Pattern Recognition Letters
journal homepage: www.elsevier.com

Thermal Crowd Counting by Distilling Multi-modal Knowledge

Xiaoxu Liu, Yi Shi, Ruichao Hou**, Tongwei Ren

State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, 210008, China

Article history:

Keywords:
crowd counting
thermal modality
multi-modal
knowledge distillation

ABSTRACT

Crowd counting aims to estimate the number of pedestrians across various scenes. Compared to RGB images, thermal images demonstrate greater robustness in complex scenarios such as low illumination and adverse weather conditions. However, this task remains underexplored due to challenges like the misidentification of heat-emitting objects. To address this issue, we propose Multi-modal Knowledge Distillation (MKD), a novel thermal crowd counting method that transfers multi-modal knowledge from an RGB-thermal (RGB-T) teacher model into a thermal-based student model. Specifically, we combine feature distillation and response distillation to transfer low-level semantic knowledge and fine-grained fused knowledge simultaneously. Furthermore, to address cases where the teacher model underperforms the student model, we design an adaptive weighting module that dynamically adjusts instance-level loss weights during training, enabling the student to prioritize more valuable samples. Extensive experiments on RGBT-CC and DroneRGBT validate the effectiveness of MKD. In particular, MKD reduces the G(0) on RGBT-CC from 12.27 to 11.35. Meanwhile, compared with the RGB-T teacher, our student achieves about a 30% reduction in both parameters and FLOPs, while maintaining state-of-the-art accuracy among thermal-based methods.

1. Introduction

Crowd counting aims to estimate the number of pedestrians in unconstrained scenes[1]. Existing studies can be broadly grouped into RGB-based crowd counting [2, 3], thermal-based crowd counting [4, 5], and RGB-thermal (RGB-T) fusion crowd counting [6, 7].

RGB-based methods benefit from rich appearance cues but often degrade in low illumination, severe occlusion, and adverse weather. In contrast, thermal images are robust to illumination changes and can better highlight pedestrians in complex scenarios. Prior works [6, 7] also demonstrate that thermal images may provide stronger support for density map estimation than RGB images in such cases. Moreover, thermal imaging captures heat patterns rather than identifiable visual textures, which is advantageous for privacy preservation [4].

Recent RGB-T fusion models exploit modality complementarity and typically achieve stronger performance [8, 9, 10, 11, 12, 13, 14, 15]. In a broader context, multi-modal foundation models have also advanced rapidly, demonstrating effective cross-modal alignment across diverse modalities [16]. However, fusion models may suffer severe degradation when one modality becomes unreliable[17], under extremely poor illumination, noisy RGB cues can even harm fusion, leading to worse results than a thermal-based crowd counting model, as illustrated in Fig. 1(a).

Despite its practical value, thermal-only crowd counting remains underexplored, mainly due to several modality-specific challenges. First, public thermal crowd datasets are relatively limited in scale compared with RGB benchmarks, which restricts supervised learning and generalization. Second, thermal images are prone to false positives from heat-emitting or reflective distractors (*e.g.*, lamps, heated objects), especially

**Corresponding author.
e-mail: lxx@smail.nju.edu.cn (Xiaoxu Liu), yishi@smail.nju.edu.cn (Yi Shi), rchou@nju.edu.cn (Ruichao Hou), rentw@nju.edu.cn (Tongwei Ren)

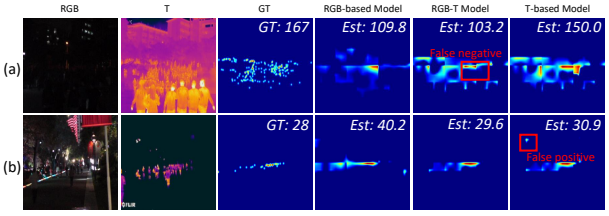

Fig. 1: Examples of crowd counting. (a) The RGB-T model detects fewer pedestrians under extremely poor illumination conditions. (b) The thermal-based crowd counting model misidentifies lamps as pedestrians.

without auxiliary RGB cues, as shown in Fig. 1(b). Third, thermal images often contain weaker fine-grained textures and lower spatial details than RGB, making dense localization and separation of close-by pedestrians more difficult. Finally, thermal-only learning usually lacks strong multi-modal supervision signals that could provide complementary semantics.

These observations motivate a practical yet underexplored direction: leveraging RGB-T knowledge during training to improve thermal-only inference, *i.e.*, transferring complementary information from RGB-T models while requiring only thermal inputs at test time. To this end, we propose a thermal-based crowd counting network via Multi-modal Knowledge Distillation (MKD). Motivated by cross-modal distillation [18, 19, 20, 21], we distill multi-modal knowledge from an RGB-T teacher to a thermal student with two objectives: (1) feature distillation to transfer branch-level representations of RGB and T streams, and (2) response distillation to transfer fused knowledge after multi-modal interaction. Moreover, we introduce an adaptive weighting module to adjust distillation strength per instance, emphasizing reliable teacher guidance.

We evaluate MKD on RGBT-CC [8] and DroneRGBT [22]. Results show that MKD significantly boosts thermal-based crowd counting models, achieving state-of-the-art performance among thermal-only methods while reducing FLOPs and parameters by about 30% compared with the RGB-T teacher.

The main contributions can be summarized as follows:

- We propose a novel thermal crowd counting method that enhances the performance by transferring multi-modal knowledge, without incurring additional inference costs.
- We design two distillation methods at the feature and response levels to effectively transfer both low-level semantic knowledge and fine-grained fused knowledge.
- We propose an adaptive weighting module that prioritizes high-quality knowledge and suppresses the influence of noisy samples.

