Multi-modal Crowd Counting via Modal Emulation Supplementary

BMVC 2024 Submission 115

Table 1: Performance of different inputs and illumination conditions on RGBT-CC.

Illumination	Input data	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
ALL	RGB	32.75	41.16	49.64	60.00	68.39
	T	20.31	24.97	29.18	36.57	29.96
	RGB-T	11.23	14.98	18.91	26.54	19.85
Brightness	RGB	20.52	28.54	37.14	48.44	40.44
	T	23.85	28.09	32.31	39.70	34.06
	RGB-T	12.66	15.98	20.29	28.56	22.39
Darkness	RGB	45.36	54.16	62.52	71.91	88.39
	T	16.67	21.75	25.95	33.35	25.05
	RGB-T	9.76	13.94	17.50	24.46	16.84

Effectiveness of Multi-modal Fusion. We investigate the effectiveness of multi-modal fusion, as shown in Table 1. On one hand, we observe a decrease in counting accuracy when using only RGB images as input, with GAME(0) for 32.75 and RMSE for 68.39. This decrease is due to the sensitivity of RGB images to changes in illumination conditions. On the other hand, utilizing only thermal images as input improves model performance. This highlights the effectiveness of thermal images. When considering both RGB and thermal images simultaneously, our method has a significant performance improvement (*i.e.* GAME(0) is 11.23 and RMSE is 19.85). Experimental results show that the fusion of RGB and thermal images outperforms unimodal methods in various lighting conditions. By leveraging the characteristics of thermal and RGB modalities, our modal emulation-based method can effectively fuse the strengths of both modalities.