, MIANet achieves superior crowd counting performance as evidenced by substantial improvements in RMSE and GAME metrics. Ablation experiments highlight the efficacy of key components including MSA, CDA, MIA, and Grid Block. MIANet reconciles the complementarity inherent without introducing redundant information and addresses non-uniform crowd density in image pairs in modality interactions for feature fusion.

MIANet relies on the integration of both RGB and thermal images to achieve optimal performance. The absence of either image type can significantly reduce its effectiveness and limit its range of applications. Furthermore exploring the reduction of MIANet's network size through knowledge distillation appears to be a promising research direction.

References

[1] H. Xu, Z. Cai, R. Li, and W. Li, "Efficient citycam-to-edge cooperative learning for vehicle counting in ITS," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 16600–16611, Sep. 2022.

[2] N. Li, F. Chang, and C. Liu, "Spatial–temporal cascade autoencoder for video anomaly detection in crowded scenes," *IEEE Trans. Multimedia*, vol. 23, pp. 203–215, 2021.

[3] S. Wang, Y. Lyu, Y. Xu, and W. Wu, "MSCDP: Multi-step crowd density predictor in indoor environment," *Neurocomputing*, vol. 544, Aug. 2023, Art. no. 126296.

[4] L. Ciampi, C. Gennaro, F. Carrara, F. Falchi, C. Vairo, and G. Amato, "Multi-camera vehicle counting using edge-AI," *Expert Syst. Appl.*, vol. 207, Nov. 2022, Art. no. 117929.

[5] J. Shao, C. C. Loy, and X. Wang, "Scene-independent group profiling in crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2219–2226.

[6] J. Shao, K. Kang, C. C. Loy, and X. Wang, "Deeply learned attributes for crowded scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4657–4666.

[7] Y. Xu et al., "Adaptive feature fusion networks for origin-destination passenger flow prediction in metro systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 5296–5312, May 2023.

[8] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 46–58, Jan. 2020.

[9] N. Jiang et al., "Anti-UAV: A large-scale benchmark for vision-based UAV tracking," *IEEE Trans. Multimedia*, vol. 25, pp. 486–500, 2023.

[10] Y. Chen and Y. Lou, "A unified multiple-motion-mode framework for socially compliant navigation in dense crowds," *IEEE Trans. Autom. Sci. Eng.*, vol. 19, no. 4, pp. 3536–3548, Oct. 2022.

[11] H. Dong, M. Zhou, Q. Wang, X. Yang, and F.-Y. Wang, "State-of-the-art pedestrian and evacuation dynamics," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1849–1866, May 2019.

[12] Y. Xiao, J. Xu, M. Chraibi, J. Zhang, and C. Gou, "A generalized trajectories-based evaluation approach for pedestrian evacuation models," *Saf. Sci.*, vol. 147, Mar. 2022, Art. no. 105574.

[13] R. Wang, Y. Hao, L. Hu, J. Chen, M. Chen, and D. Wu, "Self-supervised learning with data-efficient supervised fine-tuning for crowd counting," *IEEE Trans. Multimedia*, vol. 25, pp. 1538–1546, 2023.

[14] X. Jiang et al., "Density-aware multi-task learning for crowd counting," *IEEE Trans. Multimedia*, vol. 23, pp. 443–453, 2021.

[15] G. Gao, J. Gao, Q. Liu, Q. Wang, and Y. Wang, "CNN-based density estimation and crowd counting: A survey," 2020, *arXiv:2003.12783*.

[16] C. Wang, H. Zhang, L. Yang, S. Liu, and X. Cao, "Deep people counting in extremely dense crowds," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1299–1302.

[17] M. Fu, P. Xu, X. Li, Q. Liu, M. Ye, and C. Zhu, "Fast crowd density estimation with convolutional neural networks," *Eng. Appl. Artif. Intell.*, vol. 43, pp. 81–88, Aug. 2015.

[18] Y. Zhang, D. Zhou, S. Chen, S. Gao, and Y. Ma, "Single-image crowd counting via multi-column convolutional neural network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 589–597.

[19] L. Boominathan, S. S. S. Kruthiventi, and R. V. Babu, "CrowdNet: A deep convolutional network for dense crowd counting," in *Proc. ACM Multimedia Conf. (MM)*, 2016, pp. 640–644.

[20] D. B. Sam, S. Surya, and R. V. Babu, "Switching convolutional neural network for crowd counting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5744–5752.

