The 14th International Conference on Multimedia Retrieval
ICMR 2024
June 10-14, 2024
Dusit Thani Laguna Phuket, Thailand

南京大學
NANJING UNIVERSITY

MAGUS
MediA recoGnition and UnderStanding

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

Yaqun Fang, Yi Shi, Jia Bei*, Tongwei Ren
State Key Laboratory for Novel Software Technology, Nanjing University

MAGUS
MediA recoGnition and UnderStanding

南京大學
NANJING UNIVERSITY

CONTENTS

NANJING UNIVERSITY

Homepage:

https://magus.ink

Author Information

Yaqun Fang

fangyq@smail.nju.edu.cn

Yi Shi

yishi@smail.nju.edu.cn

Bei Jia

beijia@nju.edu.cn

Tongwei Ren

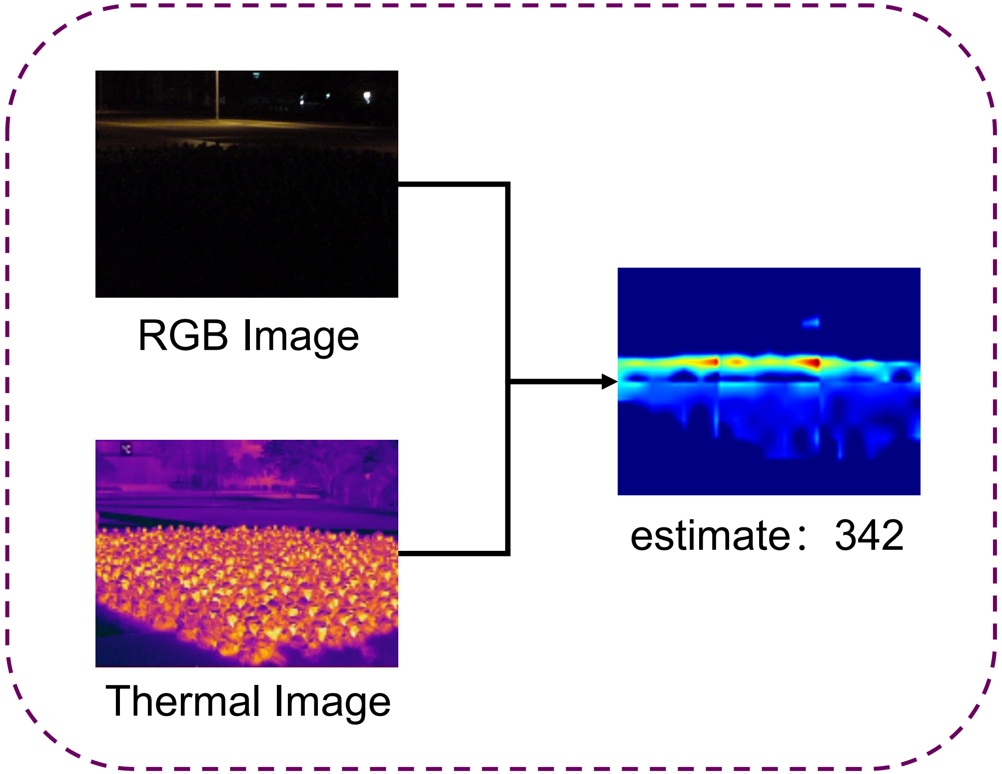
rentw@nju.edu.cn

Introduction & Related Works

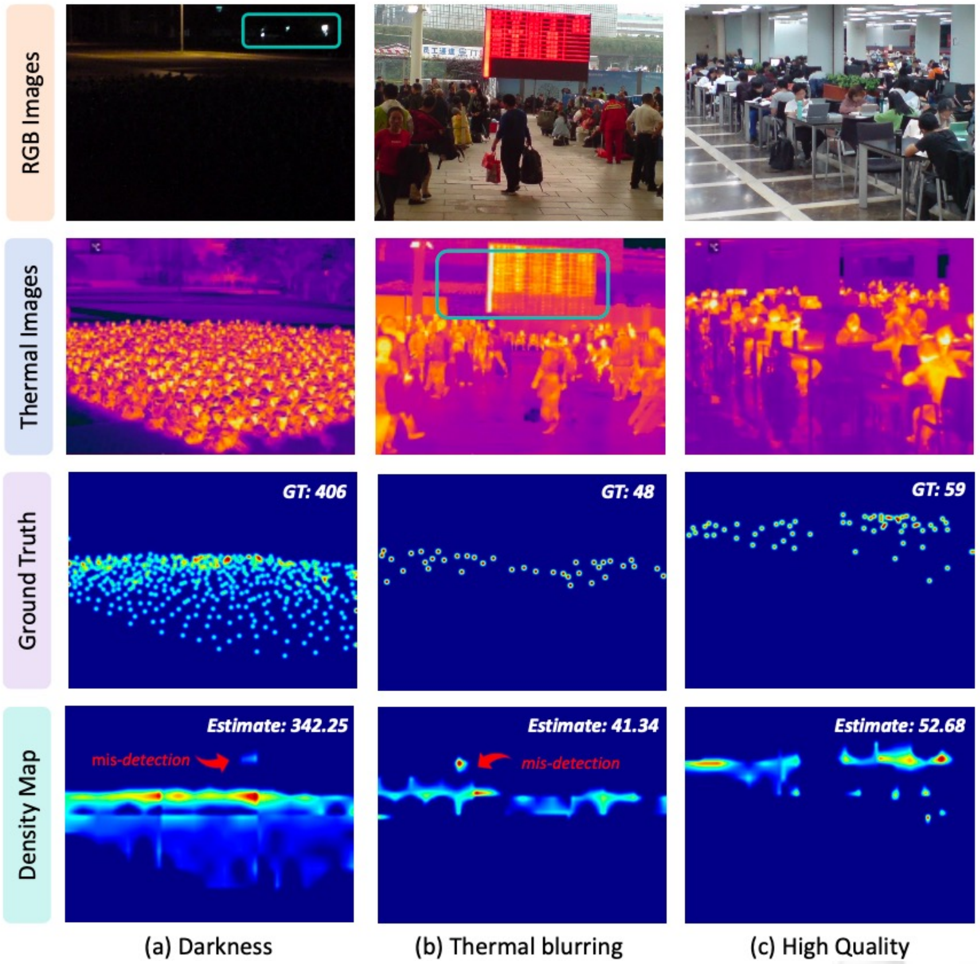
- Problem
- Related Work
- Contribution

Introduction

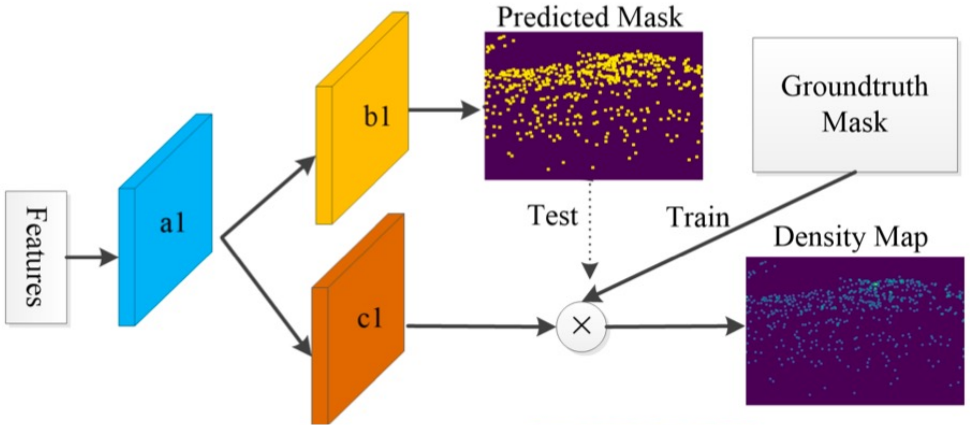
Problem

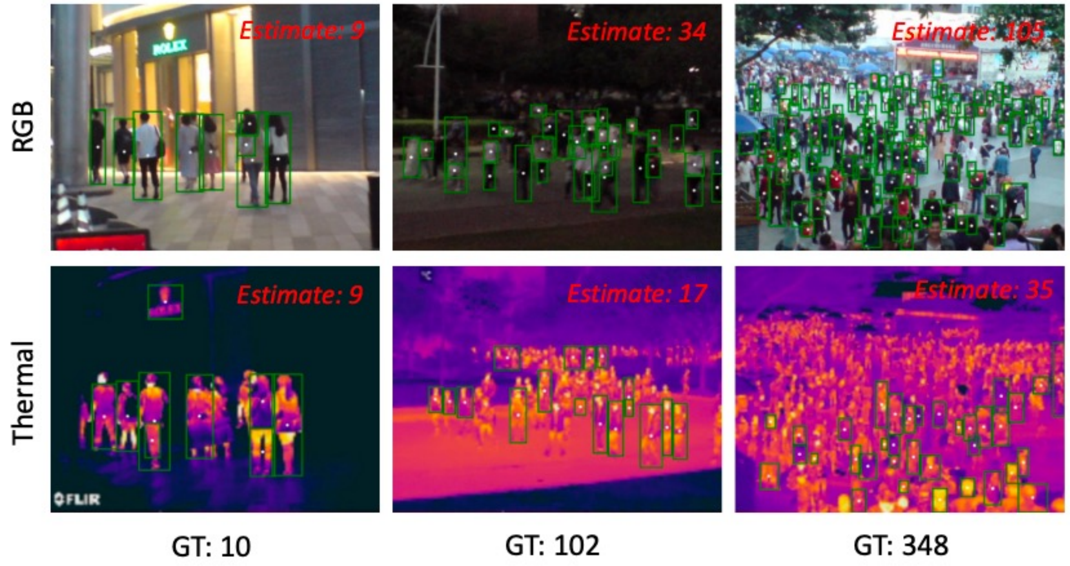


RGB-Thermal Crowd Counting



Related Work





- Need segmentation labels
- Face difficulty in transferring between datasets with different crowd sizes

