Multi-modal Crowd Counting via a Broker Modality

Haoliang Meng[1], Xiaopeng Hong[1,2⋆], Chenhao Wang[1], Miao Shang[1], and Wangmeng Zuo[1,2]

[1] Harbin Institute of Technology
[2] Pengcheng Laboratory
menghaoliang2002@163.com hongxiaopeng@ieee.org 22b903078@stu.hit.edu.cn
miaos0522@gmail.com wmzuo@hit.edu.cn

Abstract. Multi-modal crowd counting involves estimating crowd density from both visual and thermal/depth images. This task is challenging due to the significant gap between these distinct modalities. In this paper, we propose a novel approach by introducing an auxiliary broker modality and on this basis frame the task as a triple-modal learning problem. We devise a fusion-based method to generate this broker modality, leveraging a non-diffusion, lightweight counterpart of modern denoising diffusion-based fusion models. Additionally, we identify and address the ghosting effect caused by direct cross-modal image fusion in multi-modal crowd counting. Through extensive experimental evaluations on popular multi-modal crowd counting datasets, we demonstrate the effectiveness of our method, which introduces only 4 million additional parameters, yet achieves promising results. The code is available at `https://github.com/HenryCilence/Broker-Modality-Crowd-Counting`.

Keywords: Multi-modal crowd counting · Intermediate modality · Modality fusion · Diffusion model

1 Introduction

Multi-modal crowd counting [17, 24, 39] is a challenging task which integrates information from various sources to provide a comprehensive understanding of complex crowds leveraging the complementary and commonality nature of diverse modalities. However, there are two major challenges bringing difficulty for feature interaction and fusion. **On the one hand**, since vision and thermal images capture reflectance information in different bands of the spectrum, they look naturally different. Vision images excel at capturing color and texture information of visible light while thermal images are adept at containing information from invisible electromagnetic radiation. Therefore, vision images contain hardly any human figures under poor illumination conditions, while thermal images can clearly reflect the outline of each person under such conditions. Moreover, heating negative objects look almost the same in thermal images but can be easily

⋆ Corresponding author.

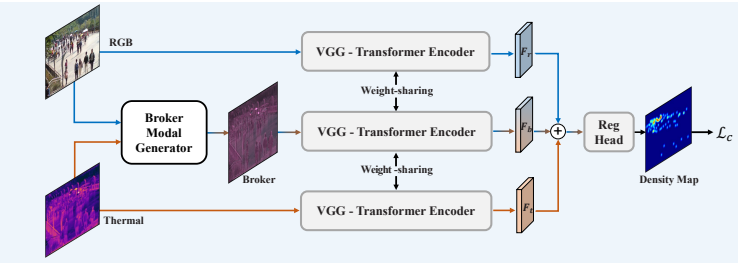


Fig. 1: Illustration of the framework of our method. By leveraging the Broker Modal Generator (BMG) to introduce an auxiliary broker modality, we frame the dual-modal visual-thermal crowd counting task as a triple-modal learning problem.

distinguished in vision images. **On the other hand**, RGB and thermal images are not strictly aligned in position due to the difference in shooting angle. Directly fusing the two modalities causes a serious ghosting effect. The misalignment of RGB-thermal image pairs may lead to inaccuracy during the counting process, since ghosting may cause people to be counted twice. Therefore, simply extracting and fusing features of these two modalities during cross-modal learning may not perform well.

In order to solve these challenges, researchers have proceeded with abundant cross-modal feature interactions through the counting model. For example, Liu *et al*. [24] propose a modality-shared branch to fully capture the shared information of different modalities. Pan *et al*. [36] feed thermal features to RGB image features to complete feature interaction. Liu *et al*. [29] utilize a learnable count token to interact and fuse two-modal information at the scale level. However, all such methods directly perform feature interactions between the two original modalities and fail to achieve satisfactory performance. In contrast, we introduce an intermediate auxiliary modality to assist cross-modal learning.

Motivated by the analysis above, this paper presents a novel multi-modal crowd-counting approach. The basic idea is to transform a learning task from two modalities with a significant gap into a learning task involving three modalities by introducing an intermediate auxiliary modality, as shown in Fig. 1. This intermediate modality, which we term a 'broker' modal, serves to bridge and align the features of the two original modalities for smoother, more coherent, and effective fusion. While in the literature, methods for constructing intermediate modalities based on contrast learning have been used in similar tasks such as cross-modal person re-identification [11,47,59], it is evident that contrast learning cannot be directly applied to regression tasks.

To bridge this gap and generate the intermediate broker modality, we propose leveraging image fusion techniques, as the resulting fused image integrates information and characteristics from both modalities. The pivotal challenge lies in selecting the most appropriate fusion approach [7, 51, 61]. Among the available methods, recent denoising diffusion model-based fusion methods such as [62] stand out for their ability to preserve both structural and detailed information

from the source images and achieve promising fusion results. However, these methods inherently entail a large-scale and computationally demanding nature, rendering them inefficient for direct integration as a module within our framework for generating intermediate modalities. To address this challenge, we introduce the idea of constructing a non-diffusion, lightweight counterpart of the denoising diffusion-based fusion model through a distillation-then-finetuning strategy. This model is initially built via a distillation process guided by denoising diffusion fusion models, followed by fine-tuning to tailor it to the crowd counting task. It inherits the high fusion capability of the denoising diffusion model while also addressing the low efficiency in training and fine-tuning associated with the diffusion model. Remarkably, we discover that such a lightweight non-diffusion model can accurately fuse the two modalities like a denoising diffusion model while being efficiently trained. Moreover, we identify the ghosting effect caused by directly fusing the two modalities and demonstrate that the designed distillation-then-finetuning learning scheme of our framework alleviates this phenomenon. We conduct comprehensive experiments on the popular multi-modal crowd counting benchmarks. Extensive experimental results demonstrate that our proposed approach significantly improves the counting performance.

The main contributions can be further summarized as follows:

- We introduce an intermediate auxiliary modality and frame the dual-modal visual-thermal crowd counting task as a triple-modal learning problem.
- We carefully design a broker modal generator, as a non-diffusion, lightweight counterpart of the denoising diffusion-based fusion model.
- We provide an in-depth analysis of the ghosting effect caused by directly fusing the two modalities and provide a solution to alleviate this phenomenon.