[21] Y. Li, X. Zhang, and D. Chen, "CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1091–1100.

[22] N. Liu, Y. Long, C. Zou, Q. Niu, L. Pan, and H. Wu, "ADCrowdNet: An attention-injective deformable convolutional network for crowd understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3225–3234.

[23] X. Jiang et al., "Crowd counting and density estimation by trellis encoder–decoder networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 6133–6142.

[24] Y. Xue, Y. Li, S. Liu, X. Zhang, and X. Qian, "Crowd scene analysis encounters high density and scale variation," *IEEE Trans. Image Process.*, vol. 30, pp. 2745–2757, 2021.

[25] J. Wan, Z. Liu, and A. B. Chan, "A generalized loss function for crowd counting and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1974–1983.

[26] Y.-J. Ma, H.-H. Shuai, and W.-H. Cheng, "Spatiotemporal dilated convolution with uncertain matching for video-based crowd estimation," *IEEE Trans. Multimedia*, vol. 24, pp. 261–273, 2022.

[27] E. Bondi, L. Seidenari, A. D. Bagdanov, and A. Del Bimbo, "Real-time people counting from depth imagery of crowded environments," in *Proc. 11th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Aug. 2014, pp. 337–342.

[28] C. Arteta, V. Lempitsky, and A. Zisserman, "Counting in the wild," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 483–498.

[29] D. Song, Y. Qiao, and A. Corbetta, "Depth driven people counting using deep region proposal network," in *Proc. IEEE Int. Conf. Inf. Autom. (ICIA)*, Jul. 2017, pp. 416–421.

[30] D. Lian, X. Chen, J. Li, W. Luo, and S. Gao, "Locating and counting heads in crowds with a depth prior," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9056–9072, Dec. 2022.

[31] P. Chen, J. Gao, Y. Yuan, and Q. Wang, "MAFNet: A multi-attention fusion network for RGB-T crowd counting," 2022, *arXiv:2208.06761*.

[32] T. Peng, Q. Li, and P. Zhu, "RGB-T crowd counting from drone: A benchmark and MMCCN network," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 497–513.

[33] B. Zhang, Y. Du, Y. Zhao, J. Wan, and Z. Tong, "I-MMCCN: Improved MMCCN for RGB-T crowd counting of drone images," in *Proc. 7th IEEE Int. Conf. Netw. Intell. Digit. Content (IC-NIDC)*, Nov. 2021, pp. 117–121.

[34] L. Liu, J. Chen, H. Wu, G. Li, C. Li, and L. Lin, "Cross-modal collaborative representation learning and a large-scale RGBT benchmark for crowd counting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4821–4831.

[35] H. Tang, Y. Wang, and L.-P. Chau, "TAFNet: A three-stream adaptive fusion network for RGB-T crowd counting," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, May 2022, pp. 3299–3303.

[36] Z. Wu, L. Liu, Y. Zhang, M. Mao, L. Lin, and G. Li, "Multimodal crowd counting with mutual attention transformers," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[37] H. Li, S. Zhang, and W. Kong, "Learning the cross-modal discriminative feature representation for RGB-T crowd counting," *Knowl.-Based Syst.*, vol. 257, Dec. 2022, Art. no. 109944.

[38] W. Zhou, Y. Pan, J. Lei, L. Ye, and L. Yu, "DEFNet: Dual-branch enhanced feature fusion network for RGB-T crowd counting," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 24540–24549, Dec. 2022.

[39] S. Gu and Z. Lian, "A unified RGB-T crowd counting learning framework," *Image Vis. Comput.*, vol. 131, Mar. 2023, Art. no. 104631.

[40] D. Lian, J. Li, J. Zheng, W. Luo, and S. Gao, "Density map regression guided detection network for RGB-D crowd counting and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1821–1830.

[41] S.-D. Yang, H.-T. Su, W. H. Hsu, and W.-C. Chen, "DECCNet: Depth enhanced crowd counting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 4521–4530.

[42] H. Li, S. Zhang, and W. Kong, "RGB-D crowd counting with cross-modal cycle-attention fusion and fine-coarse supervision," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 306–316, Jan. 2023.

[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[44] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[45] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[46] Z. Ma, X. Wei, X. Hong, and Y. Gong, "Bayesian loss for crowd count estimation with point supervision," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6142–6151.

