Appendices

Junjie Guo[1], Chenqiang Gao[2*], Fangcen Liu[1], Deyu Meng[3], and Xinbo Gao[1]

[1] School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China.
{gjj893866738, liufc67}@gmail.com, gaoxb@cqupt.edu.cn
[2] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, Guangdong 518107, China. gaochq6@mail.sysu.edu.cn
[3] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, Shanxi, 710049, China, and School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi, China. dymeng@mail.xjtu.edu.cn

1 Loss Function

The training loss of our model follows the DETR-like detectors, which is defined as:

$$\mathcal{L} = \mathcal{L}_{cls} + \mathcal{L}_{box} + \mathcal{L}_{dn}, \tag{1}$$

where $\mathcal{L}_{cls}$ is the IoU-aware classification loss following RT-DETR [3], $\mathcal{L}_{box}$ is composed of L1 loss and generalized IoU loss for bounding box regression and $\mathcal{L}_{dn}$ is the loss for denoising training [4]. In addition, we also calculate the loss of each decoder layer as the auxiliary optimization loss.

2 Limitation

Our method is effective in addressing common misalignment in the majority of cases. However, it may not handle extreme misalignment situations well, as the object features of the modality will be lost when objects exceed the range of the 4D reference point.

Table A: Comparison of parameters and computation.

Model	Arch	Para	GFLOPs
Yolov7	Single	37.2M	105.2
Deformable-DETR(4 scale)	Single	40.1M	196.5
DINO(4 scale)	Single	47.5M	279.0
CFT	Multi	206.3M	224.6
ICAFusion	Multi	120.3M	192.6
Ours	Multi	**77.5M**	**244.8**

* Corresponding author.

Table B: Comparisons on the M^{3}FD Dataset (Our submission partition)

Model	Data Type	mAP50	mAP75	mAP
Yolov7	MetaFusion [5]	66.4	-	40.9
Yolov7	CDDFuse [6]	67.5	-	42.7
Deformable-DETR	MetaFusion [5]	57.6	36.2	33.5
Deformable-DETR	CDDFuse [6]	60.2	34.6	34.2
DINO	MetaFusion [5]	71.2	47.3	45.1
DINO	CDDFuse [6]	72.7	47.9	46.3
Ours	IR+RGB	**80.2**	**56.0**	**52.9**

Table C: Comparisons on the M^{3}FD Dataset (EAEFNet partition).

Model	Data Type	mAP50	mAP75	mAP
EAEFNet [2]	IR+RGB	80.10	-	-
SA-CBAM [1]	IR+RGB	81.46	-	-
Ours	IR+RGB	**91.6**	**67.8**	**62.8**

3 Model complexity

Tab. A shows that our method is competitive on both parameters and computation(FLOPs). Even compared with single modality DETR methods, the complexity increase of our method is not very obvious.

4 More Comparasons

We additionally compare with four new methods: CDDFuse (CVPR2023) [6], Meta-Fusion(CVPR2023) [5], EAEFNet(RAL 2023) [2] and SA-CBAM(WACV 2024) [1], where the former two adopt the dataset partition of the paper submission, while the rest are the same as the EAEFNet for fairness. Both Tab. B and C show the superiority of our method.

References

1. Deevi, S.A., Lee, C., Gan, L., Nagesh, S., Pandey, G., Chung, S.J.: Rgb-x object detection via scene-specific fusion modules. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 7366–7375 (2024)
2. Liang, M., Hu, J., Bao, C., Feng, H., Deng, F., Lam, T.L.: Explicit attention-enhanced fusion for rgb-thermal perception tasks. IEEE Robotics and Automation Letters **8**(7), 4060–4067 (2023)
3. Lv, W., Xu, S., Zhao, Y., Wang, G., Wei, J., Cui, C., Du, Y., Dang, Q., Liu, Y.: Detrs beat yolos on real-time object detection. arXiv preprint arXiv:2304.08069 (2023)
4. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)

5. Zhao, W., Xie, S., Zhao, F., He, Y., Lu, H.: Metafusion: Infrared and visible image fusion via meta-feature embedding from object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13955–13965 (2023)
6. Zhao, Z., Bai, H., Zhang, J., Zhang, Y., Xu, S., Lin, Z., Timofte, R., Van Gool, L.: Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5906–5916 (2023)