Multi-modal Crowd Counting via Modal Emulation

Chenhao Wang[1]
22B903078@stu.hit.edu.cn
Xiaopeng Hong[1,†]
hongxiaopeng@ieee.org
Zhiheng Ma[2]
zh.ma@siat.ac.cn
Yupeng Wei[1]
23s003027@stu.hit.edu.cn
Yabin Wang[3]
iamwangyabin@stu.xjtu.edu.cn
Xiaopeng Fan[1]
fxp@hit.edu.cn

[1] Faculty of Computing, Harbin Institute of Technology, P. R. China
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Science, P. R. China
[3] School of Cyberspace and Engineering, Xi'an Jiaotong University, P. R. China

Abstract

Multi-modal crowd counting is a crucial task that uses multi-modal cues to estimate the number of people in crowded scenes. To overcome the gap between different modalities, we propose a modal emulation-based two-pass multi-modal crowd-counting framework that enables efficient modal emulation, alignment, and fusion. The framework consists of two key components: a *multi-modal inference* pass and a *cross-modal emulation* pass. The former utilizes a hybrid cross-modal attention module to extract global and local information and achieve efficient multi-modal fusion. The latter uses attention prompting to coordinate different modalities and enhance multi-modal alignment. We also introduce a modality alignment module that uses an efficient modal consistency loss to align the outputs of the two passes and bridge the semantic gap between modalities. Extensive experiments on both RGB-Thermal and RGB-Depth counting datasets demonstrate its superior performance compared to previous methods. Code available at https://github.com/Mr-Monday/Multi-modal-Crowd-Counting-via-Modal-Emulation.

1 Introduction

Crowd counting is an essential research topic in the field of machine perception. The goal of this task is to accurately estimate the number of people in an image. Over the past decade, crowd counting has been widely used in various fields [19, 35]. Existing crowd counting methods mainly focus on the visual features of RGB images [12, 15, 16, 18, 22, 23, 24] may have limitations when confronted with intricate environments like occlusions and shadows.

† Corresponding authors.

Recently, multi-modal crowd counting has gained increasing attention for its ability to address the limitations of using only the visible modality. Fusing thermal or depth images with RGB images can significantly improve counting performance, especially in challenging scenes with low light and occlusion. Most of the previous multi-modal crowd counting approaches [9, 44] are based on convolutional structures for multi-modal fusion, and show that integrating thermal or depth images with RGB images improves crowd counting performance. Nevertheless, simple fusion methods [20, 30] are limited in their ability to fully capture the complementarity between modalities. Recent studies have shown the effectiveness of transformer models in multi-modal tasks [21, 29]. For instance, Wu et al. [34] introduce a Mutual Attention Transformer (MAT) that uses a cross-attention mechanism to capture the complementarity of different modalities for crowd counting. However, cross-attention mechanisms in most existing methods may primarily capture multi-modal interactions rather than explore modal alignment, which may limit the full fusion of multi-modal data.

This paper tackles the multi-modal counting problem from a new perspective, arguing that a superior multi-modal feature encoder should be capable of both *fusing* and *emulating* modal features. By transforming the input of one modality into the features of another through simple and efficient operations, we can assume that the encoder can comprehend and align two distinct modalities well enough.

Based on the above analysis, in this paper, we propose a modal emulation-based multi-modal crowd counting approach that leverages a two-pass learning paradigm to perform efficient modal emulation, alignment, and fusion, as illustrated in Figure 1. We conduct extensive experiments on two widely used RGB-T and RGB-D multi-modal crowd-counting benchmarks. The results show that our proposed method outperforms previous methods and demonstrates the effectiveness of our method in leveraging multi-modal information for crowd counting. The technical contributions can be further summarized as follows:

- We propose a two-pass learning paradigm for multi-modal crowd counting. In addition to the normal *multi-modal inference* pass, we propose a *cross-modal emulation* pass that encourages the model to coordinate different modalities. This two-pass paradigm makes our approach distinct from traditional methods.
- We propose a modality alignment loss to align the outputs of the two passes and bridge the semantic gap between different modalities.
- We develop a hybrid cross-modal attention module, which consists of a straight attention mechanism that focuses more on global attention and a modulated attention mechanism that emphasizes local attention, to enhance multi-modal fusion power.