2 Related Work

2.1 Multi-modal Crowd Counting

With the development of deep learning technology, crowd counting has been increasingly studied and has become a significant computer vision task. There are three main technical routes among popular crowd counting methods: based on detection [22,23,42,45], based on regression [27,53] and based on density map [20,21,25,26,31,32]. Up to now, crowd counting methods for single modality have been well-studied. Researchers have made great achievements in ameliorating model structure [19, 21, 32] and loss function [18, 31, 33].

In recent years, with the development of multi-modal learning in various fields [2, 15, 40, 54], multi-modal crowd counting [3, 13, 38, 48, 56] has received widespread attention. Researchers use additional modalities such as thermal or depth together with visual images to enhance the performance of crowd counting methods. For example, Peng *et al.* [39] propose a drone-based RGB-thermal multi-modal crowd counting dataset covering various attributes and design a multi-modal crowd counting network to learn cross-modal representation. Liu *et*

al. [24] propose an RGB-thermal multi-modal crowd counting dataset containing complex environments and design a cross-modal collaborative representation learning framework to capture the complementary information of different modalities fully. Liu *et al*. [29] use count-guided multi-modal fusion and modal-guided count enhancement to better excavate multi-modal information. Lian *et al*. [17] collect a large-scale RGB-D crowd counting dataset and propose a regression-guided detection network for simultaneously crowd counting and localization. Li *et al*. [13] propose a cross-modal cyclic attention fusion model that integrates RGB and depth features in a cyclic attention manner to comprehensively model and fuse information from both modalities, which tackles the negative effect of the arbitrary crowd distribution on the counting task.

Most previous approaches focus on extracting features from two modalities and then interacting and fusing them to generate the density map. However, the significant gap between visible and infrared images brings difficulties to feature interaction and fusion. In this paper, we introduce an auxiliary intermediate modality to reduce the gap and participate in feature interaction and fusion.

2.2 Intermediate Modality in Infrared-Visible Cross-Modal Tasks

In various infrared-visible cross-modal tasks, one of the major challenges is the modality discrepancy between visible and infrared images. Aiming to mitigate the challenging modality discrepancy, quite a few recent approaches leverage intermediate spaces to bridge vision and infrared modalities. They design an intermediate modality generator to siphon off knowledge from visible and infrared images and output the intermediate images to assist the cross-modal process.

Li *et al*. [11] first propose an X modality generated by a lightweight network as an assistant to reduce modality discrepancy. Afterward, generators of various structures have been designed. For example, Zhang *et al*. [59] propose a non-linear middle modality generator to project vision and infrared images into a unified middle modality image space and generate middle-modality images. Alehdaghi *et al*. [1] introduce an Adaptive Generation of Privileged Intermediate Information to adapt and generate a virtual domain that bridges discriminant information between vision and infrared modalities. Wang *et al*. [46] introduce an image-level sub-network to unify the representations for images with different modalities. Recently, Zhao *et al*. [60] propose an infrared and visible image fusion method via meta-feature embedding for object detection, utilizing a fusion modality to assist the cross-modal detection process.

However, most approaches generate intermediate modality through contrastive learning and are designed for cross-modal person re-identification. Defining positive and negative pairs for contrastive learning in a dense regression task such as crowd counting is not straightforward. Consequently, applying such a scheme to cross-modal crowd counting is not feasible without significant modification. This presents an opportunity for us to introduce a new intermediate modality.

3 Method

3.1 Counting Framework

By introducing an auxiliary broker modality, we frame the dual-modal visual-thermal crowd counting task as a triple-modal learning problem. As illustrated in Fig. 1, our counting model consists of three parts: a broker modal generator (BMG), a feature extractor, and a regression head. The feature extractor is implemented by an architecture of 'VGG-19 - Transformer encoder' and is weight-sharing for all modalities. We first use BMG to generate the broker modality $F \in \mathbb{R}^{3\times H\times W}$, then feed it to the feature extractor along with two source modalities R and $T \in \mathbb{R}^{3\times H\times W}$.

$$F = g\left(R, T\right), \tag{1}$$

$$F_r = \zeta_v\left(\zeta_c\left(R\right)\right), F_b = \zeta_v\left(\zeta_c\left(F\right)\right), F_t = \zeta_v\left(\zeta_c\left(T\right)\right), \tag{2}$$

where g denotes the broker modal generator, R and T denote RGB and thermal images, and $F_r, F_b, F_t \in \mathbb{R}^{D\times h\times w}$ are features from three modalities. ζ_c and ζ_v denote VGG-19 and Transformer encoder in the feature extractor, respectively.

Finally, we sum the features up and send the summation to the regression head R to obtain a one-channel density map $D^{est} \in \mathbb{R}^{1\times h\times w}$.

Considering the performance and robustness, we adopt the popular Bayesian Loss [31] as the loss function to align discrete point annotations and the continuous density map.

$$\mathcal{L}_c = \sum_{i=1}^{M} |1 - \langle \frac{\mathcal{N}\left(z_i, \sigma^2\mathbf{1}_{2\times 2}\right)}{\sum_{n=1}^{M}\mathcal{N}\left(z_n, \sigma^2\mathbf{1}_{2\times 2}\right)}, D^{est}\rangle|, \tag{3}$$

where M is the total crowd count of the image and z_n is the n-th annotated head point position in the image. $\langle\cdot,\cdot\rangle$ denotes the inner product operation.

3.2 Broker Modality Generator

We introduce an auxiliary broker modal to bridge RGB and thermal modalities. Unlike most contrastive learning-based methods mentioned above [11, 59], our auxiliary modal is constructed on multi-modal image fusion. We choose DDFM [62], one state-of-the-art multi-modal image fusion method as the reference to construct our Broker Modal Generator (BMG). Nonetheless, the denoising diffusion model usually suffers from high computation costs in both training and prediction, making it inefficient to serve as a part of an end-to-end counting mode. Moreover, it directly fuses original images without registration, causing the ghosting effect in fusion images. To alleviate this issue in an end-to-end training and prediction multi-modal counting framework, we design a lightweight non-diffusion generator BMG as its counterpart and suit it for the task. We design

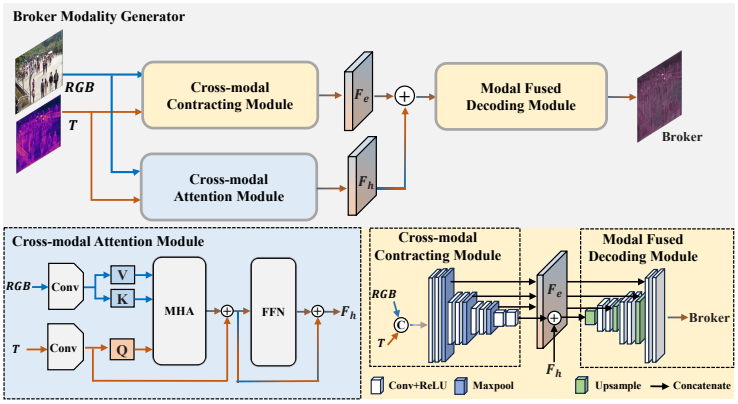


Fig. 2: The framework of our broker modal generator. Specifically, the generator consists of a cross-modal contracting module and a modal fused decoding module for image reconstruction, and a cross-modal attention module to enhance modality alignment.

a two-stage learning scheme for BMG, which consists of a distillation stage and a fine-tuning stage. In the following sections, we will elaborate on the structure of the broker modality generator and its learning scheme.