As the level of congestion increases, the incidence of missed detections becomes more pronounced.

Introduction

Contribution

In this paper, we propose a novel method which utilize SAM to generate semantic map, and guide the interaction between modalities using semantic features.

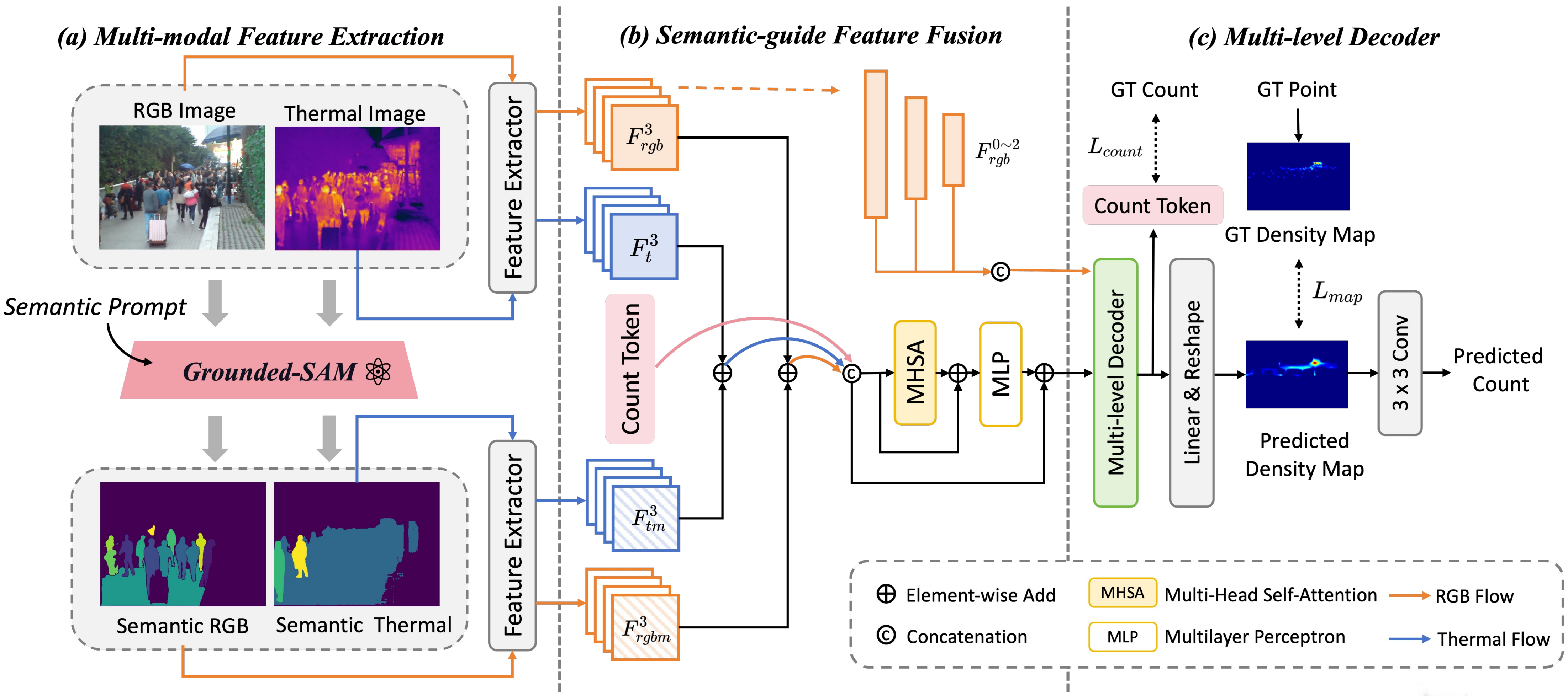
(1) Our research is the inaugural effort to integrate SAM into RGB-T crowd counting. Leveraging SAM, we innovatively generate semantic maps in both RGB and thermal modalities.

