Multi-modal Crowd Counting via Modal Emulation

Chenhao Wang, Xiaopeng Hong*, Zhiheng Ma, Yupeng Wei ,
Yabin Wang and Xiaopeng Fan
22b903078@stu.hit.edu.cn hongxiaopeng@ieee.org

Motivation

Problem

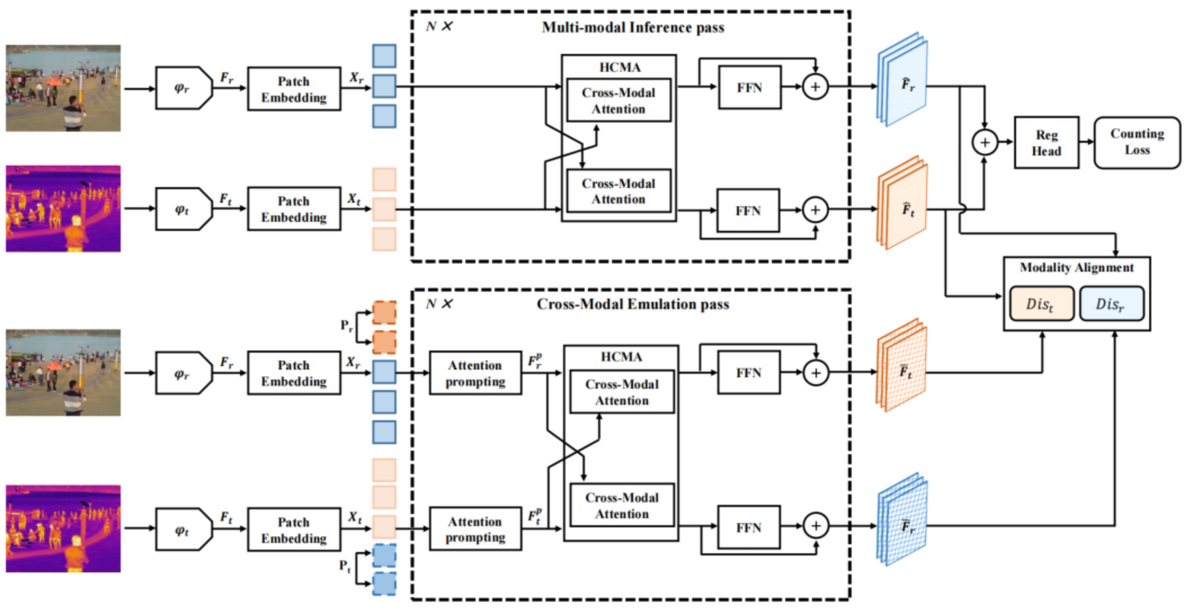
- Most multi-modal crowd counting methods may not fully explore modal alignment, which may limit data fusion.
- Some fusion methods are limited in their ability to capture the global-local information between modalities.

Basic Idea

- A robust multi-modal feature encoder should be able to fuse and emulate modal features simultaneously.
- By converting inputs from one modality to features of another modality, the encoder can be considered to be able to fully understand and align two different modalities.

