Semantic-guided RGB-Thermal Crowd Counting with Segment Anything Model

Yaqun Fang
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
fangyq@smail.nju.edu.cn

Yi Shi
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
yishi@smail.nju.edu.cn

Jia Bei*
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
beijia@nju.edu.cn

Tongwei Ren
State Key Laboratory for Novel Software Technology, Nanjing University
Nanjing, China
rentw@nju.edu.cn

ABSTRACT

RGB-Thermal (RGB-T) crowd counting leverages the complementary nature of visible light and thermal modalities for accurate counting. However, real-world scenarios often introduce challenges, such as misidentifying background elements like trees and lampposts as individuals, leading to inaccurate counts. Existing methods utilize segmentation as an preliminary procedure, which is constrained by segmentation accuracy. In this paper, we propose a novel method, utilizing the Segment Anything Model (SAM), to distinguish between the foreground and background of images. Specifically, we begin by utilizing SAM to obtain the semantic map of the original image. Subsequently, we extract the modality features and semantic features corresponding to the RGB and thermal modalities through multimodal feature extraction. These features are then fused using the Semantic-guide Feature Fusion module. Finally, the Multi-level Decoder is employed to generate the density map and the ultimate counting results. Our approach achieves state-of-the-art performance on the RGBT-CC dataset.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision problems**; *Artificial intelligence.*

KEYWORDS

RGB-T Crowd Counting, Segment Anything Model, Transformer, Self Attention

ACM Reference Format:
Yaqun Fang, Yi Shi, Jia Bei, and Tongwei Ren. 2024. Semantic-guided RGB-Thermal Crowd Counting with Segment Anything Model. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3652583.3658108

*Corresponding author.

ICMR '24, June 10–14, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0619-6/24/06.
https://doi.org/10.1145/3652583.3658108

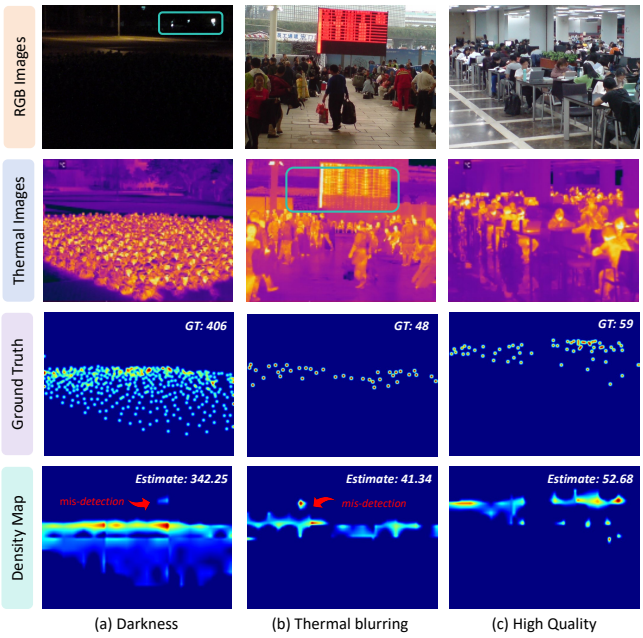

Figure 1: Examples of special cases in RGB-T crowd counting tasks. (a) Darkness. (b) Thermal blurring. (c) High-quality. Blue boxes indicate possible human-like backgrounds. Arrows represent background false detection areas in the density map.

1 INTRODUCTION

RGB-Thermal (RGB-T) crowd counting task is a branch of the crowd counting task, aiming to infer the number of people in images using RGB and thermal modal images [2, 12, 34]. In real-world scenarios, there often arises a problem of missing modality information [5, 6, 11, 29]. As shown in Figure 1 (a) and (b), under conditions such as

darkness or thermal blurring, the RGB and thermal modalities can complement each other. This mutual support ensures the accuracy of crowd counting even in challenging environments.

Current works primarily focus on facilitating interaction between these two modalities [3, 25, 28, 35], yet overlook a critical issue: the segmentation of foreground and background areas, and how to leverage this segmentation to guide feature interaction within effective regions. As depicted in Figure 1, it can be observed that in certain dark or thermally blurred conditions, background regions may exhibit human-like characteristics, which could potentially impact the counting results. To further analyze these false detections, we draw circles using the ground truth points as centers and a set pixel distance (with radii of 12, 16, 20 pixels) to differentiate between foreground and background areas. The Mean Absolute Error (MAE) is then calculated separately for these areas. It is found that in some scenarios, the background MAE constitutes a significant proportion of the total error, and both missed detections in the foreground region and false detections in the background region can occur. Therefore, distinguishing between the foreground and background for counting purposes is meaningful and contributes to the accuracy of the overall count.

Early research attempted to focus on crowd regions by overlaying foreground masks with original images, thereby filtering out background areas [7, 33]. These methods typically treat segmentation as an additional branch, utilizing segmentation labels to train a segmenter. The segmentation results are then used as a background filter for the input image, intermediate features, or the output feature map. However, these methods typically require real segmentation labels of crowd areas to train the segmenter, which consumes a significant amount of time and labor. Moreover, they are susceptible to the varying data distributions across different datasets , for example, difficulty in transferring between datasets with different crowd sizes, and are not well-suited to the complex and dynamic scenarios encountered in real-world.