2. Related Works

2.1. Crowd Counting

Traditional crowd counting methods [23, 1, 2] relying solely on RGB modality often struggle under challenging weather and lighting conditions. The thermal image has been widely adopted as a complementary modality to enhance robustness. Recent works have explored RGB-T crowd counting, leveraging the strengths of RGB and thermal modalities. For example, Peng *et al.* [22] propose a drone-based RGB-T crowd counting dataset, DroneRGBT, and MMCCN with multi-scale feature learning and modality alignment modules. Liu *et al.* [8] proposed the RGBT-CC dataset and IADM to collaboratively learn cross-modal representation with a dual information propagation mechanism. Zhang *et al.* [9] developed CSCA, employing spatial-wise cross-modal attention to capture correlations between multi-modal features. Wu *et al.* [10] proposed MAT, which utilizes a cross-modal mutual attention mechanism to leverage the complementary information of two modalities. Liu *et al.* [7] introduced a multi-scale token transformer to enhance the thermal feature using the RGB modality. Fang *et al.* [14] utilized SAM to guide the interaction between modalities with segmentation maps.

Most of these methods leverage two modalities as inputs to adapt to different scenarios. However, the RGB-T crowd counting approaches suffer severe performance degradation when the RGB information is unreliable or ineffective. In this paper, we propose a novel distillation framework to build a thermal-based crowd counting model by transferring multi-modal knowledge.

2.2. Multi-modal Knowledge Distillation

Knowledge distillation (KD) [24] was originally introduced for model compression by transferring knowledge from a high-capacity teacher to a lightweight student, and has since been widely used in multi-modal tasks, including RGB-T salient object detection [18], RGB-T image fusion [25] and RGB-T segmentation [19]. Recently, several studies have explored KD for RGB-T crowd counting [26, 27, 28, 29, 30]. Zhou *et al.* [26] distill a strong teacher into MJPNet-S*, while VPMFNet [27] transfers knowledge from an RGB model to strengthen the vision branch of a multi-branch fusion network. Meng *et al.* [28] introduce an auxiliary broker modality and pretrain a BMG network by distilling a diffusion-based teacher. Mu *et al.* [29] propose cooperative mutual distillation for a lightweight RGB-T model, and Yang *et al.* [30] develop a relation-aware lightweight network trained with cyclic self-contrastive distillation. However, these methods still require both RGB and thermal inputs at inference.

In contrast, our work transfers multi-modal knowledge to a uni-modal student to enable thermal-only inference, targeting scenarios where RGB images are unavailable or unreliable and improving robustness in challenging conditions.

3. Method

Thermal images often produce false positives without RGB information, while RGB-T approaches mitigate this issue by leveraging the complementary strengths of RGB and thermal data. To improve the performance of thermal-based crowd counting models, we distill knowledge from a pretrained RGB-T teacher model by transferring both low-level features before fusion and responses after multi-modal fusion. The framework of the proposed method is depicted in Fig. 2. During the training phase, paired RGB and thermal images are fed into the teacher network, while the student model only

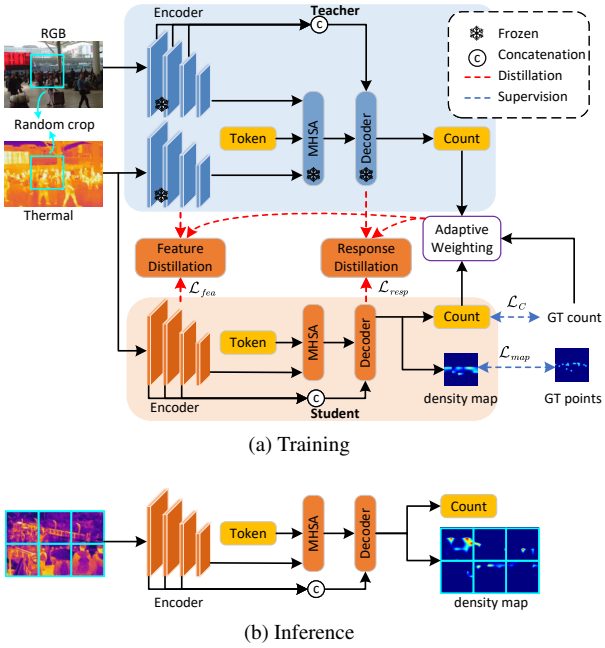

Fig. 2: The framework of our MKD method. (a) During training, feature distillation and response distillation are performed to transfer the multi-modal knowledge to the student. An adaptive weighting module adjusts the distillation weight at the instance level. (b) During inference, the model only takes thermal images as inputs.

takes the corresponding thermal image as input. The teacher model is kept frozen, and the student model is trained to learn from the teacher through two proposed distillation losses. Additionally, an adaptive weighting module dynamically adjusts instance-level distillation weights, allowing the student to focus on more valuable knowledge. Following [7], random cropping is applied to the input images during training. During inference, predictions are made on each cropped region, and the results are aggregated to produce the final output.

3.1. Network Architecture

We adopt MSDTrans [7] as the teacher model $\mathcal{T}$. Given paired RGB-T inputs (I_r, I_t), multi-scale thermal features $\{F^{\mathcal{T}}_{t,i}\}^4_{i=1}$ and RGB features $\{F^{\mathcal{T}}_{r,i}\}^4_{i=1}$ are obtained via two PVT encoders [31], where i denotes the feature layer index. A learnable count token $F^{\mathcal{T}}_c$, together with the high-level features $F^{\mathcal{T}}_{t,4}$ and $F^{\mathcal{T}}_{r,4}$, is passed through a multi-head self-attention (MHSA) module to derive enhanced features $\tilde{F}^{\mathcal{T}}_c$, $\tilde{F}^{\mathcal{T}}_t$ and $\tilde{F}^{\mathcal{T}}_r$. Subsequently, a multi-scale deformable transformer decoder [32] facilitates feature fusion across scales. Specifically, the enhanced thermal feature $\tilde{F}^{\mathcal{T}}_t$ and the token $\tilde{F}^{\mathcal{T}}_c$ are concatenated as query. The enhanced RGB feature $\tilde{F}^{\mathcal{T}}_r$ along with the low-layer RGB features $F^{\mathcal{T}}_{r,i}(i = 1, 2, 3)$ serve as key and value. Finally, a prediction head generates outputs, which consist of a linear layer that generates the final count prediction $C^{\mathcal{T}}$ and a convolutional layer that generates the final density map $D^{\mathcal{T}}$.