The structure of BMG. As illustrated in Fig. 2, the generator consists of three parts: A cross-modal contracting module (CMC), a modal fused decoding module (MFD), and a cross-modal attention module (CMA). The cross-modal contracting module f_e and modal fused decoding module f_d are implemented by U-Net [41] encoder and decoder, respectively, while the cross-modal attention module is implemented by a 1-layer, 4-head cross-attention Transformer encoder. Firstly, we use two 3×3 convolutional layers ξ_r and ξ_t to extract the feature of RGB and thermal images. Then, we concatenate them together as the input of CMC f_e and obtain the encoding result $F_e \in \mathbb{R}^{d \times h \times w}$:

$$F_e = f_e\left(\left[\xi_r\left(R\right); \xi_t\left(T\right)\right]\right), \tag{4}$$

where $[\,\cdot\,;\,\cdot\,]$ denotes the channel concatenating operation.

Furthermore, we use cross-attention to enhance the encoding results of CMC. We use convolutional layers to extract feature maps of two original modalities and flatten them to a sequence of patch embedding X_r, $X_t \in \mathbb{R}^{hw \times d}$. We use T modal embedding as Query(Q) and RGB modal embedding as Key(K) and Value(V) for the multi-head cross-attention operation. Note that the residual connection in cross-attention is contributed by the T modal.

$$\begin{aligned} X_r &= \xi\left(R\right), \; X_t = \xi\left(T\right), \\ Q &= X_t W_q, \; K = X_r W_k, \; V = X_r W_v, \\ H &= Att\left(Q, K, V\right) + X_t, \\ F_h &= FFN(H) + H, \end{aligned} \tag{5}$$

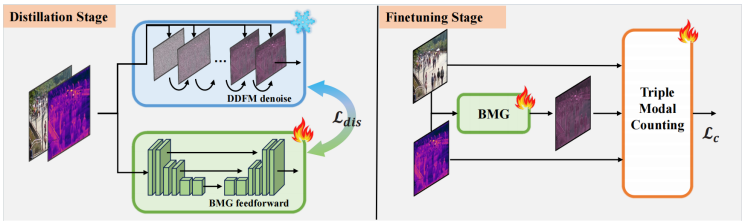


Fig. 3: Illustration of the distillation-then-finetuning strategy. In the distillation stage, BMG is initialized using a distillation process guided by DDFM. In the fine-tuning stage, BMG is tuned together with the feature extractor to suit the counting task.

where ξ is the convolutional layer in the cross-modal attention module, W_q,W_k,W_v are the projection weights for multi-head attention layer, FFN denotes the feed-forward network in Transformer encoder, and

$$Att\left(Q,K,V\right)=Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V. \tag{6}$$

Finally, we add the enhancement result $F_h \in \mathbb{R}^{d\times h\times w}$ to the encoding result F_e as the input of MFD and obtain the multi-modal fusion results $F \in \mathbb{R}^{3\times H\times W}$:

$$F=f_d\left(F_e+F_h\right). \tag{7}$$

The distillation-then-finetuning learning scheme of BMG. The learning process consists of a distillation stage and a fine-tuning stage. During the distillation stage, we pre-train BMG by distilling DDFM as its counterpart. During the fine-tuning stage, we tune BMG together with the feature extractor to better suit the counting task and bridge two original modalities.

The distillation stage. Before counting, we first pre-train BMG using MSE Loss in the first stage to enable it to acquire the infrared-visible fusion capability, as shown in Fig. 3. We achieve this through distilling DDFM[3] [62], one state-of-the-art image fusion model. During the distillation stage, BMG approximates the high fusion capability of DDFM with a much lower scale and efficiency.

$$\mathcal{L}_{dis}=\sum_{i=1}^{K}||g\left(R,T\right)-p\left(R,T\right)||_2^2, \tag{8}$$

where p denotes DDFM, and K is the number of training samples.

The fine-tuning stage. In the fine-tuning stage, we generate the broker modal by BMG and send it to the feature extractor together with the original modalities, framing the dual-modal counting task as a triple-modal learning problem.

[3] For the infrared-visible fusion task, the DDFM model retains both structural and detailed information from the source images, meeting the visual fidelity requirements as well. Thus it is used to pre-train our broker modal generator based on its favorable multi-modal fusion results.

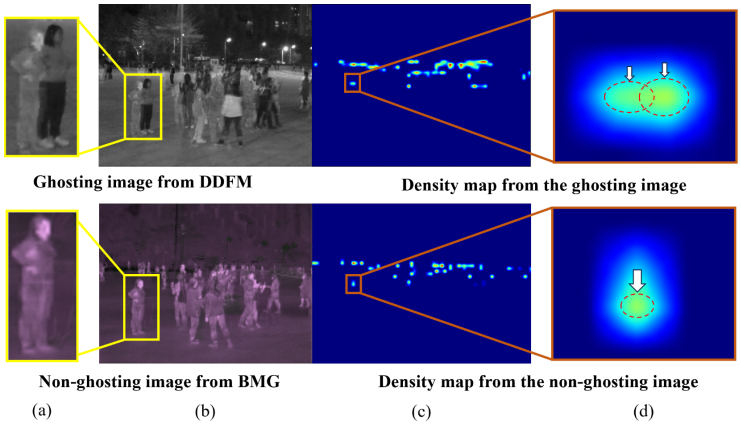


Fig. 4: Comparison of ghosting and non-ghosting intermediate images and the corresponding density maps. (b) are the intermediate images generated by DDFM directly and by our BMG, and (a) shows the comparison of the ghosting region in (b). (c) are the corresponding density maps generated from the scenes in (b), and (d) shows the difference of the density maps in the ghosting region.