With the advent of the era of large models and big data, the powerful segmentation model, Segment Anything Model (SAM) [8], has garnered significant attention from researchers. Pre-trained on a wide range of datasets, SAM demonstrated excellent zero-shot segmentation capabilities, enabling it to segment previously unseen images based on prompts including points, boxes, masks, or text. This has proven effective in numerous downstream tasks [1, 17, 20, 23]. Ma *et al.* were the pioneers in exploring the use of SAM for few-shot counting [18]. For any given input image, they use a bounding box of an object to be counted as a prompt for the SAM, obtaining a mask for the target object. This mask is then compared for similarity with the mask generated for the entire image. Objects with similarity scores exceeding a certain threshold are counted as the same object. Their experiments revealed that SAM, without further fine-tuning, significantly underperforms other few-shot counting methods, especially for small and crowded objects. In crowd counting tasks, the individuals to be counted are often small and densely packed. As is shown in Figure 2, directly employing SAM for counting individual instances can be challenging to achieve good performance. The starting point of our work is to explore how to leverage the segmentation strengths of large models like SAM to enable them to achieve outstanding performance in RGB-T crowd counting tasks.

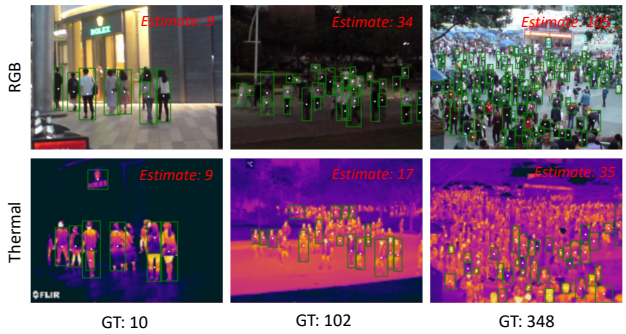

Figure 2: Examples of counting using SAM as detector directly. It is evident that as the level of congestion increases, the incidence of missed detections becomes more pronounced.

In this paper, we propose a noval method which utilize SAM to generate semantic map, and guide the interaction between modalities using semantic features. By introducing semantic regions, the issue of false detections in the background can be reduced. Additionally, the feature fusion guided by semantics enhances the counting performance in crowded region. To our knowledge, this paper is the first to apply SAM for RGB-T crowd counting. Specifically, we first apply SAM separately to the RGB and thermal modalities, using text with semantic information as prompts to obtain semantic maps for each modality. The benefit of using SAM is to avoid the need for manual annotation of segmentation labels. Additionally, due to the zero-shot learning capability of SAM, it is easily transferable to datasets with varying crowd sizes. Next, we fuse the modal feature with semantic feature through a semantic-guide feature fusion module. Finally, the fused features and lower-level image features are fed into a multi-level decoder to generate counting tokens and density maps, then generate the final counting results. Our method effectively integrates semantic and modal information to enhance the accuracy and robustness of the counting process and achieves state-of-the-art results on the RGBT-CC benchmark.

Overall, our work makes the following contributions: (1) Our research is the inaugural effort to integrate SAM into RGB-T crowd counting. Leveraging SAM, we innovatively generate semantic maps in both RGB and thermal modalities. (2) We employ semantic features to guide and enhance the representation of modal features within both RGB and thermal modalities. This approach significantly boosts the effectiveness of cross-modal feature fusion, leading to enhanced performance in crowd counting tasks.

2 RELATED WORK

2.1 RGB-T Crowd Counting

Thermal images exhibit insensitivity to variations in illumination and possess a robust capability to penetrate certain particulate matters such as dark and fog. Consequently, they can serve as supplementary information to RGB images in the context of crowd counting. Peng *et al.* introduced the DroneRGBT dataset, which

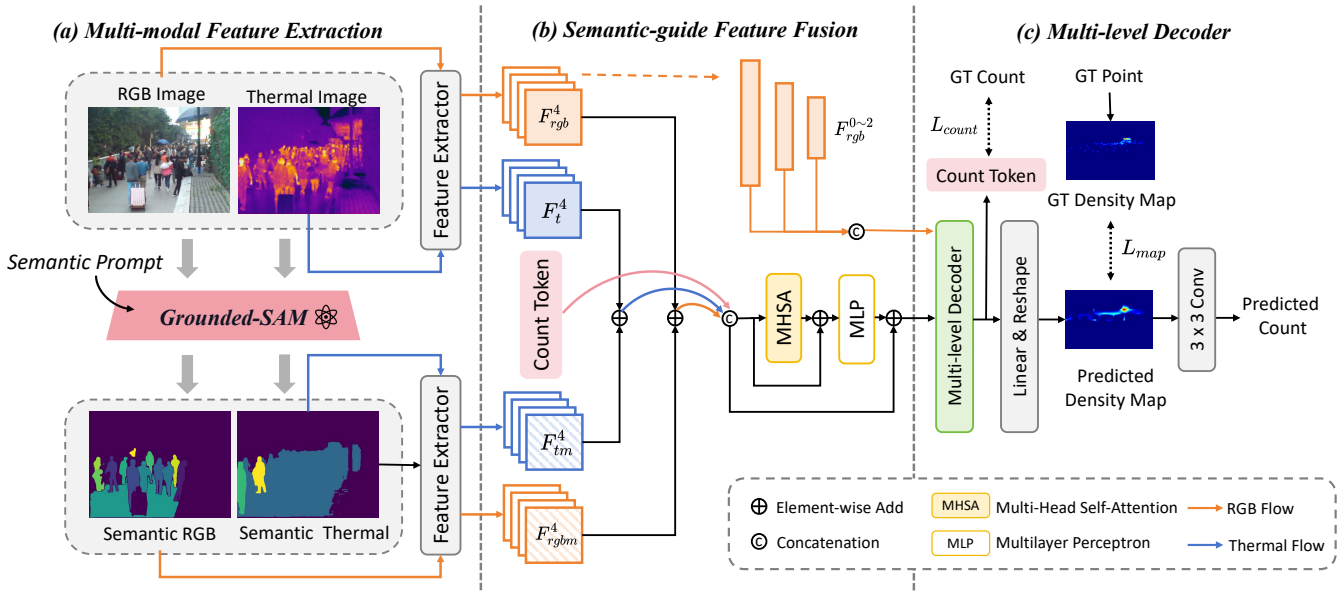

Figure 3: The framework of the proposed method.

