Supplementary Material of
Free Lunch Enhancements for Multi-modal Crowd Counting

Haoliang Meng[1], Xiaopeng Hong[1,2]*, Zhengqin Lai[1,2] and Miao Shang[1]
[1]Harbin Institute of Technology, [2]Pengcheng Laboratory
menghaoliang2002@163.com, hongxiaopeng@hit.edu.cn, {quenlenu,miaos0522}@gmail.com

1. Additional Experiments

In this section, we conduct some extra experiments to demonstrate the effectiveness of our approach.

Stage	Cosine Sim. ↑	MSE ↓	PSNR (dB) ↑
w/o PPCA, before fine-tune	0.2446	1.0262	20.70
with PPCA, before fine-tune	**0.4067**	**0.3967**	**23.47**
w/o PPCA, after fine-tune	0.4186	0.1404	23.81
with PPCA, after fine-tune	**0.4807**	**0.0480**	**26.93**

Table 1. The similarity between visual and thermal features of the same scene on RGBT-CC.

1.1. The impact of PPCA on cross-modal feature similarity

We evaluate the similarity between visual and thermal features of the same scene extracted by the backbone network to demonstrate the effectiveness of PPCA. Tab. 1 shows the average similarity on RGBT-CC [3]. We evaluate feature similarity through cosine similarity [8], mean square error (MSE) [1], and peak signal noise ratio (PSNR) [2]. As the table indicates, backbones pre-trained on general-purpose single-modal databases extract features with a lower similarity, which means that they capture less common information between modalities of the same scene. After PPCA, the similarity between cross-modal features improves, facilitating cross-modal learning. Similarly, after fine-tuning for multi-modal crowd counting, the model underwent PPCA capture more common information between modalities as well. In general, PPCA improves the feature similarity between visual and thermal images of the same scene, thus enhancing the model's ability to capture shared information across modalities.

*Corresponding author

1.2. Performance on different multi-modal crowd counting methods

Our proposed approach is plug-and-play and compatible with existing multi-modal crowd counting methods. Tab. 2 shows the performance of our approach combined with existing open-source methods in Table 1 of the main paper on RGBT-CC dataset [3]. As the table indicates, our approach consistently improves the model's performance on all five metrics, which validates its effectiveness and compatibility.

References

[1] Tianfeng Chai, Roland R Draxler, et al. Root mean square error (rmse) or mean absolute error (mae). *Geoscientific model development discussions*, 7(1):1525–1534, 2014. 1

[2] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 1

[3] 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. 1, 2

[4] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[5] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 2

[6] Haoliang Meng, Xiaopeng Hong, Chenhao Wang, Miao Shang, and Wangmeng Zuo. Multi-modal crowd counting via a broker modality. *arXiv preprint arXiv:2407.07518*, 2024. 2

[7] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[8] Peipei Xia, Li Zhang, and Fanzhang Li. Learning similarity with cosine similarity ensemble. *Information sciences*, 307: 39–52, 2015. 1

Method	Venue	Backbone	GAME(0)	GAME(1)	GAME(2)	GAME(3)	RMSE
IADM [3]	CVPR 2021	VGG-19 [7]	15.61	19.95	24.69	32.89	28.18
Ours(IADM)			**12.02**	**15.86**	**20.13**	**26.47**	**22.48**
$\mathbf{MC^3Net}$ [9]	TITS 2023	ConvNeXt-S [5]	11.47	15.06	19.40	27.95	20.59
Ours($\mathbf{MC^3Net}$)			**10.82**	**13.93**	**18.30**	**26.01**	**19.91**
BM [6]	ECCV 2024	Swin-T [4]	10.24	13.34	17.19	23.06	18.34
Ours(BM)			**9.57**	**12.62**	**16.27**	**21.85**	**17.05**

Table 2. Performances of our approach on different multi-modal crowd counting methods on RGBT-CC.

[9] Wujie Zhou, Xun Yang, Jingsheng Lei, Weiqing Yan, and Lu Yu. $\mathrm{MC^3Net}$: Multimodality cross-guided compensation coordination network for rgb-t crowd counting. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 2