Cross-Modal Collaborative Representation Learning and a Large-Scale RGBT Benchmark for Crowd Counting

Lingbo Liu[1], Jiaqi Chen[1], Hefeng Wu[1], Guanbin Li[1,2], Chenglong Li[3], Liang Lin[1,4]
[1] School of Computer Science and Engineering, Sun Yat-sen University, China
[2] Pazhou Lab, Guangzhou, China [3] Anhui University, China [4] DarkMatter AI Research, China
{liulingbo918, wuhefeng}@gmail.com, chenjq87@mail2.sysu.edu.cn, liguanbin@mail.sysu.edu.cn,
lcl1314@foxmail.com, linliang@ieee.org

1. More Results of Unimodal Data

In the main text, we have reported the results of different methods that take multimodal data as input. In this supplementary file, we also report the unimodal performance of different backbone networks (e.g., MCNN [5], SANet [1], CSRNet [2] and BL [4]).

As shown in Table 1, when only taking RGB images as input, these backbone networks perform poorly on the proposed RGBT-CC benchmark, because they fail to recognize people in poor illumination conditions, such as backlight and night. The performance is greatly improved when using thermal images. Nevertheless, both the RGB results and thermal results are worse than multimodal results. In particular, all backbone networks achieve the best performance when capturing the RGB-thermal complementarities with the proposed Information Aggregation-Distribution Module (IADM). Moreover, we also perform unimodal experiments on the ShanghaiTechRGBD dataset [3]. As shown in Table 2, the unimodal results of all backbone networks are consistently worse than their multimodal results. These experiments demonstrate the effectiveness of multimodal data for crowd counting.

2. Representation Visualization

In this supplementary file, we also visualize and compare the generated features before and after applying the proposed IADM. Here we take BL [4] as the backbone network. As shown in Fig. 1 and Fig. 2, after applying IADM, both modality-specific and modality-shared representations have been enhanced in various illumination conditions. This demonstrates that our method can indeed capture the complementary information of multimodal data effectively to facilitate the task of crowd counting.

References

[1] Xinkun Cao, Zhipeng Wang, Yanyun Zhao, and Fei Su. Scale aggregation network for accurate and efficient crowd counting. In *ECCV*, pages 734–750, 2018. 1, 2

[2] Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *CVPR*, pages 1091–1100, 2018. 1, 2

[3] 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 *CVPR*, pages 1821–1830, 2019. 1

[4] Zhiheng Ma, Xing Wei, Xiaopeng Hong, and Yihong Gong. Bayesian loss for crowd count estimation with point supervision. In *ICCV*, pages 6142–6151, 2019. 1, 2

[5] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. In *CVPR*, pages 589–597, 2016. 1, 2

Table 1. Performance of unimodal data and multimodal data on the RGBT-CC benchmark.

Backbone	Input	Feature Learning	GAME(0) ↓	GAME(1) ↓	GAME(2) ↓	GAME(3) ↓	RMSE ↓
MCNN [5]	RGB	-	36.83	43.12	49.85	58.60	71.16
	T	-	22.92	26.65	31.33	37.58	38.92
	RGBT	Early Fusion	21.89	25.70	30.22	37.19	37.44
		IADM	**19.77**	**23.80**	**28.58**	**35.11**	**30.34**
SANet [1]	RGB	-	35.97	41.45	46.75	54.89	70.52
	T	-	22.89	25.83	29.48	36.02	42.33
	RGBT	Early Fusion	21.99	24.76	28.52	34.25	41.60
		IADM	**18.18**	**21.84**	**26.27**	**32.95**	**33.72**
CSRNet [2]	RGB	-	33.94	40.76	47.31	57.20	69.59
	T	-	21.64	26.22	31.65	38.66	37.38
	RGBT	Early Fusion	20.40	23.58	28.03	35.51	35.26
		IADM	**17.94**	**21.44**	**26.17**	**33.33**	**30.91**
BL [4]	RGB	-	33.32	39.19	44.58	54.11	67.50
	T	-	19.93	23.31	27.32	34.64	34.08
	RGBT	Early Fusion	18.70	22.55	26.83	34.62	32.67
		IADM	**15.61**	**19.95**	**24.69**	**32.89**	**28.18**

Table 2. Performance of unimodal data and multimodal data on the ShanghaiTechRGBD benchmark.

Backbone	Input	Feature Learning	GAME(0) ↓	GAME(1) ↓	GAME(2) ↓	GAME(3) ↓	RMSE ↓
MCNN [5]	RGB	-	10.76	13.81	19.02	25.15	14.66
	D	-	28.36	42.95	53.41	64.92	38.74
	RGBD	Early Fusion	11.12	14.53	18.68	24.49	16.49
		IADM	**9.61**	**11.89**	**15.44**	**20.69**	**14.52**
BL [4]	RGB	-	8.83	11.67	15.85	22.85	12.96
	D	-	26.19	30.04	34.58	41.56	37.23
	RGBD	Early Fusion	8.94	11.57	15.68	22.49	12.49
		IADM	**7.13**	**9.28**	**13.00**	**19.53**	**10.27**
SANet [1]	RGB	-	6.89	8.79	11.89	16.48	9.98
	D	-	25.62	30.68	37.03	44.31	35.94
	RGBD	Early Fusion	5.74	7.84	10.47	14.30	8.66
		IADM	**4.71**	**6.49**	**9.02**	**12.41**	**7.35**
CSRNet [2]	RGB	-	4.96	7.09	9.97	13.55	7.44
	D	-	28.53	55.46	67.99	76.41	39.06
	RGBD	Early Fusion	4.92	6.78	9.47	13.06	7.41
		IADM	**4.38**	**5.95**	**8.02**	**11.02**	**7.06**