represents the inaugural RGB-T crowd counting dataset captured from a drone's perspective [22]. Liu *et al.* proposed the RGBT-CC dataset, a comprehensive and extensive RGB-T crowd counting dataset [11]. Moreover, they present a two-stream cross-modal representation learning framework, which is a baseline approach within the RGBT-CC benchmark. Subsequent research in this field primarily branches into three directions:

Focus on interaction between RGB and thermal modalities. CSCA [32] introduces a cross-modal spatial-channel attention block, enhancing the interaction between RGB and thermal modalities. CCANet [15] also leverages cross-modal channel and spatial attention mechanisms to capture complementary features from different modalities. TAFNet [25] proposes a tri-stream network approach, considering the combination of RGB and thermal modalities as the main input stream, while aggregating information from both modalities. MAT [28] utilizes a cross-modal mutual attention mechanism to foster inter-modal information exchange. EAEFNet [10] introduces an Explicit Attention-Enhanced Fusion module, which is designed to facilitate the effective integration of information from different modalities.

Focus on aggregation of multi-scale features. CSA-Net [15], addressing for the first time the multi-scale issue in RGB-T crowd counting, proposes a scale-aware feature aggregation method. CGINet [21] introduces a hierarchical interaction method, promoting the aggregation of cross-modal features at various scales. DEFNet [35] employs a multi-level decoder to integrate features from different levels.

Focus on guidance of auxiliary information. LIURGBT [16] proposes a count-guided multi-modal fusion module, which utilizes a multi-scale token transformer to interact two modal information under the guidance of count information. Their work demonstrates that introducing auxiliary guidance information can help improve counting accuracy.

These works primarily focus on promoting the interaction between modalities and the fusion of multi-scale features. However, these previous works did not consider the interference of the background on the counting results, meaning they did not adequately differentiate between foreground and background areas, which could lead to some instances of false detection.

2.2 Segment Anything Model

SAM is a foundational model in the field of visual segmentation [8]. It has been trained on a massive dataset comprising 11 million images and 1.1 billion masks, imbuing it with extensive domain knowledge. This model primarily utilizes prompt engineering to train a large-scale pre-trained model capable of segmentation based on prompts.

It holds potential for application in downstream segmentation tasks and can be combined with other visual tasks to form new solutions for various visual challenges. Yu *et al.* designs Inpaint Anything [30], a pipeline to solve inpainting-related problems by combining the advantages of SAM. Liu *et al.* enables users to specify which style region to select and which content regions to apply during style transfer with SAM [13].

In the field of counting, Ma *et al.* is the first to explore the use of the SAM for object counting [18] and finds that it falls behind the SOTA baselines, especially for small and congested objects [31]. Thus, further improvement for SAM in some special scenes is still needed. However, the rich prior semantic knowledge contained in SAM can be utilized to provide a rough range estimation for crowd areas, thereby aiding in the counting process.

2.3 Mask-guide Crowd Counting

Previous works in crowd counting have explored the utilization of foreground segmentation maps as supplementary guidance for crowd counting, with a specific focus on significant regions of

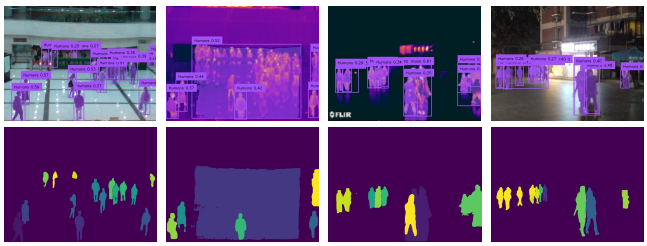

Figure 4: Visualization of semantic maps generated by Grounded-SAM. The first row is modality images, and the second row is semantic maps generated by Grounded-SAM.

interest. Zhao *et al.* employed segmentation task to assist in filtering out background interference for crowd counting [33]. Modolo *et al.* proposed utilizing a mask at the output layer to suppress prediction results in the background region, effectively improving detection accuracy [19]. Jiang *et al.* proposed to use a dedicated network branch to predict the object/non-object mask and then combine its prediction with the input image to produce the density map [7].

Most of these approaches require generating ground truth for segmentation as labels to guide the model's segmentation process or utilize existing segmentation models to suppress the background in the original image. In our work, we primarily leverage the rich semantic knowledge embedded in large-scale visual models to segment the crowd region using textual cues, eliminating the need for explicit labels.

3 METHODOLOGY

3.1 Overview

The framework of the proposed method is depicted in Figure 3. The overall framework primarily comprises three parts: Multi-modal Feature Extraction, Semantic-guide Feature Fusion, and Multi-level Decoder. Initially, given pairs of original RGB and thermal images, Grounded-SAM [24] is used to generate corresponding modal semantic segmentation maps with semantic inputs as prompts. These are then input into feature extractors to separately extract four-stage semantic and modal features for both RGB and thermal modalities. In the Semantic-guide Feature Fusion module, the highest-level semantic features, modal features, and count tokens are fused. The fused highest-level features, along with low-level features from the RGB modality, are input into the Multi-level Decoder, which outputs the count token and density map. Finally, a regressor is used to obtain the final predicted crowd count.

3.2 Multi-modal Feature Extraction

For a given pair of RGB and thermal images, we initially utilize the Grounded-SAM tool[1] to generate crowd region segmentation maps containing semantic information, based on semantic prompts. This tool is developed based on Grounding DINO [14] and Segment Anything [8]. As illustrated in Figure 4, Grounded-DINO initially generates object detection boxes based on semantic prompts. Subsequently, SAM produces areas of semantic segmentation within these boxes.