(2) We employ semantic features to guide and enhance the representation of modal features within both RGB and thermal modalities. This approach significantly boosts the effectiveness of cross-modal feature fusion, leading to enhanced performance in crowd counting tasks.

Method & Experiment

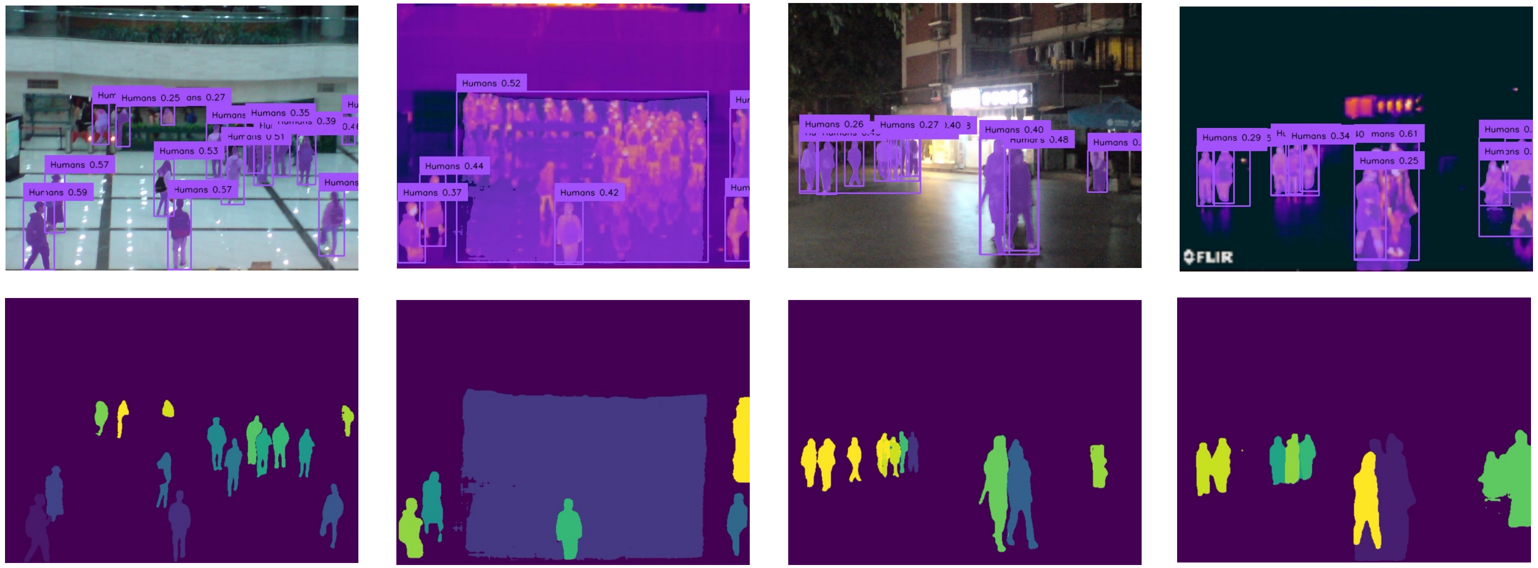
- Method
- Experiment

Overview



(a) Multi-modal Feature Extraction
(b) Semantic-guide Feature Fusion
(c) Multi-level Decoder
RGB Image
Thermal Image
Feature Extractor
Semantic Prompt
Grounded-SAM
Semantic RGB
Semantic Thermal
Feature Extractor
F^3_{rgb}
F^3_t
Count Token
F^3_{tm}
F^3_{rgbm}
$F^{0\sim2}_{rgb}$
MHSA
MLP
GT Count
GT Point
L_{count}
Count Token
GT Density Map
L_{map}
Multi-level Decoder
Linear & Reshape
3 x 3 Conv
Predicted Count
Predicted Density Map
Element-wise Add
Concatenation
MHSA Multi-Head Self-Attention
MLP Multilayer Perceptron
RGB Flow
Thermal Flow