[47] Y. Zhang, S. Choi, and S. Hong, "Spatio-channel attention blocks for cross-modal crowd counting," in *Proc. 16th Asian Conf. Comput. Vis.*, 2022, pp. 90–107.

[48] H. Li, J. Zhang, W. Kong, J. Shen, and Y. Shao, "CSA-Net: Cross-modal scale-aware attention-aggregated network for RGB-T crowd counting," *Expert Syst. Appl.*, vol. 213, Mar. 2023, Art. no. 119038.

[49] S. Zhang, W. Wang, W. Zhao, L. Wang, and Q. Li, "A cross-modal crowd counting method combining CNN and cross-modal transformer," *Image Vis. Comput.*, vol. 129, Jan. 2023, Art. no. 104592.

[50] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 8024–8035.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–11.

[52] M. Li, M. Soltanolkotabi, and S. Oymak, "Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2020, pp. 4313–4324.

[53] J.-F. Wang, T.-L. Zhang, and B.-J. Fu, "A measure of spatial stratified heterogeneity," *Ecol. Indicators*, vol. 67, pp. 250–256, Aug. 2016.

[54] J. Guo, J. Wang, C. Xu, and Y. Song, "Modeling of spatial stratified heterogeneity," *GISci. Remote Sens.*, vol. 59, no. 1, pp. 1660–1677, Dec. 2022.

[55] J. Wang et al., "Geographical detectors-based health risk assessment and its application in the neural tube defects study of the Heshun region, China," *Int. J. Geographical Inf. Sci.*, vol. 24, no. 1, pp. 107–127, Jan. 2010.

[56] J. Wang and X. Li. (2016). *Geodetector*. [Online]. Available: http://geodetector.cn/

[57] J. Wang et al., "Statistical modeling of spatially stratified heterogeneous data," *Ann. Amer. Assoc. Geographers*, vol. 114, no. 3, pp. 499–519, Mar. 2024.

[58] K. Sreedhar, J. Clemons, R. Venkatesan, S. W. Keckler, and M. Horowitz, "Vision transformer computation and resilience for dynamic inference," in *Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw. (ISPASS)*, May 2024, pp. 192–204.

[59] V. Sovrasov. (2018). *PTFlops: A Flops Counting Tool for Neural Networks in PyTorch Framework*. [Online]. Available: https://github.com/sovrasov/flops-counter.PyTorch

[60] J. Wang, B. Gao, and A. Stein, "The spatial statistic trinity: A generic framework for spatial sampling and inference," *Environ. Model. Softw.*, vol. 134, Dec. 2020, Art. no. 104835.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

Shuyu Wang received the B.S. and M.S. degrees from Chang'an University, Xi'an, China. He is currently pursuing the Ph.D. degree with the School of Cyber Science and Engineering, Southeast University. He is with Xizang Minzu University, China. His main research interests include crowd counting and deep learning.

Weiwei Wu (Member, IEEE) received the B.Sc. degree from South China University of Technology and the dual Ph.D. degree from the Department of Computer Science, City University of Hong Kong (CityU), and the University of Science and Technology of China (USTC), in 2011. He went to the Mathematical Division, Nanyang Technological University (NTU), Singapore, for postdoctoral research, in 2012. He is currently a Professor with the School of Computer Science and Engineering, Southeast University, China. He has published over 50 peer-reviewed papers in international conferences/journals and serves as a TPC and a reviewer for several top international journals and conferences. His research interests include optimizations and algorithm analysis, wireless communications, crowdsourcing, cloud computing, reinforcement learning, game theory, and network economics.

Yinglin Li received the B.S. degree from Jilin University, Changchun, China. She is currently pursuing the Ph.D. degree with the School of Computer Science and Engineering, Southeast University. Her main research interests include crowd movement understanding, trajectory analysis, and crowd simulation.

Yuhang Xu received the B.Sc. degree from Soochow University and the Ph.D. degree from Southeast University. He works as an Engineer at North Information Control Research Academy Group Company Ltd., Nanjing, China. His primary areas of research encompass spatial-temporal forecasting, the development of scheduling algorithms, and the application of high-precision global navigation satellite system (GNSS) technology.

Yan Lyu received the M.S. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2013, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2016. She was a Post-Doctoral Research Fellow with Hong Kong Baptist University, Hong Kong, in 2017, and the National University of Singapore, Singapore, from 2017 to 2020. She is currently an Associate Professor with the School of Computer Science and Engineering, Southeast University, China. Her research interests include data analytics and visualization, spatial-temporal data mining, and smart city.