[1]https://github.com/IDEA-Research/Grounded-Segment-Anything

To obtain representations of information at different hierarchical levels, we employ Pyramid Vision Transformer [27] as the image encoder. It is used to extract features from four stages of the RGB and Thermal images: $F_{rgb} = \left\{F^i_{rgb} | i = 0, 1, 2, 3\right\}$, $F_t = \left\{F^i_t | i = 0, 1, 2, 3\right\}$, as well as their corresponding semantic maps: $M_{rgb} = \left\{M^i_{rgb} | i = 0, 1, 2, 3\right\}$ and $M_t = \left\{M^i_t | i = 0, 1, 2, 3\right\}$.

The extracted feature maps decrease in size but increase in the number of channels. Their sizes are 56, 28, 14, and 7, respectively, while the number of channels are 64, 128, 320, and 512, respectively. Different Multi-Layer Perceptron layers are used to uniformly transform the number of channels of these feature maps to 512. Through the Multi-modal Feature Extraction process, we acquire various levels of modal and semantic features from both the RGB and thermal modalities.

3.3 Semantic-guide Feature Fusion

To integrate semantic segmentation information into cross-modal features, we propose a Semantic-guided feature fusion module. In this module, semantic features from two modalities are individually fused with the original features. Subsequently, the features from both modalities are concatenated with a count token. By computing self-attention, a unified feature representation is obtained.

Building upon previous research, it is established that higher-level features often contain more semantic information and thus are more beneficial for global counting. Accordingly, we merge the highest-level features extracted from two modalities with the highest-level features extracted from the semantic modality:

$$G_{rgb} = F^3_{rgb} \oplus M^3_{rgb}, \tag{1}$$

$$G_t = F^3_t \oplus M^3_t, \tag{2}$$

where G_{rgb} and G_t represent the features resulting from the fusion of the original and semantic features in the RGB and thermal modalities, respectively; $\oplus$ represents element-wise addition; F^3_{rgb} and F^3_t denote the highest-level original features of the two modalities. Meanwhile, M^3_{rgb} and M^3_t correspond to the highest-level semantic features of the RGB and thermal modalities, respectively.

Subsequently, we concatenate the RGB modality's semantically fused features, the thermal modality's semantically fused features, and a learnable count token along the channel dimension to form an initial fused feature:

$$\mathcal{F} = Concat[G_{rgb}, G_t, token_{count}], \tag{3}$$

where $Concat[\cdot]$ means the concatenation of features along the channel dimension; G_{rgb} and G_t represent the fusion features of the original and semantic features in the RGB and thermal modalities, respectively; $token_{count}$ is a tensor initialized with zero values, having a size of 1×1 and a channel count identical to the preceding two components, which represents the initialized count token.

To further integrate the cross-modal features, we input this initial fused feature into a multi-head self-attention module.

Specifically, the process begins with a linear layer that generates queries, keys, and values for the input fused features. These are then reshaped and split into n heads. For each head, attention weights

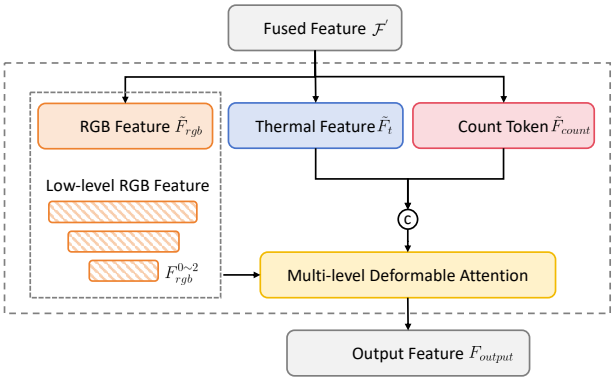

Figure 5: Detailed process of Multi-level Decoder.

are computed by:

$$A_{head}^{i} = Softmax\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) \cdot V_i, \tag{4}$$

where Q_i represents queries, K_i represents keys, V_i represents values, i represents the ith head, and A_{head}^{i} denotes the attention of the ith head. Then, the attention computed by each head is concatenated and reshaped, yielding the final attention representation:

$$\mathcal{A} = Concat[A_{head}^{1}, A_{head}^{2}, \ldots, A_{head}^{n}] \cdot \omega_0, \tag{5}$$

where $\mathcal{A}$ represents the final attention. Here, n is the total number of heads, $Concat$ refers to the concatenation operation, A_{head}^{i} denotes the attention of each head, and ω_0 is the final linear transformation matrix.

Following this, the output from the multi-head self-attention is added back to the original input features through a residual connection. This is then passed through an MLP layer, and once again added back to the original input features through another residual connection, resulting in the final fused feature:

$$\mathcal{F}' = \mathcal{F} \oplus MLP(\mathcal{F} \oplus \mathcal{A}), \tag{6}$$

where $\mathcal{F}'$ is the final fused feature, $\mathcal{F}$ is the initial fused feature, $\mathcal{A}$ represents the output of the multi-head self-attention, and $\oplus$ indicates the addition operation. The final feature effectively integrates semantic information, RGB modality information, thermal modality information, and counting information.

3.4 Multi-level Decoder

Based on previous research, it is established that the low-level features of images often encompass a wealth of information pertaining to colors, textures, and contours, among other shape-related aspects. In contrast, higher-level features are typically more imbued with semantic information. To facilitate the integration of features across different levels, this study adopts a decoder based on multi-scale deformable attention mechanism, as proposed by Liu *et al.* [16], to achieve the fusion of information across multiple scales.