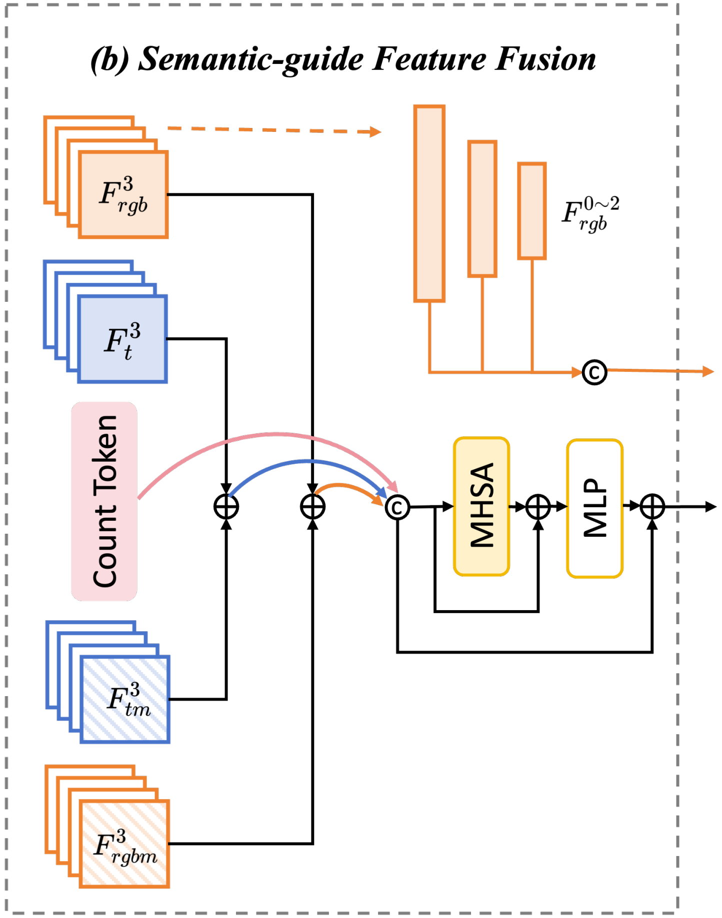
Multi-modal Feature Extraction



RGB/T images → G-SAM → Semantic maps → PVT → Feature maps

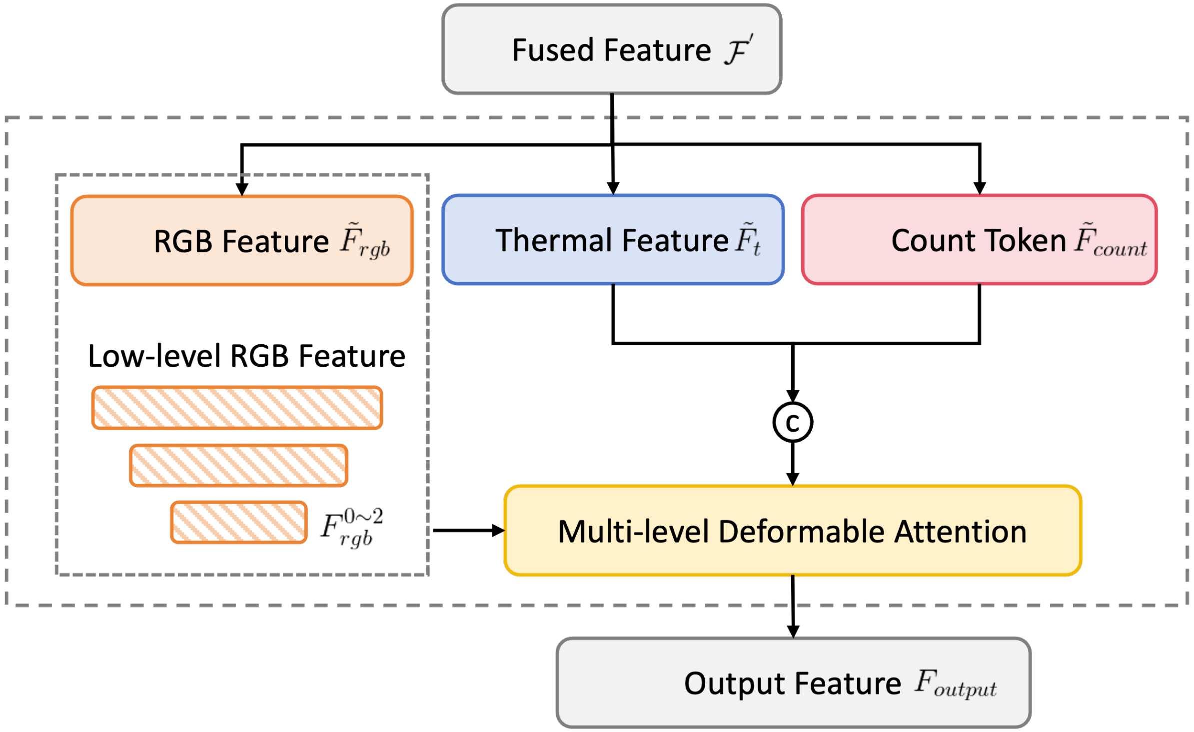
RGB/T images → PVT → Feature maps

Semantic-guide Feature Fusion



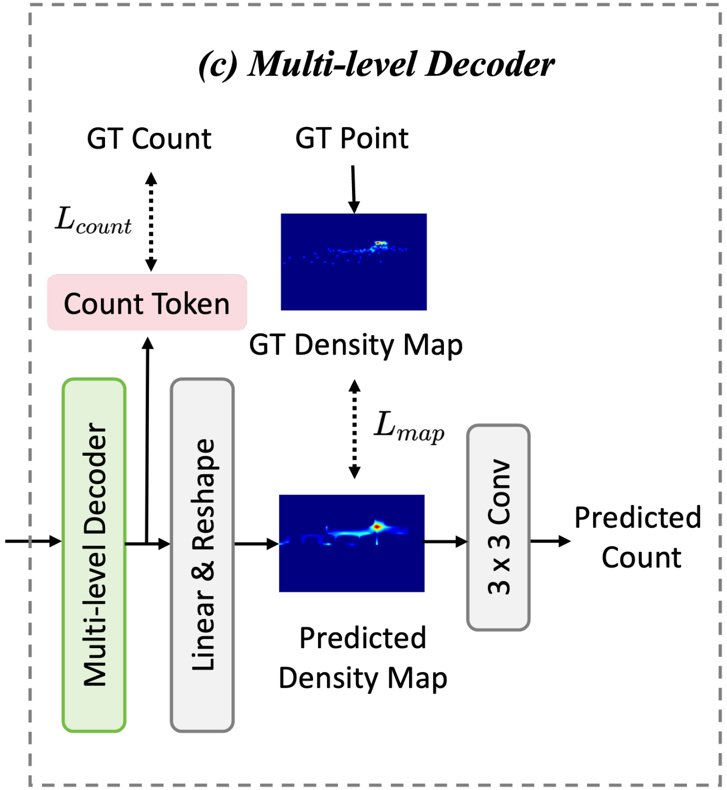
- Add semantic features to modal features
- Concat with count token
- Use Multi-head self-attention to enhance features
- Get the fused features

Multi-level Decoder



- Split fused feature into RGB feature, thermal feature and count token
- Concat thermal feature with count token
- Use Multi-level Deformable Attention to integrate low-level feature
- Get the output feature

Loss Function

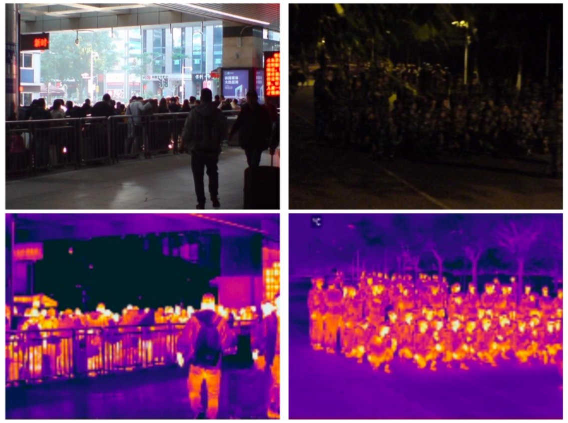


- Output feature from decoder is splited to count token and density map
- The overall loss is composed of two parts: the loss of the density map and the loss of counting