2 Related Work

2.1 Multi-modal Crowd Counting

Currently, the crowd counting task has been extensively studied [9, 11, 13, 14, 17, 23, 31, 40]. To enhance the counting accuracy, several works have introduced information from other modalities, such as thermal or depth [2, 7, 25, 26, 34, 37, 39, 44]. Lian et al. [9, 10] introduce a large-scale RGB-D crowd counting dataset and leverage a depth prior and a density map to improve the head/non-head classification in the detection network. Zhang et al. [41] adopt a CSCA method to effectively capture and integrate information from different

modalities. Zhou et al. [43] propose a dual-branch enhanced feature fusion network to fuse RGB-thermal features. Liu et al. [20] and Tang et al. [30] introduce a three-stream network for multi-modal fusion. Zhou et al. [44] propose a multimodality cross-guided compensation coordination network to predict crowd density maps by complementing different modules. However, these multi-modal approaches do not fully explore the modality alignment issue.

2.2 Transformer for Multi-modal

Many transformer-based methods were proposed for multi-modal tasks [36, 42]. Vilbert [21] and LXMERT [29] use the cross-attention mechanism to learn vision-and-language connections. Zhu et al. [45] propose a multi-modal feature pyramid transformer that fuses different modalities by intra-modal and inter-modal feature pyramid transformer. Zhang et al. [38] design a cross-modal feature rectification module to calibrate bi-modal features and a two-stage feature fusion module to enhance the information interaction.

2.3 Prompting Learning

Recently, prompting learning has achieved great success in computer vision tasks [6, 32, 33]. Zhu et al. [6] develop a visual prompt multi-modal tracking framework for various downstream multi-modal tracking tasks by learning modal-relevant prompts. Li et al. [8] utilize a prompting method to extract fusion representations between different modalities.

It is crucial to emphasize that none of the previously mentioned methods encompass the idea of cross-modal emulation, which constitutes the central focus of our paper. Consequently, the foundational motivation, the implementation, and the pertinent loss functions employed to fine-tune the prompts diverge markedly from those outlined in previous work.

3 Method

Figure 1 presents an overview of the framework, which mainly consists of a Multi-modal Inference (MMI) pass and a Cross-modal Emulation (CME) pass. The two passes share most of the network structure and weights. And we place the proposed Hybrid Cross-modal Attention (HCMA) module behind each block of the dual-channel VGG19-like network [20]. Specifically, given an RGB image and a thermal image, to maintain the specific information of each modality, we feed them into the first three blocks of VGG19 [27] φ_r and φ_t to extract modality-specific features of individual modality $F_r, F_t \in R^{C\times H\times W}$, where C, W, and H are the channel, width, and height, respectively. And then, the $2D$ feature F_r, F_t are embedded and flattened to a sequence of patch embeddings $X_r, X_t \in R^{L\times D}$, where L is the number of patches, and D is the patch dimension. To fully fuse the information of the two modalities, we introduce the MMI pass which consists of the HCMA module into the adjacent block of the VGG19 to capture global-local complementarity information. Meanwhile, the CME pass can modulate F_r features into pseudo-thermal features $\bar{F}_t$ to enhance modality alignment. Similarly, the F_t features can also be modulated into pseudo-RGB features $\bar{F}_r$. Next, the features produced by both passes are fed into the modality alignment loss, which aims to bridge the semantic gap across different modalities. Afterward, the output features of the MMI pass are linearly combined using a weighted sum and fed into a regression head to generate a prediction for the final high-fidelity crowd density map $\hat{\mathcal{D}}$. Finally, we combine the Bayesian Loss [22] to constrain the training of the overall model.