Specifically, as detailed in Section 3.3, $\mathcal{F}'$ is segmented into $\tilde{F}_{rgb}$, $\tilde{F}_t$, and $\tilde{F}_{count}$, each corresponding to the size of the original input features. As shown in Figure 5, then the process commences with the concatenation of $\tilde{F}_t'$ and $\tilde{F}_{count}$. This is followed by the aggregation of $\tilde{F}_{rgb}$ with the features extracted from the initial three layers. The culmination of this procedure involves the computation of their multi-level deformable attention, which is designated as the output. The specific computational formula employed in this methodology is as follows:

$$F_{output} = MSDAtt\left(Concat[\tilde{F}_t, \tilde{F}_{count}], \{\tilde{F}_{rgb}, F_{rgb}^{0}, F_{rgb}^{1}, F_{rgb}^{2}\}\right), \tag{7}$$

where F_{output} refers to the output of the decoder; $MSDAtt$ denotes the Multi-scales Deformable Attention Mechanism [36]; $Concat$ is used to describe the concatenation of feature maps; $\tilde{F}_t$, $\tilde{F}_{rgb}$, and $\tilde{F}_{count}$ are features derived from the segmented fusion feature $\mathcal{F}'$; F_{rgb}^{0} to F_{rgb}^{2} represent the features from the first three layers of the original RGB input; F_{output}, is divided into two distinct components: F_{map} and F_c. Here, F_{map} represents the features of the density map, while F_c corresponds to the counting features. F_c undergoes a linear transformation to generate the count token. F_{map}, after being processed through a 3×3 convolutional layer to revert to its original size, yields the final count result.

3.5 Loss Function

The overall loss is composed of two parts: the loss of the density map and the loss of counting. The density map loss originates from DM-Count [26], which uses Optimal Transport to measure the similarity between the normalized predicted density map and the normalized ground truth density map. The counting loss employs the L1 norm to supervise the predicted count against the actual count.

$$\mathcal{L}_{total} = \mathcal{L}_{map}\left(D, \hat{D}\right) + \mathcal{L}_{count}\left(C, \hat{C}\right), \tag{8}$$

where D represents the predicted density map, $\hat{D}$ represents the actual density map, C denotes the predicted count, and $\hat{C}$ signifies the actual count.

4 EXPERIMENT

4.1 Dataset and Metrics

Our method is evaluated on the publicly available RGBT-CC dataset [11] . RGBT-CC is a recently introduced benchmark, stands as a large-scale, free-view, multimodal crowd counting dataset. It comprises 2,030 pairs of RGB-T images. The dataset is partitioned into 1,030 pairs for training, 200 pairs for validation, and 800 pairs for testing purposes. This dataset presents a formidable challenge, featuring images captured under diverse illumination conditions across a variety of settings, including malls, streets, playgrounds, and stations. On average, each image in this dataset is marked with point annotations for approximately 68 pedestrians, illustrating its complexity and the richness of data it offers for crowd counting analyses.

We use the widely used Grid Average Mean Absolute Error (GAME) [4] and Root Mean Square Error (RMSE) [11] to evaluate our method. GAME at level l is computed as:

$$GAME(l) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{4^l}|\hat{P}_i^j - P_i^j|, \tag{9}$$

Table 1: Comparison results of the proposed method against the competing trackers on RGBT-CC test set. The best results are highlighted in Bold and the second results are indicated with an underline.

Methods	Publisher	Year	$GAME(0) \downarrow$	$GAME(1) \downarrow$	$GAME(2) \downarrow$	$GAME(3) \downarrow$	$RMSE \downarrow$
CMCRL [11]	CVPR	2021	15.61	19.95	24.69	32.89	28.18
MAT [28]	ICME	2022	12.35	16.29	20.81	29.09	22.53
LIURGBT [16]	BMVC	2022	<u>10.90</u>	<u>14.81</u>	<u>19.02</u>	**26.14**	<u>18.79</u>
DEFNet [35]	TITS	2022	11.90	16.08	20.19	27.27	21.09
CSCA [32]	ACCV	2022	14.32	18.91	23.81	32.47	26.01
TAFNet [25]	ISCAS	2022	12.38	16.98	21.86	30.19	22.45
CCANet [15]	TMM	2023	13.93	18.13	22.08	28.26	24.71
CSANet [9]	ESA	2023	12.45	16.46	21.48	30.62	21.64
CGINet [21]	EAAI	2023	12.07	15.98	20.06	27.73	20.54
EAEFNet [10]	RAL	2023	11.19	14.99	19.20	27.13	19.39
Ours			**10.51**	**14.52**	**18.92**	<u>26.28</u>	**17.71**

where $\hat{P}_i^j$ represents the predicted value of the j^{th} region of the i^{th} image, P_i^j indicates the ground truth corresponding to $\hat{P}_i^j$, 4^l means the number of the divided non-overlapping regions of the image, and N is the number of paired images in testing dataset. GAME sums the counting errors in all the regions. In particular, GAME(0) is equivalent to MAE. RMSE is computed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_i - P_i)^2}, \tag{10}$$

where $\hat{P}_i$ represents the predicted value of the i^{th} image, P_i indicates the ground truth corresponding to $\hat{P}_i$.

4.2 Implementation Details

Data preprocessing. The original RGB and thermal images in the dataset possess a resolution of 640×480 pixels. To facilitate processing, these images have undergone a reshaping operation, altering their dimensions to 672×448 pixels. This modification is to enable the subdivision of the images into patches, each measuring 224×224 pixels.

Acquisition of semantic information. In order to acquire masks imbued with semantic information via SAM, our approach utilizes a tool constructed on the foundation of SAM and Grounding DINO for the extraction of crowd region masks. More specifically, we employ the pretrained models *groundingdino_swint_ogc.pth* and *sam_vit_h_4b8939.pth*. The text prompt "*Humans*" is used in this process. The parameters set for the extraction include a box threshold of 0.25, a text threshold of 0.25, and a Non-Maximum Suppression (NMS) threshold of 0.8. These thresholds are integral in delineating the boundaries of human presence in the images, thereby allowing for the precise extraction of crowd region masks.