The student $\mathcal{S}$ follows the MSDTrans design as the thermal-only branch, reusing the same core components: a PVT encoder to extract multi-scale thermal features $\{F^{\mathcal{S}}_{t,i}\}^4_{i=1}$, the learnable count token and MHSA module to produce $\tilde{F}^{\mathcal{S}}_c$ and $\tilde{F}^{\mathcal{S}}_t$, and the multi-scale deformable decoder and prediction head to generate $C^{\mathcal{S}}$ and $D^{\mathcal{S}}$. The key modification is that $\mathcal{S}$ removes the RGB branch entirely and uses only thermal representations to drive the decoder. Specifically, the query is constructed as $[\tilde{F}^{\mathcal{S}}_t; \tilde{F}^{\mathcal{S}}_c]$, while the key and value are formed by the enhanced thermal feature $\tilde{F}^{\mathcal{S}}_t$ and low-level thermal features $\{F^{\mathcal{S}}_{t,i}\}^3_{i=1}$. Importantly, all distillation-related modules are introduced only during training and are discarded at inference, so the student inference path remains a lightweight thermal-only MSDTrans-style network.

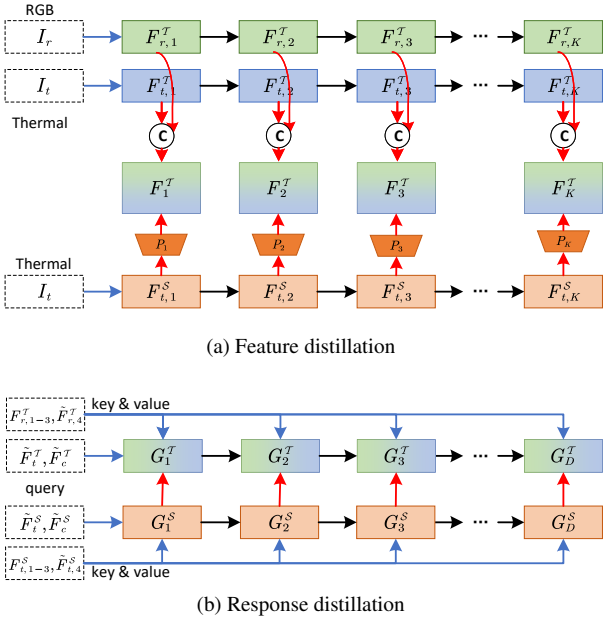

Fig. 3: Detailed design of the proposed two distillation methods.

3.2. Multi-stage Knowledge Distillation

Compared to thermal images, the RGB-T models achieve better performance by combining complementary information from the RGB and thermal images. Therefore, we utilize the knowledge from the pretrained RGB-T model to guide the thermal-based model through multi-stage knowledge distillation. Specifically, we transfer low-level image features before fusion and the response after fusion.

Low-level image features serve as critical inputs to the decoder, acting as key and value that carry semantic information essential for subsequent feature fusion. Therefore, we apply knowledge distillation to these low-level features. As illustrated in Fig. 3(a), the teacher's features from the thermal and RGB branches are concatenated along the channel dimension, which can be formulated as $F^{\mathcal{T}}_i = \text{cat}\left(\left[F^{\mathcal{T}}_{t,i}, F^{\mathcal{T}}_{r,i}\right]\right)$. To align the dimensions of the student and teacher features, following FitNet[33], we utilize a projector $P_i(\cdot)$ to each feature of the student $F^{\mathcal{S}}_{t,i}$, which is composed of a convolutional layer, a batch normalization layer and a ReLU activation layer. The feature distillation loss is defined by the L_2-distance between the features of the student and teacher as follows:

$$\mathcal{L}_{fea} = \sum_{i=1}^{K} L_2(P_i(F^{\mathcal{S}}_{t,i}), F^{\mathcal{T}}_i), \tag{1}$$

where K represents the number of encoder layers, and K is set to 4 in our experiments.

In addition to feature-level transfer, we distill the responses from the teacher model, which encapsulate fused multi-modal knowledge. As observed in recent works [34, 35], the decoder contains different knowledge at different stages. Thus, instead of relying solely on the final output, we utilize the outputs of multiple decoder layers to supervise the student. This process is illustrated in Fig. 3(b). Each decoder layer incrementally refines the density map and count token predictions based on the outputs of the previous layer. The response distillation loss is formulated as:

$$\mathcal{L}_{resp} = \sum_{i=1}^{D} L_2(G_i^{\mathcal{T}}, G_i^{\mathcal{S}}), \tag{2}$$

where D denotes the number of decoder layers, D is set to 6 in our experiments; $G_i^{\mathcal{T}}$ and $G_i^{\mathcal{S}}$ represent the outputs of the i-th decoder layer of the teacher and student model, respectively.

3.3. Adaptive Weighting Module

In the RGB-T crowd counting task, the confidence of the two modalities fluctuates depending on specific scenarios. For example, under completely dark conditions, RGB images may degrade performance, potentially causing the teacher model to underperform relative to the student model. Traditional distillation methods treat all samples equally, which can lead to suboptimal performance in such cases.