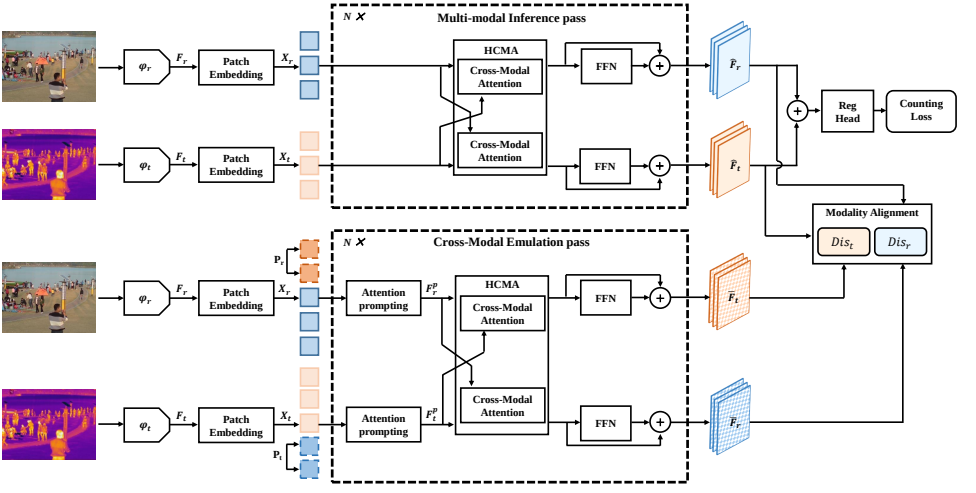

Figure 1: **Illustration of the proposed framework**. Specifically, our framework consists of two passes: the Multi-modal Inference (MMI) pass and the Cross-modal Emulation (CME) pass. The MMI pass uses a hybrid cross-modal attention module to fuse global and local modalities. The CME pass shares the structure and weights with the MMI but emulates features of one modality into another, *i.e.*, $F_r \rightarrow \bar{F}_t$ and $F_t \rightarrow \bar{F}_r$, using an additional attention prompting module. The process of emulation fosters the coordination of different modalities. Moreover, a loss function for modality alignment is employed to bridge the semantic gap that exists between these modalities.

3.1 Multi-modal Inference

In the Multi-modal Inference pass, we design the HCMA module, which comprises two types of attention mechanisms: straight cross-modal attention and modulated cross-modal attention, as shown in Figure 2.

Straight Cross-modal Attention

To capture long-range contextual information by fusing global information from both modalities, we introduce the Straight Cross-modal Attention (SCMA) mechanism based on Multi-head Attention (MHA) [1], as shown in Figure 2 (a). Specifically, different modal patch embeddings X_r and X_t are linearly projected to produce their queries, keys, and values, respectively, which are denoted as Q_r, K_r, V_r and Q_t, K_t, V_t. Then, we perform straight cross-modal attention, which can be calculated as follows:

$$H_r = s\left(\frac{Q_r K_t^T}{\sqrt{d}}\right)V_t, \quad H_t = s\left(\frac{Q_t K_r^T}{\sqrt{d}}\right)V_r \tag{1}$$

where H terms the output of the attention head, s terms the softmax function, and $\frac{1}{\sqrt{d}}$ is the scaling factor based on the query/key dimension d. Finally, the outputs of each head are concatenated and fed to a series of operations including dropouts and residual concatenation, and then reshaped into 2D features to obtain the fused global features F_r^g and F_t^g, where $F_r^g, F_t^g \in R^{C^{'} \times H \times W}$.

Modulated Cross-modal Attention

We introduce the Modulated Cross-modal Attention (MCMA) mechanism to fuse local details in different modalities and obtain modulated complementary features, as shown in Figure 2 (b). First, the patch embeddings X_r and X_t are reshaped into 2D feature maps $\tilde{F}_r$ and

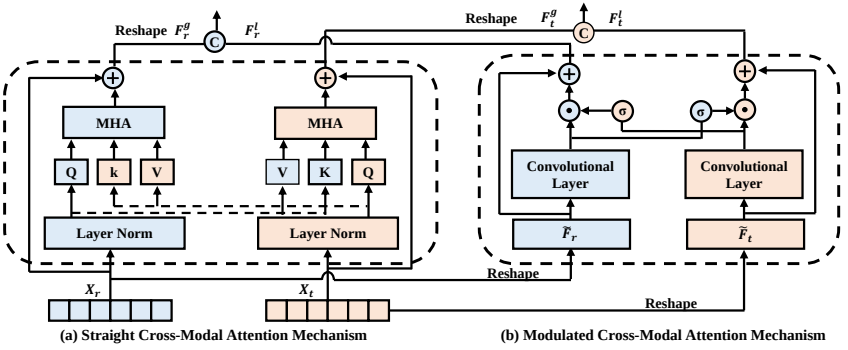

Figure 2: Architecture of the Hybrid Cross-Modal Attention Module. (a) Straight Cross-modal Attention is used for global multi-modal fusion. (b) Modulated Cross-modal Attention is used to fuse local details, where the $\odot$, σ and $\copyright$ denote Hadamard product, Sigmoid function, and Concatenation operation, respectively.

$\tilde{F}_t \in R^{C^{'} \times H \times W}$, respectively. These feature maps are then fed to the modulated cross-modal attention mechanism to capture local complementary information:

$$F_r^l = \phi_r(\tilde{F}_r) \odot \sigma(\phi_t(\tilde{F}_t)) + \tilde{F}_r, F_t^l = \phi_t(\tilde{F}_t) \odot \sigma(\phi_r(\tilde{F}_r)) + \tilde{F}_t, \tag{2}$$

where ϕ, $\odot$, and σ denote a two-layer convolutional layer, Hadamard product, and Sigmoid function, respectively.