Training details. Our networks are trained using the Adam optimizer with a batch size of 16, an initial learning rate of 2e-5, with 30 epochs for warm up. The model is trained for 500 epochs with a learning rate decay of 1e-4. Both the training and testing of our model are conducted using PyTorch1.11.0 on an NVIDIA RTX 3090 GPU with 24GB memory.

4.3 Comparison with State-of-the-Arts

We compared ten methods on the RGBT-CC benchmark, which are as follows: CMCRL [11], CSCA [32], CCANet [15], CSANet [9], TAFNet [25], MAT [28], CGINet [21], DEFNet [35], EAEFNet [10] and LIURGBT [16].

Diverging from other comparative methods, our method is specifically designed to incorporate semantic information to assist in the fusion of cross-modal features. Semantic information plays a crucial role in distinguishing between foreground areas with crowds and background regions. The differentiation is instrumental in enabling the model to focus more accurately on counting within crowd areas, while effectively suppressing the interference caused by background information.

Quantitative evaluation. Our method, along with various others, was evaluated on the RGBT-CC benchmark, as demonstrated in Table 1. Our approach achieved the best performance in terms of four key metrics: GAME(0), GAME(1), GAME(2), and RMSE. In GAME(3), it ranked second. Evaluating the overall counting performance, the MAE, represented by GAME(0), showed a reduction of 0.39, and the RMSE decreased by 1.08. These results substantiate the effectiveness of incorporating semantic information in enhancing the accuracy of crowd counting.

The significant improvements highlight the advantage of integrating semantic context into the analysis. By doing so, the model more effectively differentiates between crowd and non-crowd areas, leading to a more focused and precise counting. This advancement demonstrates the potential of semantic information in refining the accuracy of crowd counting methodologies.

Qualitative evaluation. The comparative visualization of our method against others is illustrated in Figure 6. Our method was compared with CMCLR, CSCA, and LIURGBT methods across a variety of scenarios. It is evident that our approach yields more accurate counting results in typical environments such as dark, crowded, and thermally blurred scenes. Additionally, an analysis of the density map distributions reveals that our method concentrates the predictions more effectively in crowd areas. By integrating semantic information, our approach enhances the accuracy of the

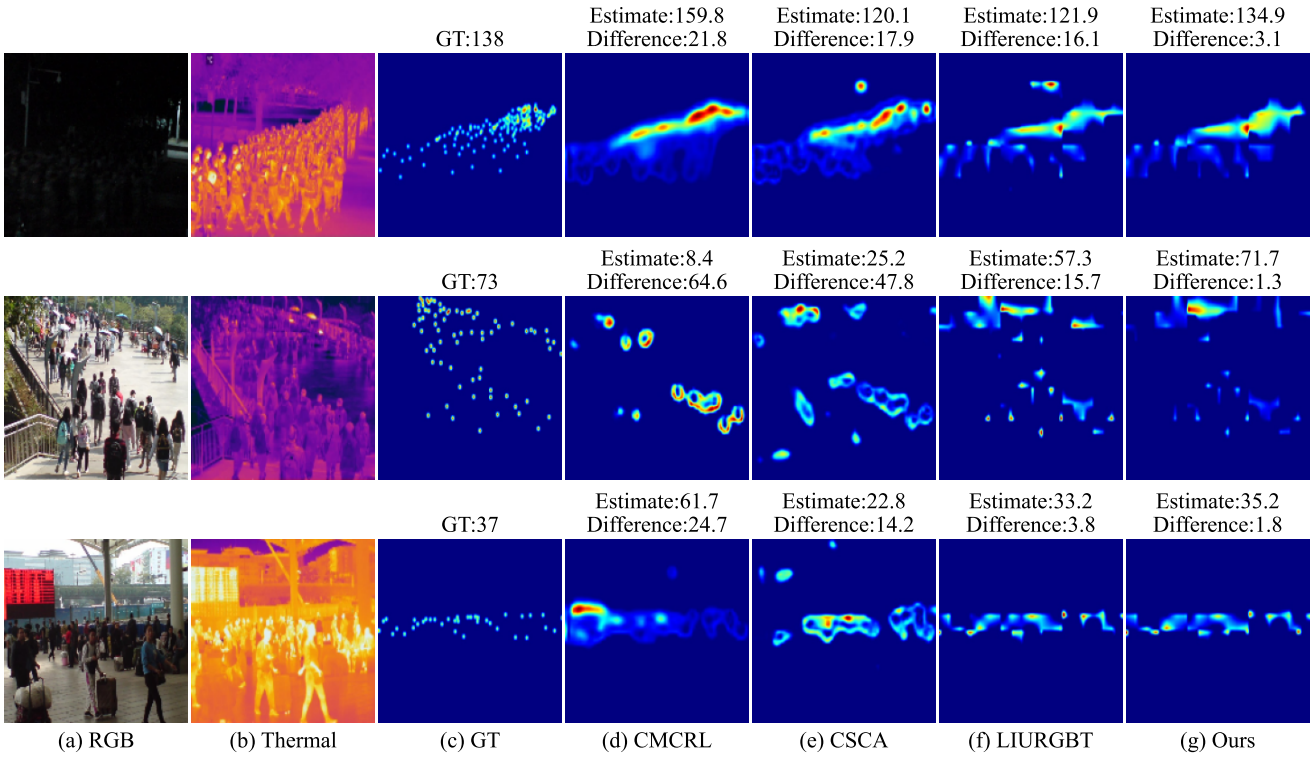

Figure 6: Visual examples in three different scenarios: dark, crowded, and thermally blurred environments. "Estimate" refers to the predicted number of people, while "Difference" denotes the deviation from the actual count.

Table 2: Ablation experiments concerning the integration of semantic maps on RGBT-CC test set. $\mathcal{S}_{rgb}$ indicates whether the semantic maps of the RGB modality is integrated, and $\mathcal{S}_t$ denotes the integration of the semantic maps for the thermal modality. The best results are highlighted in Bold.