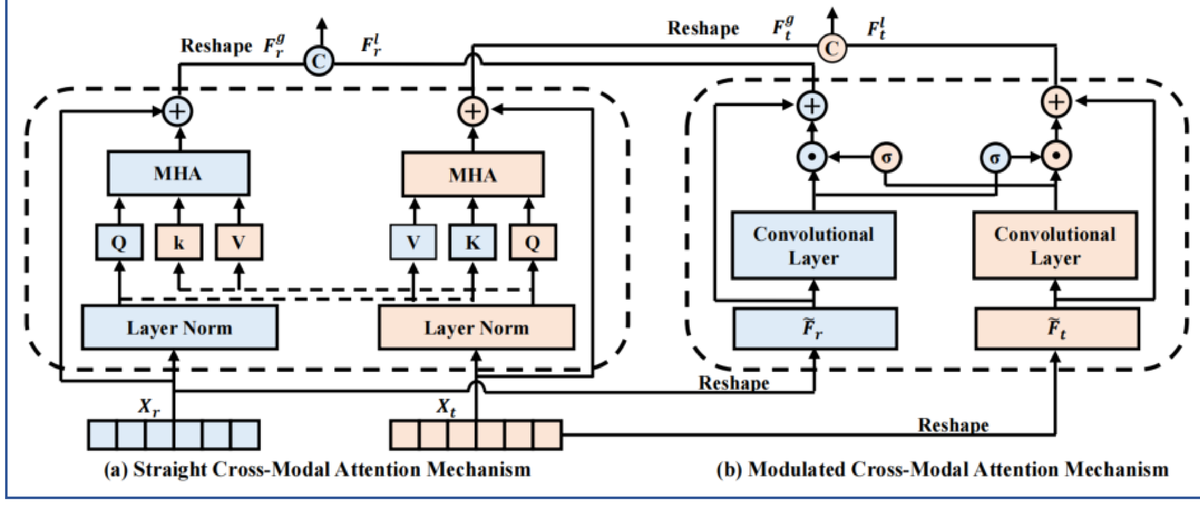
Approach Overview

- A two-pass learning paradigm for multi-modal crowd counting:
  - ➢ The multi-modal inference pass is designed to fuse global-local modal information.
  - ➢ The cross-modal emulation pass aims to coordinate different modalities.



Multi-modal Inference Pass (MMI)

- The hybrid cross-modal attention module contains two types of attention mechanisms:
  - ➢ Straight cross-modal attention mechanism focuses more on global attention.
  - ➢ Modulated cross-modal attention mechanism emphasizes local attention.



(a) Straight Cross-Modal Attention Mechanism (b) Modulated Cross-Modal Attention Mechanism

Cross-modal Emulation Pass (CME)

- The CME based on attention prompting inserts prompts into the multi-head self-attention layer. The function of attention prompting is as follows:

$$F_r^P = s(\frac{Q_r^P[P_r^k, K_r^P]^T}{\sqrt{d}})[P_r^v, V_r^P], \quad F_t^P = s(\frac{Q_t^P[P_t^k, K_t^P]^T}{\sqrt{d}})[P_t^v, V_t^P]$$

- Through the CME pass, the RGB / thermal features can be transformed to resemble the pseudo thermal / RGB features $\bar{F}_t$ / $\bar{F}_r$:

$$[\bar{F}_t, \bar{F}_r] = CME(F_r^P, F_t^P)$$

Experiments

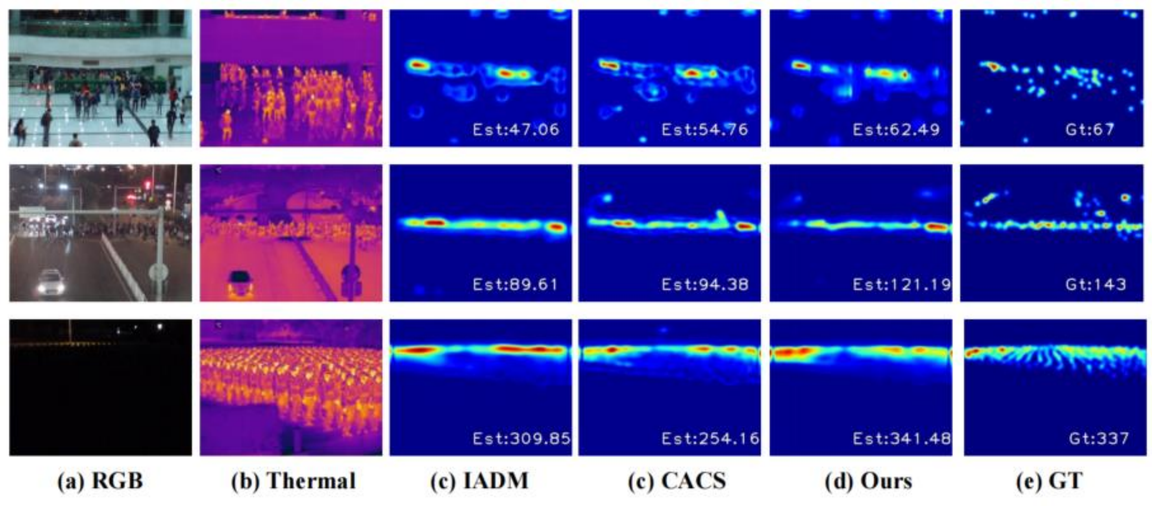
1. Comparison with existing methods on RGBT-CC dataset

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

2. Comparison with existing methods on RGBD dataset

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

3. Visualization results for generating crowd density maps with different models on RGBT-CC dataset



(a) RGB (b) Thermal (c) IADM (c) CACS (d) Ours (e) GT

Conclusions

- A two-pass framework for multimodal crowd counting is proposed to enhance the modal alignment and fusion.
- Experiments on RGB-T and RGB-D datasets demonstrate competitive performance.