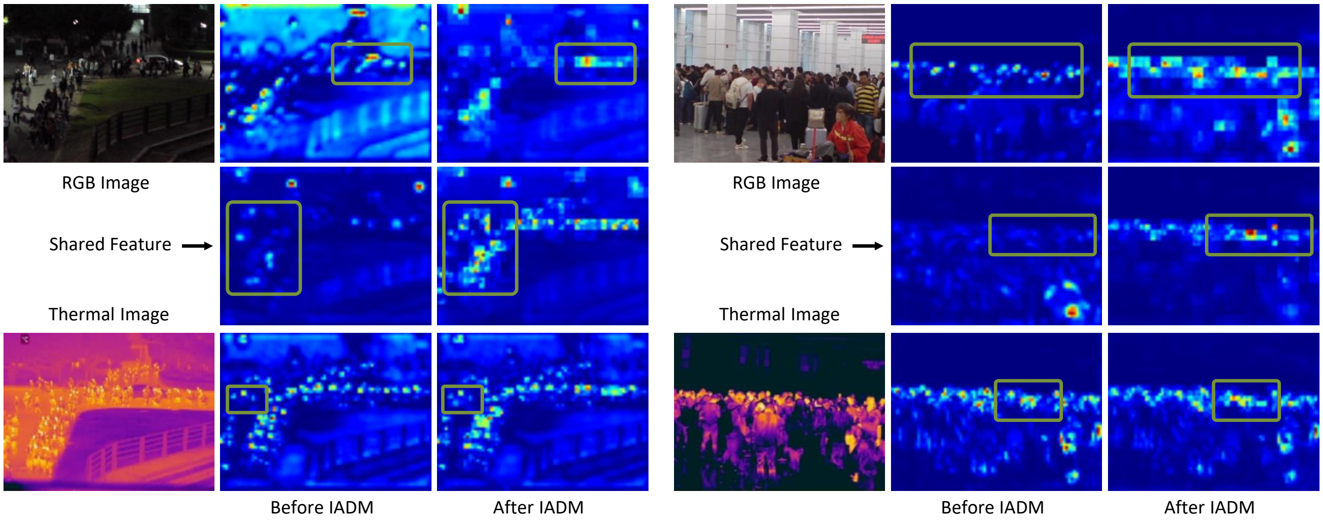

Figure 1. Visualization of the **Conv3_3** features before and after IADM. The first row is the features of the input RGB images, while the third row is the features of the input thermal images. The middle row shows the shared features. We can observe that all modality-specific and modality-shared representations have been enhanced after the proposed IADM. (*Best viewed in color.*)

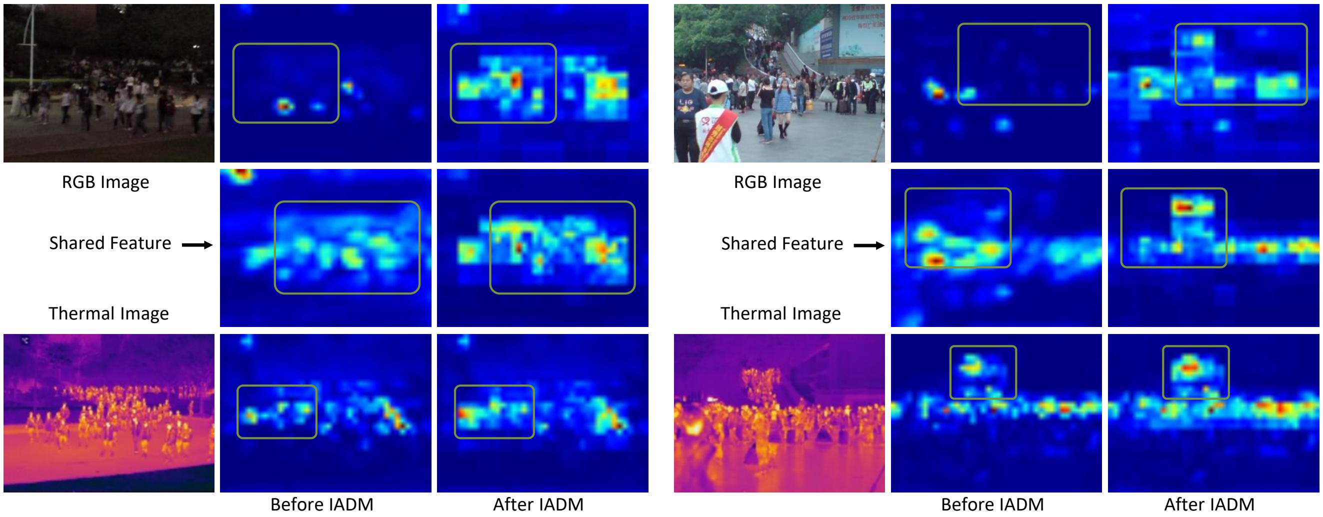

Figure 2. Visualization of the **Conv4_3** features before and after IADM. The first row is the features of the input RGB images, while the third row is the features of the input thermal images. The middle row shows the shared features. We can observe that all modality-specific and modality-shared representations have been enhanced after the proposed IADM. (*Best viewed in color.*)