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

Experiment

Dataset and Metrics



RGBT-CC Dataset

- Dataset
  - RGBT-CC Dataset (2030 pairs)
- Metrics
  - Grid Average Mean Absolute Error (GAME)
  - Root Mean Square Error (RMSE)

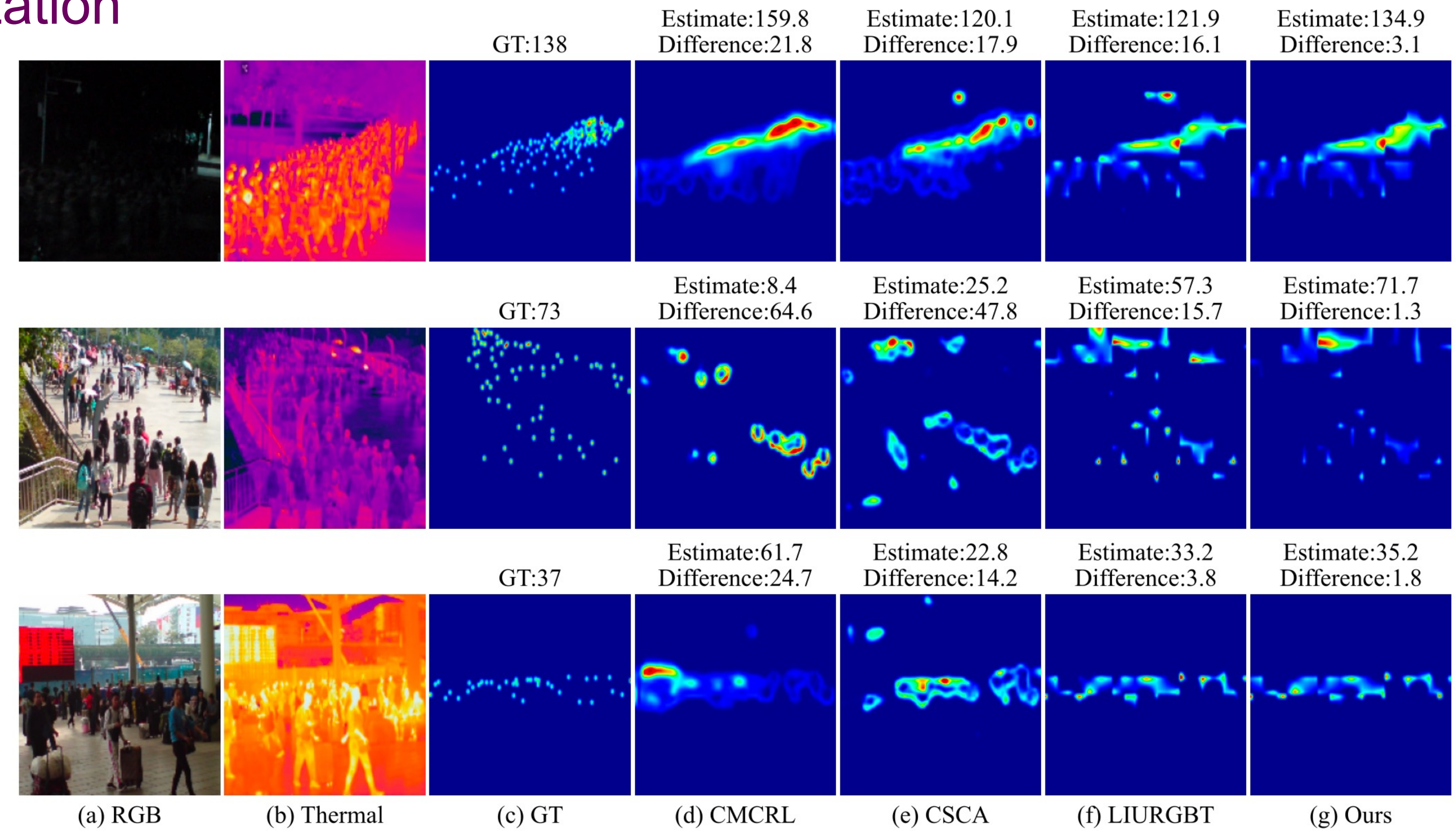
$$GAME(l) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{4^l}|\hat{P}_i^j - P_i^j|, \qquad RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_i - P_i)^2},$$