To combine complementary global and local features, we simply concatenate them together along the channel dimension and introduce a convolutional layer with 1×1 convolution kernel to reduce the concatenated features to the $C^{'}$ dimension. Then, we feed them into a two-layer feed-forward network (f) to obtain the fused feature:

$$\hat{F}_r = f_r([F_r^g, F_r^l]),\ \hat{F}_t = f_t([F_t^g, F_t^l]), \tag{3}$$

where the [,] defines the concatenation operation.

We further diversify the contributions of different modality features and assign weights to each modality based on their importance or relevance. The two modalities are then fused as a weighted sum and input to the regression head γ:

$$\hat{\mathcal{D}} = \gamma(\alpha \hat{F}_r + \beta \hat{F}_t), \tag{4}$$

where $\alpha, \beta \in [0, 1]$ are learnable parameters.

3.2 Cross-modal Emulation

As we argued in the introduction, modal emulation, by which the feature of one modality is converted into the one of another modality, is an important means of allowing models to fully comprehend and align different modalities. Motivated by this idea, we propose a cross-modal emulation pass to realize cross-modal modulation between the RGB and thermal features.

We design the CME based on attention prompting [32] which inserts prompts to the multi-head self-attention layer. We split the prompts P_r of RGB modal features into sub-prompts P_r^k, P_r^v with the same sequence length, and prepend them to the key K_r^p and value V_r^p vectors while keeping query Q_r^p vectors. Q_r^p, K_r^p and V_r^p vectors are generated by the RGB features X_r. Then, we can define the function of attention prompting as:

$$F_r^p = s\left(\frac{Q_r^p[P_r^k, K_r^p]^T}{\sqrt{d}}\right)[P_r^v, V_r^p], \tag{5}$$

Similarly, we can also get the attention prompting of thermal features:

$$F_t^p = s\left(\frac{Q_t^p[P_t^k, K_t^p]^T}{\sqrt{d}}\right)[P_t^v, V_t^p], \tag{6}$$

Finally, through the CME pass ψ, the RGB features can be transformed to resemble the thermal features, which are called the *pseudo* thermal features $\bar{F}_t$. Similarly, we can also obtain *pseudo* RGB features $\bar{F}_r$:

$$[\bar{F}_t, \bar{F}_r] = \psi(F_r^p, F_t^p). \tag{7}$$

By converting one modality to another, the CME pass effectively modulates the feature representations of different modalities and enhance their alignment to better fuse information from different modalities. Notably, the CME pass is executed only in the training phase. Therefore, it does not increase the model size and extra overhead in the testing phase.

3.3 Overall loss function

Modality Alignment Loss. To align the outputs of the two passes to bridge the semantic gap between modalities, we use a Consistency Loss:

$$\mathcal{L}_{CL} = \sum_{i=1}^{M}\left(Dis\left(\hat{F}_r^i, \bar{F}_r^i\right) + Dis\left(\hat{F}_t^i, \bar{F}_t^i\right)\right), \tag{8}$$

where M is the number of training samples, the $Dis(\cdot)$ is the distance metric and we simply use the Euclidean distance in our experiments.

Counting Loss. We adopt the Bayesian Loss (BL) [22] for crowd counting:

$$\mathcal{L}_{BL} = \sum_{i=1}^{M}\left|1 - \langle\hat{\mathcal{D}}, \frac{\mathcal{N}(\mathcal{D}_i, \sigma^2\mathbf{I}_{2\times 2})}{\sum_{j=1}^{M}\mathcal{N}(\mathcal{D}_j, \sigma^2\mathbf{I}_{2\times 2})}\rangle\right|, \tag{9}$$

$\mathcal{N}(\cdot,\cdot)$ is a Normal distribution centered at the ith head point $\mathcal{D}_i$ with standard deviation σ.

Finally, the overall loss is

$$\mathcal{L} = \mathcal{L}_{BL} + \mathcal{L}_{CL}. \tag{10}$$

4 Experiments

We conduct experiments on two challenging datasets. **RGBT-CC** contains 2,030 RGB-T image pairs, each with the size of 640×480. We follow [20] and use 1,030, 200, and 800 pairs for training, validation, and testing, respectively. **ShanghaiTechRGBD** is a large-scale RGB-depth crowd counting dataset of 2,193 images [9]. Each sample includes both an RGB image and a corresponding depth map. 1,193 samples are assigned to the training set and the remaining ones for testing.