We tune BMG together with the feature extractor during the counting process to suit the counting task. During fine-tuning, BMG acquires the capability to bridge the two original modalities coherently and to alleviate the ghosting phenomenon caused by space misalignment. The influence and effect of the two-stage learning process are shown in Section 4.5.

3.3 Discussion on the ghosting effect.

As visual and thermal images are captured by two distinct sensors, there exists an angular disparity or spatial displacement between them. Therefore, fusing them directly using DDFM may lead to a ghosting effect[4] in the fusion image. For instance, in the first line of Fig. 4(a), two overlapping figures for one person are evident in the yellow rectangular area. In such cases, counting models may mistakenly interpret such a fusion region as a region of two separate entities, resulting in a density region containing overlapping Gaussian distributions, as depicted in the first line of Fig. 4(d).

Ghosting effect alleviation by two-stage learning. The proposed two-stage learning scheme, somehow surprisingly, mitigates this issue well. The key is to

[4] To better illustrate the challenges posed by this problem, we conducted experiments to assess the effectiveness of image alignment-based de-ghosting algorithms such as [4,9,30]. Our results indicate that while satisfactory outcomes can be attained with natural image pairs, the de-ghosting algorithm performs inadequately on image pairs with low imaging quality, such as the data used in this study. Further comprehensive experiments are available in the supplementary material.

decouple the distillation of denoising diffusion models from the counting process. In the distillation stage, the training process exclusively distills a heavy denoising diffusion multi-modal image fusion model (with a size of over 500M, as shown in Tab. 6) using a lightweight network (with a size of less than 4M). This single training task ensures the convergence of the training process. In the fine-tuning stage, the training process fine-tunes the pre-trained broker modal generator to the counting task, facilitating the elimination of the ghosting effect. Specifically, as the analysis above suggests, the ghosting effect may cause the model to generate overlapping Gaussian distributions and overestimate the count. By calculating the counting loss, which directly reflects the produced density map's deviation from the ground truth, the error caused by the ghosting effect is penalized during fine-tuning. Consequently, this leads to a clearer fusion image with a significantly reduced ghosting effect.

4 Experiments and Discussion

4.1 Evaluation Metrics and Datasets

We conduct our experiments on three widely used multi-modal crowd counting benchmark datasets: RGBT-CC, DroneRGBT, and ShanghaiTechRGBD. We adopt RMSE [43] and GAME [5] to evaluate the methods.

RGBT-CC [24] is the most widely used large-scale RGBT crowd counting benchmark, in which 138,389 pedestrians are annotated in 2,030 pairs of RGB-thermal images captured in unconstrained scenarios. All image pairs are captured by an optical-thermal camera in various scenarios like malls, streets, playgrounds, train stations, and metro stations.

DroneRGBT [39] is a drone-based RGBT crowd counting benchmark dataset. It contains 3,607 registered RGB-thermal image pairs with ground truth annotations and covers different attributes including height, illumination, and density. Images are captured by drone-mounted cameras, covering a wide range of scenarios like campus, street, park, parking lot, playground, and plaza.

ShanghaiTechRGBD [17] is a large-scale RGBD crowd counting dataset containing 2,193 RGB-depth image pairs with 144,512 annotations. Images are captured by a stereo camera whose valid depth ranges from 0 to 20 meters. The scenes include busy streets of metropolitan areas and crowded public parks.

4.2 Implementation Details

Network structure. CMC of BMG downsamples the input image size (H, W) to (h, w) while increasing channels from 6 to $d = 256$. MFD upsamples the features to the full input size (H, W) and reconstructs a 3-channel broker modal F. The patch number for cross-modal attention is 64 and embedding dimension $d = 256$. The number of the Transformer encoder layer is 2 and the head number is 6. The patch number of the Transformer encoder is 196 and the embedding dimension is $D = 768$. We upsample the output of the feature extractor to (h, w), and then feed it to the regression head.

Table 1: Comparison with the state-of-the-art methods on RGBT-CC.

Method	Venue	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
UCNet [56]	CVPR 2020	33.96	42.42	53.06	65.07	56.31
HDFNet [38]	ECCV 2020	22.36	27.79	33.68	42.48	33.93
MVMS [57]	CVPR 2019	19.97	25.10	31.02	38.91	33.97
BBSNet [3]	ECCV 2020	19.56	25.07	31.25	39.24	32.48
CmCaF [13]	TII 2022	15.87	19.92	24.65	28.01	29.31
BL+IADM [24]	CVPR 2021	15.61	19.95	24.69	32.89	28.18
CSCA [58]	ACCV 2022	14.32	18.91	23.81	32.47	26.01
BL+MAT [48]	ICME 2022	13.61	18.08	22.79	31.35	24.48
TAFNet [44]	ISCAS 2022	12.38	16.98	21.86	30.19	22.45
BL+MAT+SSP [48]	ICME 2022	12.35	16.29	20.81	29.09	22.53
GETANet [36]	GRSL 2024	12.14	15.98	19.40	28.61	22.17
DEFNet [63]	TITS 2022	11.90	16.08	20.19	27.27	21.09
R2T [12]	KBS 2022	11.63	16.70	22.12	32.32	21.28
MJPNet-T [64]	IoT 2024	11.56	16.36	20.95	28.91	17.83
MCN [10]	ESWA 2024	11.56	15.92	20.16	28.06	19.02
MC^{3}Net [65]	TITS 2023	11.47	15.06	19.40	27.95	20.59
C4-MIM [6]	CIS 2024	11.08	14.85	19.05	24.94	19.71
CAGNet [52]	ESWA 2024	11.06	14.73	18.94	25.76	17.83
BGDFNet [50]	TIM 2024	11.00	15.04	19.86	29.72	19.05
VPMFNet [35]	IoT 2024	10.99	15.17	20.07	28.03	19.61
MSDTrans [29]	BMVC 2022	10.90	14.81	19.02	26.14	18.79
Ours	ECCV 2024	**10.19**	**13.61**	**17.65**	**23.64**	**17.32**

Table 2: Comparison with the state-of-the-art methods on DroneRGBT.