Comparison with State-of-the-Arts

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

Experiment

Visualization



Experiment

Ablation Studies

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

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

Limitation & Conclusion

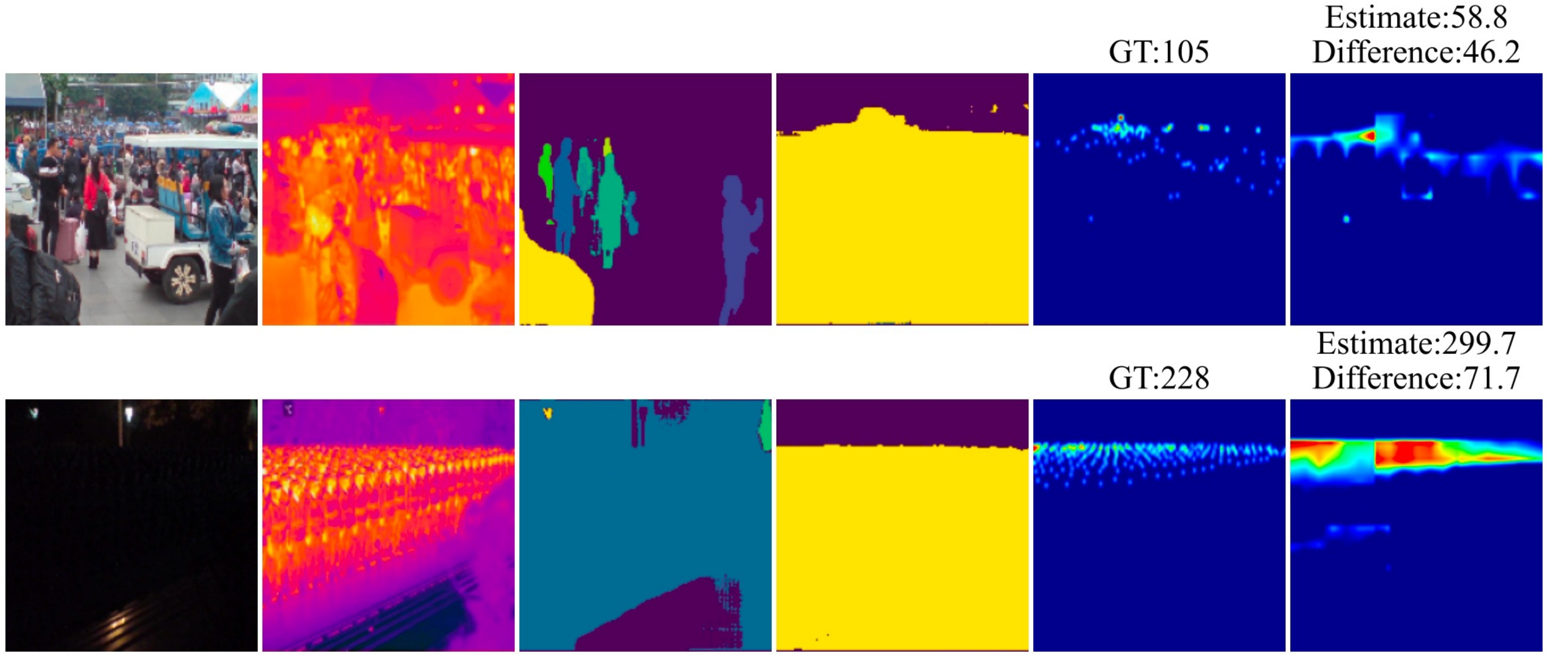
- Limitation
- Conclusion

NANJING UNIVERSITY

MAGUS
MediA recoGnition and UnderStanding

Limitation

Failure cases



(a) RGB (b) Thermal (c) RGB mask (d) Thermal mask (e) GT (f) Ours

- Challenge in scenarios contain excessive crowd
- Constrain by the quality of the original image

Conclusion

Conclusion

In this paper, we proposed a novel semantic-guided RGB-T crowd counting method, which generates semantic maps of crowd on both RGB and thermal modalities by leveraging SAM.

Our method explored the utilization of semantic features to guide and enhance the representation of modal features through the semantic-guided fusion module. With semantic information, the false-positive counting in background is reduced, while the counting accuracy in crowd regions is improved.

The experiments on the RGBT-CC dataset demonstrate that our proposed method outperforms the state-of-the-art methods.

Thank You!