Implementation details. We implement our model on the Pytorch framework with an NVIDIA RTX 3090 GPU. The CME and MMI passes share most of the network structure

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

Method	Venue	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
UCNet [39]	CVPR2020	33.96	42.42	53.06	65.07	56.31
HDFNet [26]	ECCV2020	22.36	27.79	33.68	42.48	33.93
MVMS [40]	CVPR2019	19.97	25.10	31.02	38.91	33.97
BBSNet [8]	ECCV2020	19.56	25.07	31.25	39.24	32.48
CmCaF [9]	TII2022	15.87	19.92	24.65	28.01	29.31
IADM [20]	CVPR2021	15.61	19.95	24.69	32.89	28.18
CSCA [41]	ACCV2022	14.32	18.91	23.81	32.47	26.01
TAFNet [30]	ISCAS2022	12.38	16.98	21.86	30.19	22.45
BL+MAT [32]	ICME2022	12.35	16.29	20.81	29.09	22.53
DEFNet [43]	TITS2022	11.90	16.08	20.19	27.27	21.09
MC^{3}Net [44]	TITS2023	11.47	15.06	19.40	27.95	20.59
Ours-small		11.68	16.12	20.58	28.42	**19.06**
Ours-base		**11.23**	**14.98**	**18.91**	**26.54**	**19.85**

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

Method	Venue	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
UCNet [39]	CVPR2020	10.81	15.24	22.04	32.98	15.70
DetNet [24]	CVPR2018	9.74	-	-	-	13.14
HDFNet [26]	ECCV2020	8.32	13.93	17.97	22.62	13.01
CL [4]	ECCV2018	7.32	-	-	-	10.48
BBSNet [8]	ECCV2020	6.26	8.53	11.80	16.46	9.26
BL+MAT [32]	ICME2022	5.39	6.73	8.98	13.66	7.77
RDNet [3]	CVPR2019	4.96	-	-	-	7.22
CSCA [41]	ACCV2022	4.39	6.47	8.82	11.76	6.39
IADM [20]	CVPR2021	4.38	5.95	8.02	11.02	7.06
DPDNet [10]	TPAMI2021	4.23	5.67	7.04	9.64	6.75
PESSNet [37]	TITS2023	4.10	-	-	-	6.02
Ours-small		4.73	6.48	9.74	16.44	6.88
Ours-base		**3.80**	**5.36**	7.71	12.57	**5.52**

and weights. CME is implemented by incorporating an attention prompting module, with five learnable prompts, before the first HCMA block of MMI. In Figure 1, the number of HCMA blocks N is set to 3. In our implementation, patch dimension D is set to 768. The SCMA mechanism is set to 1 layer with 4 heads, the patch size of the first HCMA block is set to 2, while the last two blocks were set to 1. This model has 160M parameters, which is considered as our *base* model. We also design a *small* model with 82M parameters, where the patch dimension D in the three HCAM blocks are set to 256, 512, and 512, respectively. In the training phase, we adopt Adam as the optimizer, the learning rate is set to 0.00001. We set the batch size to 32 on the RGBT-CC dataset and batch size to 1 on the RGB-D dataset, respectively. Normal data augmentation is applied to the input images, including random crop and flip. The input images are randomly cropped to 256×256 for RGBT-CC dataset and 1024×1024 for RGB-D dataset. The max training epoch is set to 1500. The Root Mean Square Error (RMSE) [28] and the Grid Average Mean Absolute Error (GAME) [3] are adopted to evaluate the performance.

4.1 Comparison with State-of-the-Art Methods

On the RGBT-CC dataset, the performance of all compared methods is shown in Table 1. It could be found that the proposed method achieves better performance on evaluation metrics. For example, compared to MC^{3}Net, our model significantly improves counting performance, reducing GAME(0) to 11.23 and RMSE to 19.85, respectively. Our model stands out due to

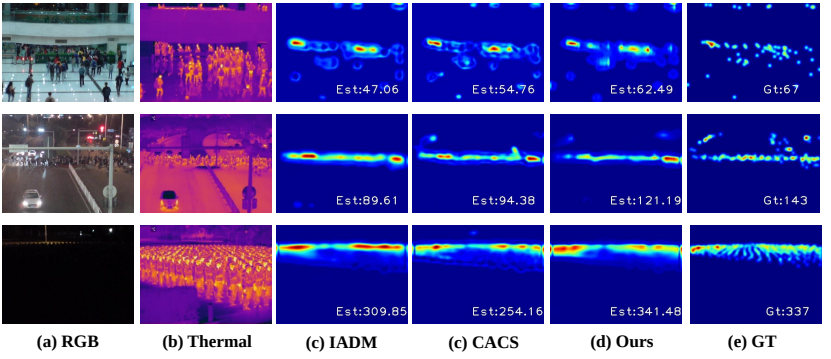

Figure 3: Visualization results for generating crowd density maps with different models.

its specific modal emulation ability to enhance the alignment and fusion of representations between RGB images and thermal images. We compare the visualization results of generating crowd density maps using different models to further validate our method in Figure 3.

