Rolling and Denoising: Rethinking Dynamic Modal Fusion for Multi-Modal Object Re-Identification

Supplementary Material

6. Introduction

In this supplementary material, we provide more comprehensive experiments and in-depth analyses to further investigate fusion imbalance and conflict denoising. Specifically, we address the following questions:

Sec.A How robust are the denoising network and fusion module under modality conflicts?

- Robustness analysis under Gaussian noise perturbations.
- Experiments on missing-modality scenarios.

Sec.B How does modality imbalance evolve during training?

- Analysis of modality weights in training.

Sec.C How is the computational efficiency of the model and the selection of hyperparams?

- Analysis of computational efficiency.
- Analysis of the TOP-K value in LMD.
- Analysis of diffusion steps used in LMD.

These analyses offer a deeper insight of our RoDI and further demonstrate the effectiveness of its core modules.

7. How robust are the denoising network and fusion module under modality conflicts?

7.1. Robustness analysis under Gaussian noise perturbations.

Fig. 7 demonstrates that RoDI† achieves the strongest resilience against progressive Gaussian corruption. With the DINOv3 backbone, RoDI† hold leading performance from $\varepsilon = 0$ to $\varepsilon = 30$, maintaining 61.5% mAP and 66.6% Rank-1, while DeMo† degrades more rapidly across both metrics. CLIP-based models such as UGG-ReID and DeMo drop off sharply, and ViT-based EDITOR breaks down even at moderate noise levels. However, compare with these methods, RoDI using the same CLIP backbone, remains consistently robust in low-quality scenarios. These results emphasize the effectiveness of our proposed denoising mechanism which not only counters noise-induced fusion conflicts but also enables the model to effectively fuse both easy and hard samples.

7.2. Experiments on missing-modality scenarios.

Tab. 5 evaluates the robustness of different methods under various missing-modality settings on RGBNT201. Among CLIP-based approaches, RoDI achieves the best average performance with 58.3% mAP and 59.2% Rank-1, surpassing PromptMA by 3.7% mAP and 4.7% Rank-1, and clearly outperforming DeMo and TOP-ReID across most configurations. In single-modality cases, RoDI attains 68.3% mAP and 68.5% Rank-1 with only RGB available, 77.9% mAP and 80.5% Rank-1 with only NIR, and 62.9% mAP and 64.5% Rank-1 with only TIR, demonstrating strong capacity to exploit each individual spectrum. When two modalities are missing, RoDI shows particularly large gains in challenging settings such as "M (RGB+TIR)" and "M (NIR+TIR)", where it significantly outperforms all baselines. In the DINOv3-based comparison, RoDI† achieves the highest average performance with 61.7% mAP and 64.0% Rank-1, improving over DeMo† by 4.5% mAP and 7.0% Rank-1, and dominating in five out of six missing-modality configurations. These results demonstrate that RoDI's denoising module is capable of generating high-quality and reliable pseudo-modal representations, while its fusion module maintains strong robustness even under severe modality dropout.

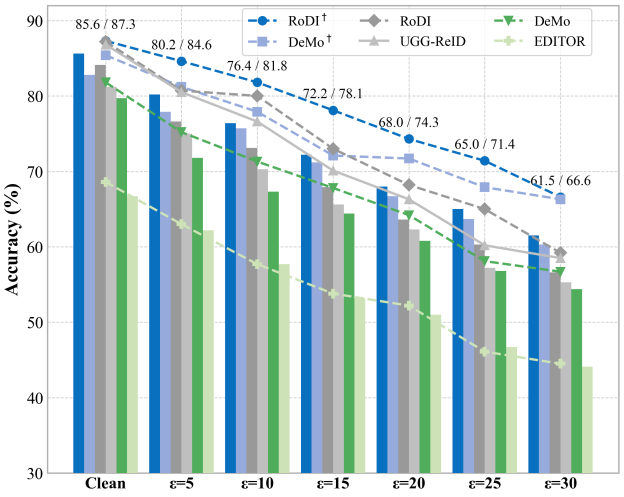

Figure 7. Robustness analysis on RGBNT201. Evaluation results with different levels of Gaussian noise added during testing after training on clean data. ε is the variance of Gaussian.