Inspired by the success of some instance-level techniques [18, 40, 41], we propose an adaptive weighting module tailored for RGB-T crowd counting. This module adjusts the distillation weight at the instance level, allowing the model to prioritize samples that convey more valuable knowledge from the teacher model. The weight w_j is formulated as follows:

$$w_j = \begin{cases} w_0, & \text{if } C \\ 1, & \text{otherwise} \end{cases}, \tag{3}$$

where $w_0 < 1$ is a constant value, C represents the condition which is set according to different strategies. In this paper, three strategies: 'global brightness', 'local brightness', and 'adaptive weight' are explored. Ultimately, we adopt the 'adaptive weight' method due to its superior performance.

Global Brightness. Illumination is a crucial factor that affects the performance of models. Therefore, we assign weights based on whether the scenario is bright or dark.

Local Brightness. During training, random cropping is applied to the input images. Even in a dark scenario, some local areas can be bright. To assess the brightness of the cropped image, we begin by converting the RGB image into grayscale. Then the darkness p_d is calculated based on the percentage of the pixels that fall below a specified threshold v. If p_d exceeds a predetermined threshold ξ, we classify the image as dark. Otherwise, it is considered bright.

Adaptive Weight. During early training epochs, the teacher model performs better than the student model, which is beneficial to the student model's learning. However, in later training epochs, there are instances where the performance of the teacher declines, thereby impeding the student's learning.

3.4. Loss Function

Finally, the distillation loss can be formulated as:

$$\mathcal{L}_{distill} = \sum_{j=1}^{N} w_j \left(\lambda_1 \mathcal{L}_{fea,j} + \lambda_2 \mathcal{L}_{resp,j}\right), \tag{4}$$

where $\mathcal{L}_{fea,j}$ and $\mathcal{L}_{resp,j}$ stands for the feature distillation loss and the response distillation loss of the j-th instance, $\lambda_1 = 0.5$ and $\lambda_2 = 1$ are balancing weights of the proposed two distillation loss terms, and N is the size of the mini-batch.

We leverage the task loss used in previous methods [7, 14], which is composed of density map loss $\mathcal{L}_{map}$ originating from DM-Count [1] and a count loss $\mathcal{L}_C$ computed with the L_1 norm. It can be formulated as:

$$\mathcal{L}_{task} = \mathcal{L}_{map}(D^{\mathcal{S}}, \hat{D}) + \mathcal{L}_C(C^{\mathcal{S}}, \hat{C}), \tag{5}$$

where $D^{\mathcal{S}}$ and $C^{\mathcal{S}}$ represent the predicted density map, and count of the student model, $\hat{D}$ and $\hat{C}$ denote the ground truth density map and count.

The final loss is calculated by linearly combining the distillation loss and the task loss:

$$\mathcal{L} = \mathcal{L}_{distill} + \mathcal{L}_{task}. \tag{6}$$

4. Experiment

4.1. Experimental Settings

Dataset. We evaluate our method on two popular RGB-T crowd counting benchmarks, RGBT-CC [8] and DroneRGBT [22], which are complementary in scenario coverage and challenges. RGBT-CC provides aligned RGB-thermal image pairs collected from ground-view scenes with diverse environments and noticeable illumination variation. It contains 1,030/200/800 pairs for training/validation/testing, respectively, and includes both bright and dark conditions. This illumination diversity makes RGBT-CC well suited for analyzing model robustness under low-light scenarios, where RGB cues may become unreliable. DroneRGBT focuses on UAV-based data acquisition and the associated domain characteristics. It consists of 3,807 aligned RGB-T pairs captured by drone-mounted cameras, with 1,800 pairs for training and the remaining images for testing. Following the common protocol, we further split the training set into 70%/30% for training/validation. Compared with ground-view datasets, DroneRGBT typically exhibits larger viewpoint changes, more pronounced scale variation, and diverse crowd density patterns, which are valuable for evaluating the generalization of our distillation strategy across viewpoints and scenes.

Evaluation metrics. Following previous methods, we adopt GAME[42] and RMSE as the evaluation metrics. GAME at level l is calculated as:

$$\text{GAME}(l) = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{4^l} \left| \hat{P}_i^j - P_i^j \right|, \tag{7}$$

Table 1: Comparison with state-of-the-art methods using the thermal modality on RGBT-CC and DroneRGBT. Metrics with ↓ mean lower is better.

Type	Method	Venue	RGBT-CC					DroneRGBT				
			G(0)↓	G(1)↓	G(2)↓	G(3)↓	RMSE ↓	G(0)↓	G(1)↓	G(2)↓	G(3)↓	RMSE ↓
RGB-T	IADM [8]	CVPR2021	15.61	19.95	24.69	32.89	28.18	9.77	12.91	17.08	22.61	15.76
	CSCA [9]	ACCV2022	14.32	18.91	23.81	32.47	26.01	9.51	12.12	15.84	21.57	15.19
	MAT [10]	ICME2022	12.35	16.29	20.81	29.09	22.53	–	–	–	–	–
	DEFNet [11]	TITS2022	11.90	16.08	20.19	27.27	21.09	8.69	10.67	13.44	17.48	14.19
	MSDTrans[7]	BMVC2022	10.90	14.81	19.02	26.14	18.79	6.42	8.47	10.99	14.70	10.14
	BMG [28]	ECCV2024	10.19	13.61	17.65	23.64	17.32	6.20	–	–	–	10.40
	MJPNet-T [26]	IOT2024	11.56	16.36	20.95	28.91	17.83	6.02	7.72	10.32	14.39	9.62
	BGDFNet [36]	TIM2024	11.00	15.04	19.86	29.72	19.05	–	–	–	–	–
	MMFFNet [37]	TIM2025	9.87	13.59	17.31	22.91	16.85	5.36	–	–	–	8.33
T	SANet [38]	ECCV2018	22.23	26.16	31.05	37.61	38.94	12.53	16.26	20.74	26.85	17.53
	CSRNet [39]	CVPR2018	21.64	26.22	31.65	38.66	37.38	11.11	14.24	18.13	22.94	16.40
	BL [23]	ICCV2019	17.80	22.88	28.50	37.30	30.24	8.64	11.14	14.60	19.84	14.41
	DM-Count [1]	NeurIPS2020	13.69	17.08	21.03	27.18	24.27	6.83	9.06	11.69	15.02	10.83
	GL [2]	CVPR2021	14.77	18.29	22.29	29.28	27.15	7.54	9.82	12.80	16.62	11.70
	IADM [8]	CVPR2021	16.20	20.71	25.23	33.30	28.30	11.49	14.40	18.40	23.74	17.35
	CSCA [9]	ACCV2022	14.75	19.55	24.74	33.53	24.24	11.56	14.15	17.74	22.71	18.14
	MSDTrans [7]	BMVC2022	14.09	19.11	23.38	30.21	24.09	9.82	11.94	14.11	17.25	14.85
	MC^{3}Net [13]	TITS2023	14.89	18.36	22.21	30.25	25.97	–	–	–	–	–
	VPMFNet [27]	IOT2024	14.88	18.67	23.99	33.80	25.50	–	–	–	–	–
	MKD (Ours)	–	**11.35**	**15.35**	**19.66**	**26.42**	**19.77**	**6.13**	**8.05**	**10.48**	**14.38**	**9.41**