$\mathcal{S}_{rgb}$	$\mathcal{S}_t$	$GAME(0)\downarrow$	$GAME(1)\downarrow$	$GAME(2)\downarrow$	$GAME(3)\downarrow$	$RMSE\downarrow$
✗	✗	11.44	15.43	19.67	26.70	20.44
✓	✗	10.85	15.17	19.56	26.78	19.08
✗	✓	10.84	15.14	19.52	26.59	18.53
✓	✓	**10.51**	**14.52**	**18.92**	**26.28**	**17.71**

Table 3: Ablation experiments concerning the fusion strategy of semantic maps on RGBT-CC test set. The best results are highlighted in Bold.

Strategy	$GAME(0)\downarrow$	$GAME(1)\downarrow$	$GAME(2)\downarrow$	$GAME(3)\downarrow$	$RMSE\downarrow$
Multiply	10.92	14.96	19.51	26.62	19.70
Concat	10.95	15.30	19.68	27.10	18.49
Avg	11.48	15.55	19.96	27.94	19.30
Avg+Concat	11.00	14.76	19.02	**26.24**	19.33
Ours	**10.51**	**14.52**	**18.92**	26.28	**17.71**

predictive distribution, ensuring that the focus is maintained on regions with higher concentrations of people.

However, as depicted in Figure 7, our method encounters challenges in scenarios characterized by excessive crowding and an overwhelming number of individuals, resulting in suboptimal performance. In some scenarios, our method is still constrained by the quality of the original image. When Grounded-SAM fails to accurately differentiate the regions containing humans, the counting performance remains unsatisfactory. These issues warrant more in-depth research in the future.

4.4 Ablation Studies

To ascertain the effectiveness of semantic information and the efficacy of the Semantic-guided Feature Fusion module, we conducted a series of ablation experiments. These experiments were specifically designed to address two key aspects: the integration of semantic information and the effective fusion of this semantic information into the model.

Whether to add semantic information. We conducted ablation experiments to assess the impact of integrating semantic information on detection results. These experiments included various configurations: not integrating any semantic maps, integrating the semantic maps only for the RGB modality, integrating the semantic maps only for the thermal modality, and integrating semantic maps for both modalities. The results, as shown in Table 2, indicate that the addition of semantic information contributes to an increase in detection accuracy. It is evident that the semantic information from both the RGB and thermal modalities provides beneficial enhancements to the detection process.

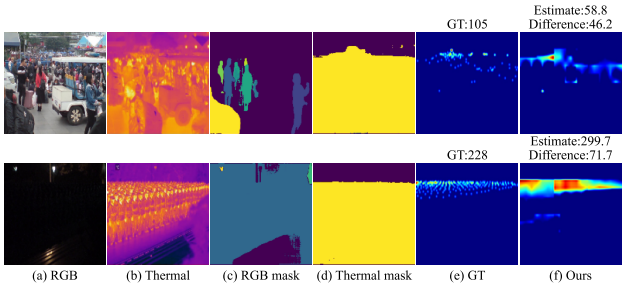

Figure 7: Visualizations of some failure cases.

How to effectively integrate semantic information. We carried out ablation experiments to evaluate the effects of different fusion strategies on detection results. These experiments tested various approaches, including the pointwise multiplication of original modality information with semantic modality information, concatenation, averaging, concatenation after averaging, and our proposed fusion module. Specifically:

- Pointwise multiplication involves element-wise multiplication of modality feature maps with semantic feature maps.
- Concatenation refers to joining the modality feature maps and semantic feature maps along the channel dimension, followed by a 1×1 convolution to restore the original shape.
- Averaging entails using the average of semantic features from both modalities added to the original modality features for fusion.
- Concatenation after averaging involves joining the average semantic features from both modalities with the original modality features.

The experimental results, as shown in Table 3, demonstrate that our proposed Semantic-guided Feature Fusion Module outperforms other fusion methods. This finding highlights the superiority of our approach in effectively combining semantic and modality-specific features, thereby enhancing the accuracy and efficacy of the detection process.

5 CONCLUSION

In this paper, we proposed a novel semantic-guided RGB-T crowd counting method, which generates semantic maps of crowd on both RGB and thermal modalities by leveraging SAM. Our method explored the utilization of semantic features to guide and enhance the representation of modal features through the semantic-guided fusion module. With semantic information, the false-positive counting in background is reduced, while the counting accuracy in crowd regions is improved. The experiments on the RGBT-CC dataset demonstrate that our proposed method outperforms the state-of-the-art methods.

ACKNOWLEDGMENTS

This work is supported by the National Science Foundation of China (62072232), Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (021714380026), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