Moreover, we evaluate the ShanghaiTech-RGBD dataset. The results in Table 2 demonstrate that our method outperforms previous advanced models in terms of the main evaluation metrics (*i.e.*, GAME(0), and RMSE). Specifically, our method achieves the lowest GAME(0) of 3.80 and RMSE of 5.52. The superior performance of our method demonstrates its generality and effectiveness in addressing multi-modal crowd counting tasks.

4.2 Detailed Discussion

Ablation Study. We conducted a comprehensive study to evaluate the contribution of each component to the overall performance of the framework, as shown in Table 3. We start with the baseline, a two-stream expansion of the BL approach [22]. Then, by incorporating the HCMA module consisting of SCMA and MCMA into the multi-modal inference process, our method consistently reduces the counting errors, specifically, by 6.25 and 9.74 in terms of GAME(0) and RMSE when compared with the baselines. The results demonstrate that the HCMA module contributes to enhancing the fusion of local information and global representation between the two modalities. Furthermore, when using the CME pass with the attention prompting module, the best performance is achieved (*i.e.*, GAME(0) is 11.23, and RMSE is 19.85). The success of the CME pass can be attributed to its ability to effectively coordinate information from different modalities. By aligning and fusing the modalities, the overall model performance is improved.

Table 3: Ablation study on RGBT-CC dataset.

MMI		CME		GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
SCMA	MCMA	AP	IP					
×	×	×	-	18.68	22.91	28.06	35.87	31.42
√	×	×	-	15.37	20.63	25.29	33.01	27.60
√	√	×	-	12.43	16.58	21.26	28.77	21.68
√	√	-	√	11.48	15.95	20.56	28.57	19.57
√	√	√	-	**11.23**	**14.98**	**18.91**	**26.54**	**19.85**

For the CME pass, we conduct experiments with different prompting techniques, namely attention prompting (AP) and input prompting (IP) [6], in Table 3. Both prompting tech-

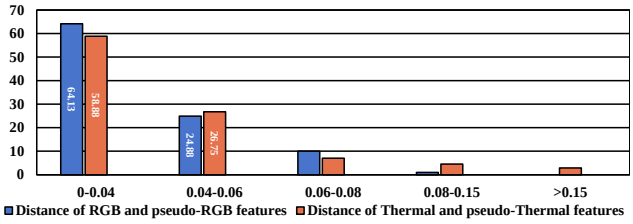

Figure 4: Distribution of the relative $L1$ distances between the real and pseudo features.

niques work for multi-modal crowd counting. However, we notice a trend where AP outperforms IP. This can be attributed to the fact that IP requires discarding the prompts in each HCMA block, which may potentially reduce its effectiveness.

We conduct experiments to compare HCMA with the vanilla cross-attention module in transformers as shown in Table 4. Specifically, we replace the HCMA module with the cross-attention module and ensure that both models have a similar number of parameters (*i.e.* 151M for this model and 160M for ours). The results indicate that the improvement achieved by our proposed method is not primarily attributable to the simple introduction of the cross-attention module or additional parameters, but instead benefited from the carefully-designed HCMA and attention prompting modules.

Table 4: Comparison to the vanilla cross-attention (VCA) module.

Model	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
VCA	15.11	19.66	24.29	31.92	27.94
Ours	**11.23**	**14.98**	**18.91**	**26.54**	**19.85**

Effectiveness of CME pass. We tabulate the distribution of the relative $L1$ distances between the thermal and pseudo-thermal features (as well as the RGB and pseudo-RGB features). The results are reported in Figure 4, where the horizontal axis indicates the ratio of the $L1$ distance to the average $L1$ norm of the real features, and the vertical axis represents the percentage of the test samples among 800. For thermal and pseudo-thermal features, 58.88% of the samples have the relative $L1$ distances below 0.04, and 92.63% of the samples are below 0.08. Similarly, for RGB and pseudo-RGB features, 64.13% of samples are below 0.04, and 99.00% of samples are below 0.08. This suggests that most of the pseudo samples only have a slight difference from the targeted samples. These results suggest that the CME pass well coordinates the two modalities.

Table 5: Impact of direct use of pseudo-features.