Method	Venue	GAME(0)	RMSE
CMDBIT [49]	TCSVT 2023	11.50	21.27
BL+IADM [24]	CVPR 2021	11.41	17.54
DEFNet [63]	TITS 2022	10.87	17.93
CSRNet [14]	CVPR 2018	8.91	13.80
GETANet [36]	GRSL 2024	8.44	13.99
CGINet [37]	EAAI 2023	8.37	13.45
MC^{3}Net [65]	TITS 2023	7.33	11.17
MMCCN [39]	ACCV 2020	7.27	11.45
I-MMCCN [55]	IC-NIDC 2021	6.91	11.26
Ours	ECCV 2024	**6.20**	**10.40**

Experimental setup and training details. We implement our model with PyTorch and train it on an NVIDIA RTX 3090 GPU with 24GB memory. We adopt the Adam optimizer with the learning rate of 1e-5 and the weight decay of 1e-4. During training, we randomly crop and flip the input images for data augmentation. The crop size is 224×224 for RGBT-CC and DroneRGBT, and 1024×1024 for ShanghaiTechRGBD. The batch size for all datasets is 1 and the maximum number of training epochs is 400.

4.3 Experimental Evaluations

We compare our proposed method with the state-of-the-art methods on the benchmark datasets described above.

Experimental results on RGB-T datasets. Experimental results on RGBT-CC and DroneRGBT are shown in Tab. 1 and Tab. 2, respectively. The ex-

Table 3: Comparison with the state-of-the-art methods on ShanghaiTechRGBD.

Method	Venue	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
UCNet [56]	CVPR 2020	10.81	15.24	22.04	32.98	15.70
DetNet [23]	CVPR 2018	9.74	-	-	-	13.14
HDFNet [38]	ECCV 2020	8.32	13.93	17.97	22.62	13.01
CL [8]	ECCV 2018	7.32	-	-	-	10.48
GETANet [36]	GRSL 2024	6.91	8.83	12.31	17.11	9.88
BL+MAT [48]	ICME 2022	6.61	8.16	10.09	16.62	9.39
BBSNet [3]	ECCV 2020	6.26	8.53	11.80	16.46	9.26
BL+MAT+SSP [48]	ICME 2022	5.39	6.73	8.98	13.66	7.77
VPMFNet [35]	IoT 2024	5.24	6.85	9.14	13.81	8.19
RDNet [17]	CVPR 2019	4.96	-	-	-	7.22
CSCA [58]	ACCV 2022	4.39	6.47	8.82	11.76	6.39
CSRNet+IADM [24]	CVPR 2021	4.38	5.95	8.02	11.02	7.06
DPDNet [16]	TPAMI 2021	4.23	5.67	7.04	9.64	6.75
AGCCM [34]	TIP 2022	3.90	-	-	-	5.86
CCANet [28]	TMM 2023	3.78	5.34	7.63	10.17	5.56
Ours	ECCV 2024	**3.72**	6.06	9.02	12.63	**5.28**

Table 4: Comparison of parameter quantity with the state-of-the-art models

Method	Venue	Param(M)↓
HDFNet [38]	ECCV 2020	45.63
UCNet [56]	CVPR 2020	31.08
BBSNet [3]	ECCV 2020	49.78
BL+IADM [24]	CVPR 2021	25.67
DEFNet [63]	TITS 2022	45.33
MC^3Net [65]	TITS 2023	113.08
Ours	ECCV 2024	40.55

perimental results demonstrate that our method clearly outperforms previous state-of-the-art models and becomes the new state-of-the-art method on both RGB-T datasets. For example, compared to MSDTrans [29], our method shows significant improvement on all five evaluation metrics on RGBT-CC. It reduces GAME(0), GAME(1), GAME(2), GAME(3), and RMSE by 0.71, 1.20, 1.37, 2.50, and 1.47, respectively. Compared to I-MMCCN [55] on DroneRGBT, it improves achievement of 0.71 and 0.86 on GAME(0) and RMSE, respectively. Furthermore, we provide visual comparison results with previous state-of-the-art methods in Fig. 5. These quantitative and qualitative experiments demonstrate that our proposed method is greatly effective and robust.

Experimental results on RGB-D dataset. Furthermore, to thoroughly evaluate the effectiveness and robustness of our proposed multi-modal crowd counting approach, we conducted experiments in the realm of RGB-depth crowd counting tasks. Experimental results on ShanghaiTechRGBD are shown in Tab. 3. Notably, our approach achieves the best GAME(0) value of 3.72 and the best RMSE value of 5.28. These compelling results underscore the robustness and versatility of our proposed method in handling the intricacies associated with multi-modal crowd counting tasks.

Table 5: Influence of the distillation-then-finetuning training scheme for BMG.

Distillation stage	Fine-tuning stage	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
×	✓	12.93	16.12	20.27	26.07	25.49
✓	×	11.29	14.52	18.70	24.41	20.85
✓	✓	**10.19**	**13.61**	**17.65**	**23.64**	**17.32**

Parameter quantity of different models. We also analyze the complexity of models by comparing the parameter quantity with state-of-the-art methods, as in Tab. 4. Despite achieving remarkable results in terms of crowd counting accuracy, it is noteworthy that our model maintains an acceptable and competitive parameter quantity when benchmarked against contemporary state-of-the-art methodologies.

In general, the results above demonstrate that our method achieves high accuracy in both RGB-T and RGB-D crowd counting tasks. These results show the remarkable robustness and effectiveness of our approach in addressing the challenging multi-modal crowd counting task.

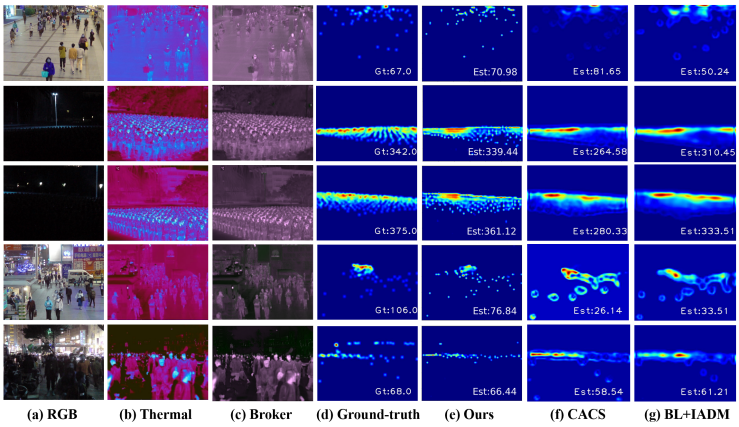


Fig. 5: Visualization results of our method compared with different models. Our method consistently outperforms compared with other previous state-of-the-art models, demonstrating the effectiveness and superiority of the proposed model in accurately estimating multi-modal crowd density.