where N is the total number of the testing samples, P_i^j and $\hat{P}_i^j$ are the estimated count and the corresponding ground truth count in the j^{th} region of the i-th image. In the tables of this paper, we use G(l) for short.

Implementation details. All experiments are conducted on a single NVIDIA RTX 3090 GPU with 24 GB memory, using an Intel Xeon E5-2680 CPU and the PyTorch framework. Following the previous method[7], we reshape the image into a shape of $(674, 448)$. During training, images are randomly cropped into a shape of $(224, 224)$. Our networks are trained for 500 epochs with the initial 30 epochs for warm-up. The mini-batch size is set at 16. We use the Adam optimizer with an initial learning rate of 2e-5 and a weight decay of 1e-4. During inference, we first resize each input image to a unified resolution of 674×448, and then partition the preprocessed full image into multiple crops with the same size as in training (224×224). Each crop is fed into the network independently to produce a local density map. We then paste all predicted local density maps back to the resized image coordinate system and stitch them into a full-image density map with overlap handling, and obtain the final crowd estimate by summing over the aggregated full-image density map. In our implementation, training typically takes about 12 hours on a single GPU.

4.2. Comparison With State-of-the-art Methods

We compare our method with state-of-the-art crowd counting methods on the RGBTCC and DroneRGBT datasets. Due to the scarcity of thermal modality approaches, we establish this baseline through the following strategies: (1) *Retrained RGB-based*: We adapt several advanced RGB crowd counting models to the thermal modality, including SANet [38], CSRNet [39], BL [23], DM-Count [1], and GL [2]. (2) *T-based*: We also select some results on the thermal modality from the ablation studies of some RGB-T crowd counting works, including MC^{3}Net [13] and VPMFNet [27]. (3) *Adapted RGB-T based*: Additionally, we generate RGB images from thermal images using PearlGAN [43] and evaluate them with RGB-T crowd counting models, namely IADM [8], CSCA [9], and MSDTrans [7]. A comprehensive comparison of state-of-the-art methods is presented in Table 1. Our model achieves state-of-the-art performance on the thermal modality. In addition, it outperforms many RGB-T models without relying on the RGB modality during inference, thereby demonstrating the effectiveness of our distillation method in effectively transferring multi-modal knowledge.

Table 2: The performance under different illumination conditions on RGBT-CC. Results displayed on a gray background indicate the performance of the model without distillation.

Scenario	Model	G(0)↓	G(1)↓	G(2)↓	G(3)↓	RMSE ↓
Dark	Teacher	10.41	14.40	17.99	24.33	17.94
	Student	11.81	15.67	19.48	25.40	23.75
	MKD	10.51	14.33	18.14	23.86	19.58
Bright	Teacher	11.38	15.20	20.02	27.89	19.58
	Student	12.72	16.84	21.81	29.60	22.52
	MKD	12.16	16.34	21.14	28.92	19.96
All	Teacher	10.9	14.81	19.02	26.14	18.79
	Student	12.27	16.26	20.66	27.53	23.13
	MKD	11.35	15.35	19.66	26.42	19.77

To further validate the effectiveness of the proposed method, we infer the models under different illumination conditions. The results are summarized in Table 2, where *teacher* is the RGB-T model [7], and *student* is the thermal-based crowd counting model proposed in this paper. Benefiting from the distilled knowledge, our method significantly improves the student model, which even surpasses the teacher model in some metrics (*e.g.*, GAME(1) and GAME(3)) in the dark

Table 3: Ablation study on the effect of different modalities on RGBT-CC.

Teacher	Student	G(0)↓	G(1)↓	G(2)↓	G(3)↓	RMSE↓
✗	T	12.27	16.26	20.66	27.53	23.13
RGB	T	11.85	16.37	20.90	27.84	**21.31**
RGB-T	T	**11.55**	**15.86**	**20.21**	**27.00**	21.56
✗	RGB	18.62	28.08	35.29	45.61	37.78
T	RGB	18.32	27.56	34.54	44.79	**35.47**
RGB-T	RGB	**18.01**	**25.97**	**32.91**	**43.55**	37.57

scenario. Moreover, we visualize some density maps generated by different models to analyze the sources of performance gains. The thermal-based crowd counting model is susceptible to interference from heated objects. The first example in Fig. 4 shows that our model successfully suppresses interference with knowledge distilled from the teacher. The following two examples demonstrate that our method exhibits greater robustness across various lighting conditions.