	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
W. PF	12.21	16.39	21.12	28.70	19.98
W/O PF	**11.23**	**14.98**	**18.91**	**26.54**	**19.85**

Nevertheless, an additional inquiry arises: can these *pseudo-modal features be directly employed* in generating the final density map output? We conduct an evaluation as shown in Table 5 to address it. When we concatenate the pseudo features and real features to feed the regression head, it can be found that the model performance is degraded. The rationale behind this phenomenon may lie in the fact that despite the pseudo-features closely approximating the features of the target modality, there still exist certain discrepancies. The direct amalgamation of one feature and its inferior counterpart does not lead to improvement but instead may result in a performance decline. In addition, if we use pseudo features for

generating the density map, it means that the data needs to go through both the CME and MMI passes during testing. Although the MMI and CME passes share parameters, this will add extra overhead.

5 Conclusion

We propose an effective emulation-based two-pass framework for multi-modal crowd counting. Our framework leverages a multi-modal inference pass that includes a hybrid cross-modal attention module, which fuses global and local complementary information from different modalities, as well as a cross-modal emulation pass that encourages the model to coordinate different modalities through attention prompting. Additionally, we introduce a modal alignment module to bridge the semantic gap between modalities. Through quantitative and qualitative experiments on RGB-T and RGB-D datasets, we demonstrate that our approach achieves competitive performance and high effectiveness for crowd counting. Our framework has promising potential to be applied to a variety of multi-modal tasks, which warrants further investigation in future research.

Acknowledgement

This work was funded in part by the National Key R&D Program of China (2021YFF0900500), the National Natural Science Foundation of China (62076195, 62376070, 62206271, 62441202, U22B2035), as well as the Fundamental Research Funds for the Central Universities (AUGA-5710011522).

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

[2] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, pages 275–292. Springer, 2020.

[3] Ricardo Guerrero-Gómez-Olmedo, Beatriz Torre-Jiménez, Roberto López-Sastre, Saturnino Maldonado-Bascón, and Daniel Onoro-Rubio. 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*, pages 423–431. Springer, 2015.

[4] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–546, 2018.

[5] Zhu Jiawen, lai Simiao, Chen Xin, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[6] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14943–14952, 2023.

[7] He Li, Shihui Zhang, and Weihang Kong. Rgb-d crowd counting with cross-modal cycle-attention fusion and fine-coarse supervision. *IEEE Transactions on Industrial Informatics*, 19(1):306–316, 2022.

[8] Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2604–2613, 2023.

[9] Dongze Lian, Jing Li, Jia Zheng, Weixin Luo, and Shenghua Gao. 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*, pages 1821–1830, 2019.

[10] Dongze Lian, Xianing Chen, Jing Li, Weixin Luo, and Shenghua Gao. Locating and counting heads in crowds with a depth prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9056–9072, 2021.

[11] Hui Lin, Zhiheng Ma, Xiaopeng Hong, Yaowei Wang, and Zhou Su. Semi-supervised crowd counting via density agency. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1416–1426, 2022.

[12] Hui Lin, Zhiheng Ma, Rongrong Ji, Yaowei Wang, and Xiaopeng Hong. Boosting crowd counting via multifaceted attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19628–19637, 2022.

[13] Hui Lin, Zhiheng Ma, Xiaopeng Hong, Qinnan Shangguan, and Deyu Meng. Gramformer: Learning crowd counting via graph-modulated transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3395–3403, 2024.

[14] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2018.

[15] Lingbo Liu, Hongjun Wang, Guanbin Li, Wanli Ouyang, and Liang Lin. Crowd counting using deep recurrent spatial-aware network. *Proceedings of the International Joint Conference on Artificial Intelligence*, 2018.

[16] Lingbo Liu, Ruimao Zhang, Jiefeng Peng, Guanbin Li, Bowen Du, and Liang Lin. Attentive crowd flow machines. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1553–1561, 2018.

[17] Lingbo Liu, Zhilin Qiu, Guanbin Li, Shufan Liu, Wanli Ouyang, and Liang Lin. Crowd counting with deep structured scale integration network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1774–1783, 2019.

[18] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Tianshui Chen, Guanbin Li, and Liang Lin. Efficient crowd counting via structured knowledge transfer. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2645–2654, 2020.

[19] Lingbo Liu, Jiajie Zhen, Guanbin Li, Geng Zhan, Zhaocheng He, Bowen Du, and Liang Lin. Dynamic spatial-temporal representation learning for traffic flow prediction. *IEEE Transactions on Intelligent Transportation Systems*, 22(11):7169–7183, 2020.

[20] Lingbo Liu, Jiaqi Chen, Hefeng Wu, Guanbin Li, Chenglong Li, and Liang 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*, pages 4823–4833, 2021.

[21] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[22] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6142–6151, 2019.