4.4 Influence of Distillation-then-finetuning Training Scheme

Here we investigate the influence of designing the distillation-then-finetuning training scheme for BMG. First, Tab. 5 shows that the two-stage training scheme greatly improves the performance of the model. If training without DDFM distillation initialization or fine-tuning BMG, the performance obviously decreases

Table 6: Comparison of different intermediate modality generating methods.

Method	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE	Time cost	Params
DDFM [62]	11.18	14.84	19.10	25.57	18.32	59s	552.81M
BMG	**10.19**	**13.61**	**17.65**	**23.64**	**17.32**	**0.022s**	**3.89M**

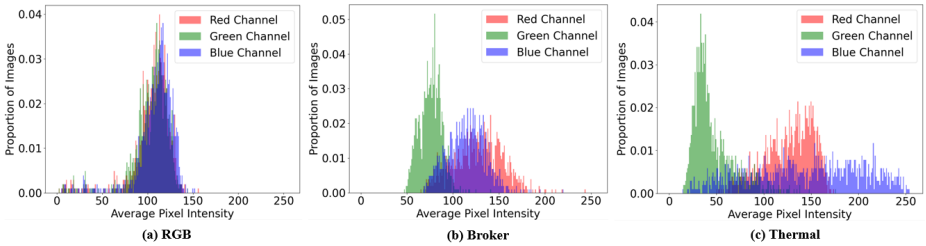


Fig. 6: Histograms of the average single-color-channel intensity of all pixels inside an image over the training set of RGBT-CC. They are computed from visible, broker, and thermal images, respectively.

(*e.g.* GAME(0) from 10.19 to 11.29 and 12.93). This validates both the distillation and fine-tuning stages in two-stage learning. Besides, Tab. 6 shows the performance of counting using different intermediate modalities, the time cost of generating one intermediate image, and the parameter quantity of different generators. Compared with DDFM, our BMG achieves superior results with a more economical cost. Moreover, the simple and elegant fine-tuning of BMG unexpectedly alleviates the ghosting phenomenon, as shown in Fig. 7

Moreover, Fig. 6 suggests that BMG finally learns a proper auxiliary modality with an intermediate look between visible and thermal images. With BMG, cross-modal learning becomes much easier as it avoids the whole counting model to deal with the gap between visible and infrared modalities independently.

4.5 Ablation Studies

We perform a comprehensive analysis and ablation study on RGBT-CC using the proposed model, as shown in Tab. 7. We here investigate the effectiveness of the structure of BMG. From Tab. 7, we notice that compared with using only original modalities, training a lightweight network by distilling DDFM to generate a broker modality enhances the capability of the counting model (*e.g.* GAME(0) is 10.49 and RMSE is 18.03). The performance is clearly improved when we further introduce cross-modal attention (*e.g.* GAME(0) is 10.19 and RMSE is 17.32). Moreover, as shown in Fig. 7, our BMG alleviates the ghosting effect in the broker modality greatly, especially when the space misalignment is significant (such as in images in the last three lines). All these results above suggest that BMG enhances the fusion capability of the counting model. With BMG, the whole model better mitigates the detrimental effects of modal misalignment during the fine-tuning stage.

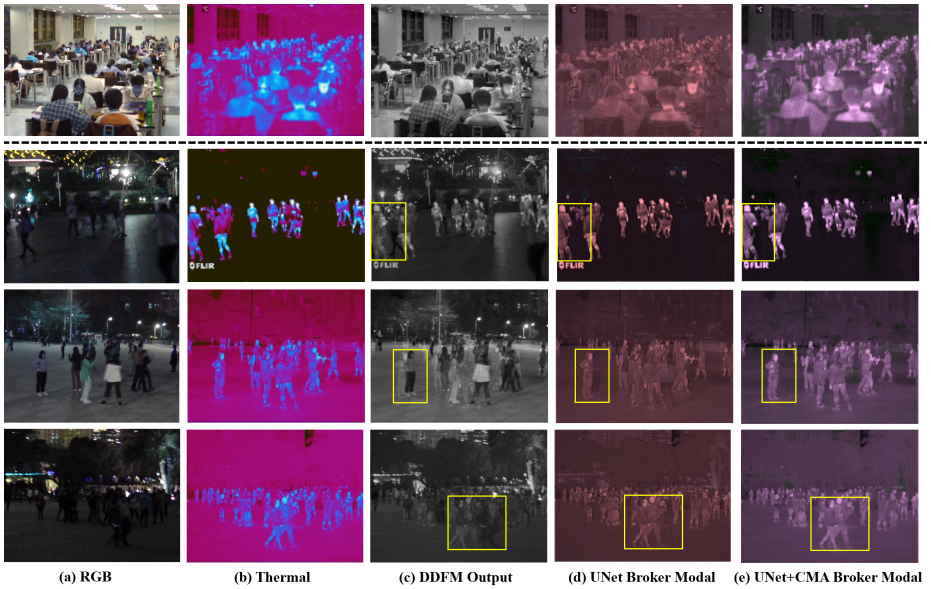


Fig. 7: Original and broker modalities generated by different generators. The first row is the case where the space misalignment is slight, and the remaining is the case where the space misalignment is relatively significant. The broker modal harmoniously bridges the attributes of the two source modalities and mitigates the gap between them. Moreover, the ghosting caused by misalignment in space is greatly eliminated during the fine-tuning stage, as shown in the yellow rectangular area.

Table 7: The impact of different components of the model on RGBT-CC dataset.

CMC & MFD	Cross-modal attention	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
×	×	14.94	18.29	22.10	29.43	28.74
✓	×	10.49	14.11	18.28	24.37	18.03
✓	✓	**10.19**	**13.61**	**17.65**	**23.64**	**17.32**

5 Conclusion

This paper presents a novel approach to multi-modal crowd counting method. By framing the dual-modal visual-thermal crowd counting task as a triple-modal learning problem and introducing a broker modality, we effectively bridge the features of the two original modalities for coherent multi-modal learning. Our proposed broker modality generator, characterized by a non-diffusion, lightweight structure, retains the strong fusion capabilities of denoising diffusion fusion models. Extensive experimental results show that with minor extra parameters, our method outperforms other competitors and achieves promising results. Additionally, we offer a thorough analysis of the ghosting effect inherently in image fusion and justify the effectiveness of the proposed distillation-then-finetuning training strategy for alleviating it.

Acknolwedgement

This work was funded in part by the National Natural Science Foundation of China (62076195, 62376070) and in part by the Fundamental Research Funds for the Central Universities (AUGA5710011522).