[1] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. 2023. Tracking anything with decoupled video segmentation. In *IEEE/CVF International Conference on Computer Vision*.
[2] Feng Dai, Hao Liu, Yike Ma, Xi Zhang, and Qiang Zhao. 2021. Dense scale network for crowd counting. In *International Conference on Multimedia Retrieval*.
[3] Zhipeng Du, Miaojing Shi, Jiankang Deng, and Stefanos Zafeiriou. 2023. Redesigning multi-scale neural network for crowd counting. *IEEE Transactions on Image Processing* 32 (2023), 3664–3678.
[4] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. 2015. Extremely overlapping vehicle counting. In *Pattern Recognition and Image Analysis*.
[5] Ruichao Hou, Tongwei Ren, and Gangshan Wu. 2022. Mirnet: A robust rgbt tracking jointly with multi-modal interaction and refinement. In *IEEE International Conference on Multimedia and Expo*.
[6] Ruichao Hou, Boyue Xu, Tongwei Ren, and Gangshan Wu. 2023. MTNet: learning modality-aware representation with transformer for RGBT tracking. In *IEEE International Conference on Multimedia and Expo*.
[7] Shengqin Jiang, Xiaobo Lu, Yinjie Lei, and Lingqiao Liu. 2019. Mask-aware networks for crowd counting. *IEEE Transactions on Circuits and Systems for Video Technology* (2019), 3119–3129.
[8] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).
[9] He Li, Junge Zhang, Weihang Kong, Jienan Shen, and Yuguang Shao. 2023. CSA-Net: Cross-modal scale-aware attention-aggregated network for RGB-T crowd counting. *Expert Systems with Applications* 213 (2023), 119038.
[10] Mingjian Liang, Junjie Hu, Chenyu Bao, Hua Feng, Fuqin Deng, and Tin Lun Lam. 2023. Explicit attention-enhanced fusion for RGB-Thermal perception tasks. *IEEE Robotics and Automation Letters* 8, 7 (2023), 4060–4067.
[11] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, and Liang Lin. 2021. Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*.
[12] Lei Liu, Jie Jiang, Wenjing Jia, Saeed Amirgholipour, Michelle Zeibots, and Xiangjian He. 2019. DENet: a universal network for counting crowd with varying densities and scales. *arXiv preprint arXiv:1904.08056* (2019).
[13] Songhua Liu, Jingwen Ye, and Xinchao Wang. 2023. Any-to-any style transfer: making Picasso and Da Vinci collaborate. *arXiv preprint arXiv:2304.09728* (2023).
[14] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).
[15] Yanbo Liu, Guo Cao, Boshan Shi, and Yingxiang Hu. 2023. CCANet: A collaborative cross-modal attention network for RGB-D crowd counting. *IEEE Transactions on Multimedia* 26 (2023), 154–165.
[16] Zhengyi Liu, Wei Wu, Yacheng Tan, and Guanghui Zhang. 2022. RGB-T multi-modal crowd counting based on transformer. In *British Machine Vision Conference*.
[17] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. 2024. Segment anything in medical images. *Nature Communications* 15, 1 (2024), 654.
[18] Zhiheng Ma, Xiaopeng Hong, and Qinnan Shangguan. 2023. Can sam count anything? an empirical study on sam counting. *arXiv preprint arXiv:2304.10817* (2023).
[19] Davide Modolo, Bing Shuai, Rahul Rama Varior, and Joseph Tighe. 2021. Understanding the impact of mistakes on background regions in crowd counting. In *IEEE/CVF Winter Conference on Applications of Computer Vision*.
[20] Lucas Prado Osco, Qiusheng Wu, Eduardo Lopes de Lemos, Wesley Nunes Gonçalves, Ana Paula Marques Ramos, Jonathan Li, and José Marcato Junior. 2023. The segment anything model (sam) for remote sensing applications: From zero to one shot. *International Journal of Applied Earth Observation and Geoinformation* 124 (2023), 103540.
[21] Yi Pan, Wujie Zhou, Xiaohong Qian, Shanshan Mao, Rongwang Yang, and Lu Yu. 2023. CGINet: Cross-modality grade interaction network for RGB-T crowd counting. *Engineering Applications of Artificial Intelligence*, 106885.
[22] Tao Peng, Qing Li, and Pengfei Zhu. 2020. Rgb-t crowd counting from drone: A benchmark and mmccn network. In *Asian conference on computer vision*.
[23] Simiao Ren, Francesco Luzi, Saad Lahrichi, Kaleb Kassaw, Leslie M. Collins, Kyle Bradbury, and Jordan M. Malof. 2024. Segment anything, From Space?. In *IEEE/CVF Winter Conference on Applications of Computer Vision*.
[24] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. *arXiv preprint arXiv:2401.14159* (2024).
[25] Haihan Tang, Yi Wang, and Lap-Pui Chau. 2022. TAFNet: A three-stream adaptive fusion network for RGB-T crowd counting. In *IEEE International Symposium on Circuits and Systems*.

[26] Boyu Wang, Huidong Liu, Dimitris Samaras, and Minh Hoai Nguyen. 2020. Distribution matching for crowd counting. *Advances in neural information processing systems* 33 (2020), 1595–1607.

[27] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In *IEEE/CVF International Conference on Computer Vision*.

[28] Zhengtao Wu, Lingbo Liu, Yang Zhang, Mingzhi Mao, Liang Lin, and Guanbin Li. 2022. Multimodal crowd counting with mutual attention transformers. In *IEEE International Conference on Multimedia and Expo*.

[29] Boyue Xu, Ruichao Hou, Jia Bei, Tongwei Ren, and Gangshan Wu. 2024. Jointly modeling association and motion cues for robust infrared UAV tracking. *The Visual Computer* (2024), 1432–2315.

[30] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. 2023. Inpaint anything: segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790* (2023).

[31] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu. 2023. A comprehensive survey on segment anything model for vision and beyond. *arXiv preprint arXiv:2305.08196* (2023).

[32] Youjia Zhang, Soyun Choi, and Sungeun Hong. 2022. Spatio-channel attention blocks for cross-modal crowd counting. In *Asian Conference on Computer Vision*.

[33] Muming Zhao, Jian Zhang, Chongyang Zhang, and Wenjun Zhang. 2019. Leveraging heterogeneous auxiliary tasks to assist crowd counting. In *IEEE/CVF conference on Computer Vision and Pattern Recognition*.

[34] Liangfeng Zheng, Yongzhi Li, and Yadong Mu. 2021. Learning factorized cross-view fusion for multi-view crowd counting. In *IEEE International Conference on Multimedia and Expo*.

[35] Wujie Zhou, Yi Pan, Jingsheng Lei, Lv Ye, and Lu Yu. 2022. DEFNet: Dual-branch enhanced feature fusion network for RGB-T crowd counting. *IEEE Transactions on Intelligent Transportation Systems* 23, 12 (2022), 24540–24549.

[36] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2020. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159* (2020).