4.3. Ablation Studies

Comparison of different teacher modalities. To analyze the effect of different teacher models, we modify the modality of the teacher. When distilling from an RGB-based teacher, we implement the method in a similar way. The distinction lies in the feature distillation phase, where the multi-scale features in the backbone are directly transferred to the student model through projectors. In this experiment, we disable the adaptive weighting module. Results in Table 3 show that utilizing the RGB-T teacher model yields superior performance than the RGB-based teacher across most metrics, indicating that transferring RGB-T knowledge encourages the student to capture multi-modal complementary information. Moreover, we conduct an extensive experiment by distilling the knowledge into an RGB-based model. The results prove that distillation also works on the RGB-based model. However, the performance of the RGB-based model lags behind that of thermal-based models due to challenges associated with poor illumination conditions.

Effect of distillation losses. To investigate the impact of the distillation methods, we conduct an ablation study on RGBT-CC in Table 4. In this experiment, we disable the adaptive weighting module to isolate the effect of the distillation losses. We first train the thermal student without distillation, and then apply feature distillation and response distillation independently. Compared with the non-distilled student, both variants bring measurable improvements, indicating that the teacher provides useful transferable supervision. We further combine the two losses in our multi-stage distillation, which achieves the best overall performance, validating the effectiveness of our design.

We also include a classical KD baseline that distills only the teacher's final output. Its performance can be worse than training without distillation, likely due to negative transfer in cross-modal distillation: the RGB-T teacher's predictions may depend on RGB cues that are unavailable to the thermal-only student at inference, making the supervision partially inconsistent and harder to imitate. In contrast, our MKD provides more transferable guidance by distilling intermediate representations together with responses, and the full model further benefits from adaptive weighting.

Table 4: Ablation study on adaptive distillation on RGBT-CC.

Setting	G(0)↓	G(1)↓	G(2)↓	G(3)↓	RMSE↓
classical KD [24]	18.30	35.13	22.56	28.40	36.80
w/o $\mathcal{L}_{resp}$	11.98	15.95	20.35	27.01	22.31
w/o $\mathcal{L}_{fea}$	11.61	15.87	20.41	27.35	21.29
MKD (Ours)	**11.35**	**15.35**	**19.66**	**26.42**	**19.77**

Table 5: Ablation study on the effect of adaptive distillation on RGBT-CC.

Strategy	G(0)↓	G(1)↓	G(2)↓	G(3)↓	RMSE↓
w/o distillation	12.27	16.26	20.66	27.53	23.13
Same weight	11.55	15.86	20.21	27.00	21.56
Global brightness	12.11	16.00	20.13	27.17	22.17
Local brightness	11.72	15.70	19.96	26.69	21.82
Adaptive weight	**11.35**	**15.35**	**19.66**	**26.42**	**19.77**

Effect of the adaptive weighting. We compare the proposed method (*adaptive weight*) with two additional strategies, alongside a baseline distillation using uniform weights (*same weight*). The first strategy assigns weights based on the brightness labels provided in the dataset, referred to as *global brightness*. The second strategy allocates weights according to the brightness of cropped images, termed *local brightness*. Specifically, we first convert the RGB image to grayscale and calculate the darkness level p_d based on the percentage of pixels that fall below a specified threshold $v(v = 40)$. If p_d exceeds a certain threshold $\xi(\xi = 0.75)$, the image is classified as dark; otherwise, it is considered bright. For a fair comparison, we set the weight w_0 to 0.1. The results are recorded in Table 5. Although all strategies outperform the model without distillation, both *global brightness* and *local brightness* perform worse than *same weight* across some metrics. This suggests that brightness may not be an optimal criterion for filtering samples. In contrast, *adaptive weight* adjusts the weight based on the count loss, which reduces the weights only when the performance of the teacher falls below that of the student, thereby achieving superior performance.

Effect of teacher models. We further study whether MKD depends on a specific RGB-T teacher. We distill from different RGB-T teachers under the same setting and report results in Table 6. A stronger RGB-T teacher may rely more heavily on RGB cues, making its predictions less transferable to a thermal-only student that cannot access those cues at inference. In addition, effective distillation requires stable, well-calibrated outputs and easily alignable intermediate features. Differences in calibration and fusion representations can reduce the learnability of the teacher's supervision despite higher RGB-T accuracy.

4.4. Complexity Analysis

We compare the computational complexity of the proposed thermal student with its RGB-T teacher and representative baselines. As reported in Table 7, the teacher (MSDTrans) contains 200.87M parameters and requires 24.58G FLOPs, whereas our method reduces the model size to 141.20M parameters and the computation to 17.22G FLOPs. This yields

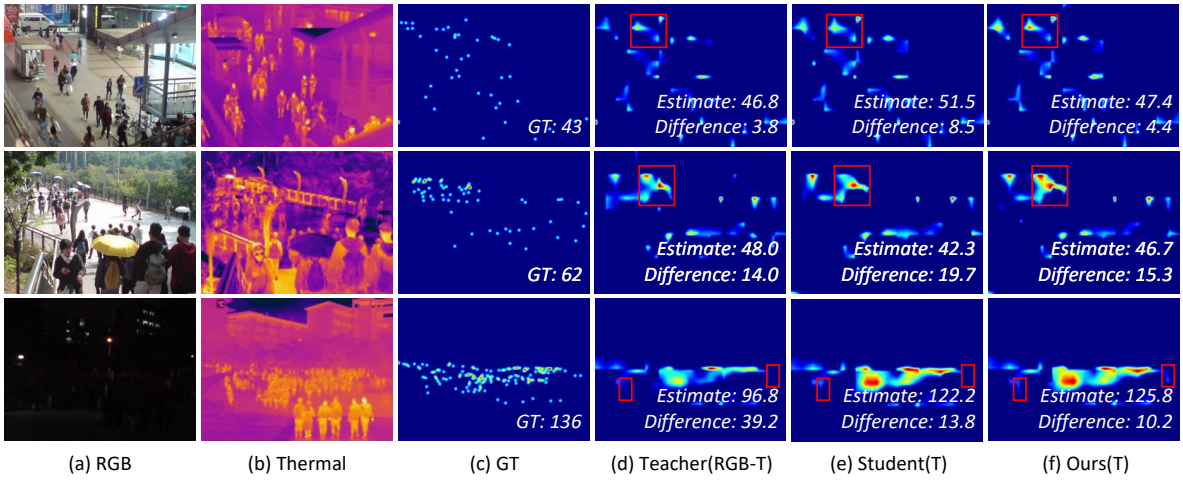

Fig. 4: Qualitative results on the test set of RGBT-CC.