References

1. Alehdaghi, M., Josi, A., Shamsolmoali, P., Cruz, R.M., Granger, E.: Adaptive generation of privileged intermediate information for visible-infrared person re-identification. arXiv preprint arXiv:2307.03240 (2023)
2. Chen, K., Chen, J.K., Chuang, J., Vázquez, M., Savarese, S.: Topological planning with transformers for vision-and-language navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11276–11286 (2021)
3. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In: European conference on computer vision. pp. 275–292. Springer (2020)
4. Gao, J., Cai, X.f.: Image matching method based on multi-scale corner detection. In: 2017 13th International Conference on Computational Intelligence and Security (CIS). pp. 125–129. IEEE (2017)
5. Guerrero-Gómez-Olmedo, R., Torre-Jiménez, B., López-Sastre, R., Maldonado-Bascón, S., Onoro-Rubio, D.: Extremely overlapping vehicle counting. In: Pattern Recognition and Image Analysis: 7th Iberian Conference, IbPRIA 2015, Santiago de Compostela, Spain, June 17-19, 2015, Proceedings 7. pp. 423–431. Springer (2015)
6. Guo, Q., Yuan, P., Huang, X., Ye, Y.: Consistency-constrained rgb-t crowd counting via mutual information maximization. Complex & Intelligent Systems pp. 1–22 (2024)
7. Huang, Z., Liu, J., Fan, X., Liu, R., Zhong, W., Luo, Z.: Reconet: Recurrent correction network for fast and efficient multi-modality image fusion. In: European Conference on Computer Vision. pp. 539–555. Springer (2022)
8. Idrees, H., Tayyab, M., Athrey, K., Zhang, D., Al-Maadeed, S., Rajpoot, N., Shah, M.: Composition loss for counting, density map estimation and localization in dense crowds. In: Proceedings of the European conference on computer vision (ECCV). pp. 532–546 (2018)
9. Jiang, Q., Liu, Y., Yan, Y., Deng, J., Fang, J., Li, Z., Jiang, X.: A contour angle orientation for power equipment infrared and visible image registration. IEEE Transactions on Power Delivery **36**(4), 2559–2569 (2020)
10. Kong, W., Liu, J., Hong, Y., Li, H., Shen, J.: Cross-modal collaborative feature representation via transformer-based multimodal mixers for rgb-t crowd counting. Expert Systems with Applications p. 124483 (2024)
11. Li, D., Wei, X., Hong, X., Gong, Y.: Infrared-visible cross-modal person re-identification with an x modality. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 4610–4617 (2020)
12. Li, H., Zhang, S., Kong, W.: Learning the cross-modal discriminative feature representation for rgb-t crowd counting. Knowledge-Based Systems **257**, 109944 (2022)
13. Li, H., Zhang, S., Kong, W.: Rgb-d crowd counting with cross-modal cycle-attention fusion and fine-coarse supervision. IEEE Transactions on Industrial Informatics **19**(1), 306–316 (2022)

14. Li, Y.C.: Dilated convolutional neural networks for understanding the highly congested scenes/y. li, x. zhang, d. chen. In: Proceedings of the IEEE conference on computer vision and pattern recognition.–IEEE. pp. 1091–1100 (2018)
15. Li, Y., Wang, H., Luo, Y.: A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In: 2020 IEEE international conference on bioinformatics and biomedicine (BIBM). pp. 1999–2004. IEEE (2020)
16. Lian, D., Chen, X., Li, J., Luo, W., Gao, S.: Locating and counting heads in crowds with a depth prior. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9056–9072 (2021)
17. Lian, D., Li, J., Zheng, J., Luo, W., Gao, S.: Density map regression guided detection network for rgb-d crowd counting and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1821–1830 (2019)
18. Lin, H., Hong, X., Ma, Z., Wei, X., Qiu, Y., Wang, Y., Gong, Y.: Direct measure matching for crowd counting. the Thirtieth International Joint Conference on Artificial Intelligence (2021)
19. Lin, H., Ma, Z., Hong, X., Shangguan, Q., Meng, D.: Gramformer: Learning crowd counting via graph-modulated transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3395–3403 (2024)
20. Lin, H., Ma, Z., Hong, X., Wang, Y., Su, Z.: Semi-supervised crowd counting via density agency. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 1416–1426 (2022)
21. Lin, H., Ma, Z., Ji, R., Wang, Y., Hong, X.: Boosting crowd counting via multifaceted attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19628–19637 (2022)
22. Liu, C., Lu, H., Cao, Z., Liu, T.: Point-query quadtree for crowd counting, localization, and more. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1676–1685 (2023)
23. Liu, J., Gao, C., Meng, D., Hauptmann, A.G.: Decidenet: Counting varying density crowds through attention guided detection and density estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5197–5206 (2018)
24. Liu, L., Chen, J., Wu, H., Li, G., Li, C., Lin, L.: 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. pp. 4823–4833 (2021)
25. Liu, L., Qiu, Z., Li, G., Liu, S., Ouyang, W., Lin, L.: Crowd counting with deep structured scale integration network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
26. Liu, L., Wang, H., Li, G., Ouyang, W., Lin, L.: Crowd counting using deep recurrent spatial-aware network. arXiv preprint arXiv:1807.00601 (2018)
27. Liu, Y., Liu, L., Wang, P., Zhang, P., Lei, Y.: Semi-supervised crowd counting via self-training on surrogate tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. pp. 242–259. Springer (2020)
28. Liu, Y., Cao, G., Shi, B., Hu, Y.: Ccanet: A collaborative cross-modal attention network for rgb-d crowd counting. IEEE Transactions on Multimedia (2023)
29. Liu, Z., Wu, W., Tan, Y., Zhang, G.: Rgb-t multi-modal crowd counting based on transformer. The 33rd British Machine Vision Conference 2022 (2022)