[23] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Learning scales from points: A scale-aware probabilistic model for crowd counting. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 220–228, 2020.

[24] Zhiheng Ma, Xiaopeng Hong, Xing Wei, Yunfeng Qiu, and Yihong Gong. Towards a universal model for cross-dataset crowd counting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3205–3214, 2021.

[25] Haoliang Meng, Xiaopeng Hong, Chenhao Wang, Miao Shang, and Wangmeng Zuo. Multi-modal crowd counting via a broker modality. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[26] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. 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*, pages 235–252. Springer, 2020.

[27] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Vishwanath A Sindagi and Vishal M Patel. Multi-level bottom-top and top-bottom feature fusion for crowd counting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1002–1012, 2019.

[29] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[30] Haihan Tang, Yi Wang, and Lap-Pui Chau. Tafnet: A three-stream adaptive fusion network for rgb-t crowd counting. In *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 3299–3303. IEEE, 2022.

[31] Yabin Wang, Zhiheng Ma, Xing Wei, Shuai Zheng, Yaowei Wang, and Xiaopeng Hong. Eccnas: Efficient crowd counting neural architecture search. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(1s): 1–19, 2022.

[32] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 631–648. Springer, 2022.

[33] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[34] Zhengtao Wu, Lingbo Liu, Yang Zhang, Mingzhi Mao, Liang Lin, and Guanbin Li. Multimodal crowd counting with mutual attention transformers. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2022.

[35] Feng Xiong, Xingjian Shi, and Dit-Yan Yeung. Spatiotemporal modeling for crowd counting in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 5151–5159, 2017.

[36] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18268–18278, 2023.

[37] Jun Yi, Yiran Pang, Wei Zhou, Meng Zhao, and Fujian Zheng. A perspective-embedded scale-selection network for crowd counting in public transportation. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[38] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[39] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. 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*, pages 8582–8591, 2020.

[40] Qi Zhang and Antoni B Chan. 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*, pages 8297–8306, 2019.

[41] Youjia Zhang, Soyun Choi, and Sungeun Hong. Spatio-channel attention blocks for cross-modal crowd counting. In *Proceedings of the Asian Conference on Computer Vision*, pages 90–107, 2022.

[42] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6068–6077, 2023.

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

[44] Wujie Zhou, Xun Yang, Jingsheng Lei, Weiqing Yan, and Lu Yu. Mc^{3}net: Multimodality cross-guided compensation coordination network for rgb-t crowd counting. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[45] Yaohui Zhu, Xiaoyu Sun, Miao Wang, and Hua Huang. Multi-modal feature pyramid transformer for rgb-infrared object detection. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

Response to the Reviewers

Response to Reviewer #smWP:
Q1, Q2, Q3, Q4, and Q5: a few typos and re-write lines 128 to 135.
A: Thanks. We have fixed the typos and rewritten lines 128 to 135.
Q6: Compared with the method of Liu et al.
A: Thanks. The method proposed by Liu et al., which is based on CNNs to fuse multi-modal information, fails to model global cross-modal relationships. However, our method not only uses the MMI pass to fuse both global and local multi-modal information but also employs the CME pass to fully align different modalities.
Response to Reviewer #N5v7:
Q1-Q7:Some formatting issues. A: Thanks. We have fixed these issues.
Response to Reviewer #Htx6:
Q1: Motivation for the proposed method.
Thanks. In multi-modal tasks, both modal fusion and modal alignment are essential. Unlike most multi-modal counting methods that directly align and fuse different modal information in a common feature space using operations like cross-attention, we propose a prompting-based cross-modal emulation and a corresponding two-pass learning paradigm to imbue the network backbone with the ability to "translate" from one modality to another. Additionally, CNN-based fusion methods often struggle to capture long-range contextual information, which limits the effectiveness of cross-modal fusion. To address this, we employ the hybrid cross-modal attention module to effectively fuse local and global information.
Q2: The fusion design differs from the LoGoNet.
Thanks. For the local fusion mechanism, LoGoNet uses a local fusion module with grid point dynamic fusion to dynamically integrate point cloud features with RGB features. In contrast, our method employs modulated cross-modal attention to fuse the local fusion information, which operates as follows $F_r = \phi_r(F_r) \odot \sigma(\phi_t(F_t)) + F_r$ and $F_r = \phi_t(F_t) \odot \sigma(\phi_t(F_t)) + F_r$. For the global fusion mechanism, our approach utilizes straight cross-modal attention, focusing on the global fusion of both the thermal (depth) modality and the RGB modality.
Q3: Minor Issues:
Thanks. We have fixed these issues.