Table 6: Comparison with different teacher models via MKD on RGBT-CC.

Method	G(0)↓	G(1)↓	G(2)↓	G(3)↓	RMSE↓
MKD (MMFFNet)	12.61	15.47	**18.74**	**23.94**	23.45
MKD (Ours)	**11.35**	**15.35**	19.66	26.42	**19.77**

Table 7: Model complexity comparison in terms of parameters and FLOPs.

Type	Method	Params (M)	FLOPs (G)
RGB-T	IADM [8]	25.67	22.20
RGB-T	MSDTrans [7]	200.87	24.58
T	CSRNet [39]	23.90	41.52
T	MC3Net [13]	113.08	31.94
T	MKD (Ours)	141.20	17.22

a reduction of 29.7% in parameters and 29.9% in FLOPs, indicating a substantially more efficient model for thermal-only inference.

5. Limitations and Future Work

While the MKD framework achieves promising results for thermal crowd counting, it is subject to several limitations that warrant further investigation. First, MKD relies on paired RGB-T data during training to enable multi-modal knowledge transfer, which may restrict its applicability in scenarios where only thermal data are available or RGB-T pairing is difficult to collect. Second, since the student is supervised by an RGB-T teacher, it may inherit the teacher's biases or failure modes, although our adaptive weighting alleviates this issue to some extent by down-weighting less reliable samples. Third, scalability to low-resolution or mobile thermal sensors is not fully explored; reduced spatial details and domain gaps may lead to degraded density estimation in real deployments. In future work, we will investigate distillation with unpaired or weakly paired cross-sensor data, bias-aware and uncertainty-aware distillation to reduce negative transfer, and resolution-robust training or domain adaptation techniques to better support low-cost mobile thermal devices.

6. Conclusion

In this work, we proposed a novel knowledge distillation method for thermal crowd counting. Specifically, we integrated feature distillation and response distillation to make full use of RGB-T knowledge. Moreover, we introduced an adaptive weighting module that dynamically adjusts the weight of the distillation loss at the instance level. Extensive experiments on RGBT-CC and DroneRGBT demonstrate consistent improvements over thermal-only baselines and achieve state-of-the-art performance among thermal-based methods, while maintaining high efficiency compared with the RGB-T teacher.

CRediT authorship contribution statement

Xiaoxu Liu: Conceptualization, Methodology, Software, Writing – original draft. Yi Shi: Investigation, Methodology, Validation. Ruichao Hou: Project administration, Writing – original draft. Tongwei Ren: Polishing, Funding acquisition, Resource.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62072232), the Key R&D Project of Jiangsu Province (BE2022138), the Fundamental Research Funds for the Central Universities (021714380026), the Innovation Project of State Key Laboratory for Novel Software Technology, Nanjing University (ZZKT2024B20), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

[1] B. Wang, H. Liu, D. Samaras, M. H. Nguyen, Distribution matching for crowd counting, Advances in neural information processing systems 33 (2020) 1595–1607.

[2] J. Wan, Z. Liu, A. B. Chan, A generalized loss function for crowd counting and localization, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 1974–1983.

[3] V. A. Sindagi, V. M. Patel, A survey of recent advances in cnn-based single image crowd counting and density estimation, Pattern Recognition Letters 107 (2018) 3–16.

[4] R. Tse, T. Wang, M. Im, G. Pau, Privacy aware crowd-counting using thermal cameras, in: Twelfth International Conference on Digital Image Processing, Vol. 11519, SPIE, 2020, pp. 323–333.

[5] A. Hassan, A. EL-SAYED, M. Moawad, An improved technique for crowd counting based on thermal bands, Menoufia Journal of Electronic Engineering Research 31 (1) (2022) 29–34.

[6] H. Tang, Y. Wang, L.-P. Chau, Tafnet: A three-stream adaptive fusion network for rgb-t crowd counting, in: 2022 IEEE international symposium on circuits and systems, IEEE, 2022, pp. 3299–3303.

[7] Z. Liu, W. Wu, Y. Tan, G. Zhang, Rgb-t multi-modal crowd counting based on transformer, British Machine Vision Conference (2022).

[8] L. Liu, J. Chen, H. Wu, G. Li, C. Li, L. Lin, Cross-modal collaborative representation learning and a large-scale rgbt benchmark for crowd counting, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 4823–4833.

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

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

[11] W. Zhou, Y. Pan, J. Lei, L. Ye, L. Yu, Defnet: Dual-branch enhanced feature fusion network for rgb-t crowd counting, IEEE Transactions on Intelligent Transportation Systems 23 (12) (2022) 24540–24549.

[12] Y. Pan, W. Zhou, X. Qian, S. Mao, R. Yang, L. Yu, Cginet: Cross-modality grade interaction network for rgb-t crowd counting, Engineering Applications of Artificial Intelligence 126 (2023) 106885.

[13] W. Zhou, X. Yang, J. Lei, W. Yan, L. Yu, Mc 3 net: Multimodality cross-guided compensation coordination network for rgb-t crowd counting, IEEE Transactions on Intelligent Transportation Systems 25 (5) (2023) 4156–4165.

[14] Y. Fang, Y. Shi, J. Bei, T. Ren, Semantic-guided rgb-thermal crowd counting with segment anything model, in: Proceedings of the 2024 International Conference on Multimedia Retrieval, 2024, pp. 570–578.