30. Ma, J., Zhou, H., Zhao, J., Gao, Y., Jiang, J., Tian, J.: Robust feature matching for remote sensing image registration via locally linear transforming. IEEE Transactions on Geoscience and Remote Sensing **53**(12), 6469–6481 (2015)
31. Ma, Z., Wei, X., Hong, X., Gong, Y.: Bayesian loss for crowd count estimation with point supervision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6142–6151 (2019)
32. Ma, Z., Wei, X., Hong, X., Gong, Y.: Learning scales from points: A scale-aware probabilistic model for crowd counting. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 220–228 (2020)
33. Ma, Z., Wei, X., Hong, X., Lin, H., Qiu, Y., Gong, Y.: Learning to count via unbalanced optimal transport. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2319–2327 (2021)
34. Mo, H., Ren, W., Zhang, X., Yan, F., Zhou, Z., Cao, X., Wu, W.: Attention-guided collaborative counting. IEEE Transactions on Image Processing **31**, 6306–6319 (2022)
35. Mu, B., Shao, F., Xie, Z., Chen, H., Jiang, Q., Ho, Y.S.: Visual prompt multi-branch fusion network for rgb-thermal crowd counting. IEEE Internet of Things Journal (2024)
36. Pan, Y., Zhou, W., Fang, M., Qiang, F.: Graph enhancement and transformer aggregation network for rgb-thermal crowd counting. IEEE Geoscience and Remote Sensing Letters (2024)
37. Pan, Y., Zhou, W., Qian, X., Mao, S., Yang, R., Yu, L.: Cginet: Cross-modality grade interaction network for rgb-t crowd counting. Engineering Applications of Artificial Intelligence **126**, 106885 (2023)
38. Pang, Y., Zhang, L., Zhao, X., Lu, H.: Hierarchical dynamic filtering network for rgb-d salient object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16. pp. 235–252. Springer (2020)
39. Peng, T., Li, Q., Zhu, P.: Rgb-t crowd counting from drone: A benchmark and mmccn network. In: Proceedings of the Asian conference on computer vision (2020)
40. Ren, S., Du, Y., Lv, J., Han, G., He, S.: Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13325–13333 (2021)
41. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. pp. 234–241. Springer International Publishing, Cham (2015)
42. Sam, D.B., Peri, S.V., Sundararaman, M.N., Kamath, A., Babu, R.V.: Locate, size, and count: accurately resolving people in dense crowds via detection. IEEE transactions on pattern analysis and machine intelligence **43**(8), 2739–2751 (2020)
43. Sindagi, V.A., Patel, V.M.: Multi-level bottom-top and top-bottom feature fusion for crowd counting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1002–1012 (2019)
44. Tang, H., Wang, Y., Chau, L.P.: Tafnet: A three-stream adaptive fusion network for rgb-t crowd counting. In: 2022 IEEE International Symposium on Circuits and Systems (ISCAS). pp. 3299–3303. IEEE (2022)
45. Wang, Y., Hou, J., Hou, X., Chau, L.P.: A self-training approach for point-supervised object detection and counting in crowds. IEEE Transactions on Image Processing **30**, 2876–2887 (2021)

46. Wang, Z., Wang, Z., Zheng, Y., Chuang, Y.Y., Satoh, S.: Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 618–626 (2019)
47. Wei, X., Li, D., Hong, X., Ke, W., Gong, Y.: Co-attentive lifting for infrared-visible person re-identification. In: Proceedings of the 28th ACM international conference on multimedia. pp. 1028–1037 (2020)
48. Wu, Z., Liu, L., Zhang, Y., Mao, M., Lin, L., Li, G.: Multimodal crowd counting with mutual attention transformers. In: 2022 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2022)
49. Xie, Z., Shao, F., Chen, G., Chen, H., Jiang, Q., Meng, X., Ho, Y.S.: Cross-modality double bidirectional interaction and fusion network for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **33**(8), 4149–4163 (2023)
50. Xie, Z., Shao, F., Mu, B., Chen, H., Jiang, Q., Lu, C., Ho, Y.S.: Bgdfnet: Bidirectional gated and dynamic fusion network for rgb-t crowd counting in smart city system. IEEE Transactions on Instrumentation and Measurement (2024)
51. Xu, H., Yuan, J., Ma, J.: Murf: Mutually reinforcing multi-modal image registration and fusion. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
52. Yang, X., Zhou, W., Yan, W., Qian, X.: Cagnet: Coordinated attention guidance network for rgb-t crowd counting. Expert Systems with Applications **243**, 122753 (2024)
53. Yang, Y., Li, G., Wu, Z., Su, L., Huang, Q., Sebe, N.: Reverse perspective network for perspective-aware object counting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4374–4383 (2020)
54. Yu, L., Chen, J., Sinha, A., Wang, M., Chen, Y., Berg, T.L., Zhang, N.: Commercemm: Large-scale commerce multimodal representation learning with omni retrieval. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 4433–4442 (2022)
55. Zhang, B., Du, Y., Zhao, Y., Wan, J., Tong, Z.: I-mmccn: Improved mmccn for rgb-t crowd counting of drone images. In: 2021 7th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC). pp. 117–121. IEEE (2021)
56. Zhang, J., Fan, D.P., Dai, Y., Anwar, S., Saleh, F.S., Zhang, T., Barnes, N.: Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8582–8591 (2020)
57. Zhang, Q., Chan, A.B.: Wide-area crowd counting via ground-plane density maps and multi-view fusion cnns. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8297–8306 (2019)
58. Zhang, Y., Choi, S., Hong, S.: Spatio-channel attention blocks for cross-modal crowd counting. In: Proceedings of the Asian Conference on Computer Vision. pp. 90–107 (2022)
59. Zhang, Y., Yan, Y., Lu, Y., Wang, H.: Towards a unified middle modality learning for visible-infrared person re-identification. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 788–796 (2021)
60. Zhao, W., Xie, S., Zhao, F., He, Y., Lu, H.: Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13955–13965 (2023)

61. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5906–5916 (2023)
62. Zhao, Z., Bai, H., Zhu, Y., Zhang, J., Xu, S., Zhang, Y., Zhang, K., Meng, D., Timofte, R., Van Gool, L.: Ddfm: denoising diffusion model for multi-modality image fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8082–8093 (2023)
63. Zhou, W., Pan, Y., Lei, J., Ye, L., Yu, L.: Defnet: Dual-branch enhanced feature fusion network for rgb-t crowd counting. IEEE Transactions on Intelligent Transportation Systems **23**(12), 24540–24549 (2022)
64. Zhou, W., Yang, X., Dong, X., Fang, M., Yan, W., Luo, T.: Mjpnet-s*: Multistyle joint-perception network with knowledge distillation for drone rgb-thermal crowd density estimation in smart cities. IEEE Internet of Things Journal (2024)
65. Zhou, W., Yang, X., Lei, J., Yan, W., Yu, L.: MC^3Net: Multimodality cross-guided compensation coordination network for rgb-t crowd counting. IEEE Transactions on Intelligent Transportation Systems (2023)