[15] M. Xu, Z. Ge, X. Jiang, G. Cui, P. Lv, B. Zhou, C. Xu, Depth information guided crowd counting for complex crowd scenes, Pattern Recognition Letters 125 (2019) 563–569.

[16] R. Shao, C. Yang, Q. Li, L. Xu, X. Yang, X. Li, M. Li, Q. Zhu, Y. Zhang, Y. Li, et al., Allspark: A multimodal spatio-temporal general intelligence model with ten modalities via language as a reference framework, IEEE Transactions on Geoscience and Remote Sensing (2025).

[17] S. Wei, C. Luo, Y. Luo, Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20039–20049.

[18] H. Zhou, B. Qiao, L. Yang, J. Lai, X. Xie, Texture-guided saliency distilling for unsupervised salient object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7257–7267.

[19] S. Gupta, J. Hoffman, J. Malik, Cross modal distillation for supervision transfer, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2827–2836.

[20] G. Radevski, D. Grujicic, M. Blaschko, M.-F. Moens, T. Tuytelaars, Multimodal distillation for egocentric action recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 5213–5224.

[21] H. Zhao, Q. Zhang, S. Zhao, Z. Chen, J. Zhang, D. Tao, Simdistill: Simulated multi-modal distillation for bev 3d object detection, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 38, 2024, pp. 7460–7468.

[22] T. Peng, Q. Li, P. Zhu, Rgb-t crowd counting from drone: A benchmark and mmccn network, in: Proceedings of the Asian conference on computer vision, 2020.

[23] Z. Ma, X. Wei, X. Hong, Y. Gong, Bayesian loss for crowd count estimation with point supervision, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 6142–6151.

[24] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv preprint arXiv:1503.02531 (2015).

[25] Z. Ding, R. Hou, Y. Men, S. Luan, Y. Liu, K. He, S. Xie, Dskfuse: Passive-active distillation learning for multi-modal image fusion via dynamic sparse kansformer, Expert Systems with Applications (2025) 130610.

[26] W. Zhou, X. Yang, X. Dong, M. Fang, W. Yan, T. Luo, Mjpnet-s*: Multistyle joint-perception network with knowledge distillation for drone rgb-thermal crowd density estimation in smart cities, IEEE Internet of Things Journal 11 (11) (2024) 20327–20339.

[27] B. Mu, F. Shao, Z. Xie, H. Chen, Q. Jiang, Y.-S. Ho, Visual prompt multibranch fusion network for rgb-thermal crowd counting, IEEE Internet of Things Journal 11 (19) (2024) 31758–31775.

[28] H. Meng, X. Hong, C. Wang, M. Shang, W. Zuo, Multi-modal crowd counting via a broker modality, in: European Conference on Computer Vision, Springer, 2024, pp. 231–250.

[29] B. Mu, F. Shao, H. Chen, X. Wang, Q. Jiang, A mutual head knowledge distillation framework for lightweight rgb-t crowd counting, IEEE Transactions on Circuits and Systems for Video Technology (2025).

[30] X. Yang, Y. Sun, Lranet: Lightweight relation-aware network based on self-comparative distillation for uav rgb-thermal crowd counting in smart cities, IEEE Internet of Things Journal (2025).

[31] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pvt v2: Improved baselines with pyramid vision transformer, Computational visual media 8 (3) (2022) 415–424.

[32] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, in: International Conference on Learning Representations, 2021.

[33] A. Romero, Fitnets: Hints for thin deep nets, in: in International Conference on Learning Representations, 2015.

[34] F. Chen, H. Zhang, K. Hu, Y.-K. Huang, C. Zhu, M. Savvides, Enhanced training of query-based object detection via selective query recollection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 23756–23765.

[35] J. Chang, S. Wang, H.-M. Xu, Z. Chen, C. Yang, F. Zhao, Detrdistill: A universal knowledge distillation framework for detr-families, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 6898–6908.

[36] Z. Xie, F. Shao, B. Mu, H. Chen, Q. Jiang, C. Lu, Y.-S. Ho, Bgdfnet: bidirectional gated and dynamic fusion network for rgb-t crowd counting in smart city system, IEEE Transactions on Instrumentation and Measurement 73 (2024) 1–16.

[37] K. Zhou, H. Yan, Mmffnet: Multi-modal feature fusion network for rgb-t crowd counting, IEEE Transactions on Instrumentation and Measurement (2025).

[38] X. Cao, Z. Wang, Y. Zhao, F. Su, Scale aggregation network for accurate and efficient crowd counting, in: Proceedings of the European conference on computer vision, 2018, pp. 734–750.

[39] Y. Li, X. Zhang, D. Chen, Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1091–1100.

[40] Q. Lan, Q. Tian, Instance, scale, and teacher adaptive knowledge distillation for visual detection in autonomous driving, IEEE Transactions on Intelligent Vehicles 8 (3) (2022) 2358–2370.

[41] Y. Yang, X. Sun, W. Diao, H. Li, Y. Wu, X. Li, K. Fu, Adaptive knowledge distillation for lightweight remote sensing object detectors optimizing, IEEE Transactions on Geoscience and Remote Sensing 60 (2022) 1–15.

[42] R. Guerrero-Gómez-Olmedo, B. Torre-Jiménez, R. López-Sastre, S. Maldonado-Bascón, D. Onoro-Rubio, Extremely overlapping vehicle counting, in: Iberian conference on pattern recognition and image analysis, Springer, 2015, pp. 423–431.

[43] F. Luo, Y. Li, G. Zeng, P. Peng, G. Wang, Y. Li, Thermal infrared image colorization for nighttime driving scenes with top-down guided attention, IEEE Transactions on Intelligent Transportation Systems 23 (9) (2022) 15808–15823.