8. How does modality imbalance evolve during training?

8.1. Analysis of modality weights in training.

Fig. 8 visualizes the evolution of modality balance weights estimated by EDFL during training. Each fusion query is initialized from its own modality and thus exhibits a

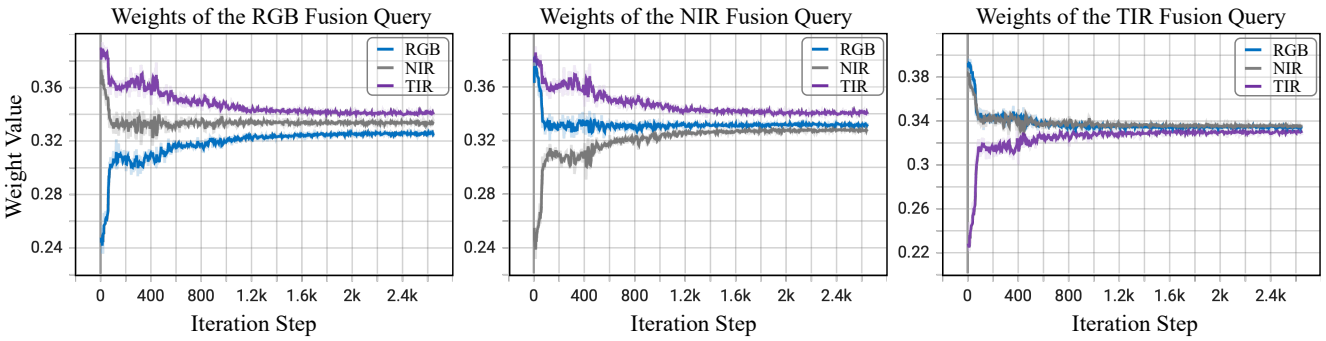

Figure 8. The evolution of modality balance weights estimated by EDFL during training.

Methods	M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)		Avg.	
	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1	mAP	R@1
PCB [28]	23.6	24.2	24.4	25.1	19.9	14.7	20.6	23.6	11.0	6.8	18.6	14.4	19.7	18.1
TOP-ReID [32]	54.4	57.5	64.3	67.6	51.9	54.5	35.3	35.4	26.2	26.0	34.1	31.7	44.4	45.4
DeMo [34]	63.3	65.3	72.6	75.7	56.2	54.1	45.6	46.5	26.3	24.9	40.3	38.5	50.7	50.8
PromptMA [48]	67.4	68.4	72.5	75.7	58.9	57.3	**51.5**	**53.0**	33.3	30.4	43.9	42.2	54.6	54.5
RoDI	**68.3**	**68.5**	**77.9**	**80.5**	**62.9**	**64.5**	47.7	49.0	**44.1**	**43.3**	**48.6**	**49.6**	**58.3**	**59.2**
DeMo† [34]	**71.3**	**71.8**	77.2	78.8	63.0	61.7	**56.5**	**57.1**	32.4	32.1	42.9	40.7	57.2	57.0
RoDI†	69.0	70.1	**80.0**	**82.7**	**66.4**	**69.5**	52.3	53.5	**49.0**	**50.4**	**53.2**	**57.7**	**61.7**	**64.0**

Table 5. Performance of the missing-modalities settings on RGBNT201. "M (X)" means missing the X spectral modality. Best results are in **bold**, the second bests are underlined.

strong bias against the other two modalities at the early stage. As training progresses, the model gradually learns to redistribute attention and incorporate information from multiple spectra. However, the upper trajectories reveal that fusion queries still exhibit noticeable bias toward the two non-origin modalities before convergence, and this bias diminishes only as the modality fusion process becomes more mature. These observations confirm that EDFL effectively exposes the dynamic imbalance across modalities and highlights the necessity of subsequent rolling-based rebalancing.

9. How is the computational efficiency of the model and the selection of hyperparams?

9.1. Analysis of computational efficiency.

Tab. 6 reports a comprehensive comparison of computational efficiency across state-of-the-art methods. Under the CLIP backbone, RoDI attains the best overall performance 84.1%/87.2% mAP/Rank-1 while using only 90.6M parameters and 38.7G FLOPs, which significantly lighter than methods such as UGG-ReID (103.2 M), PromptMA (107.9 M), and TOP-ReID (324.5 M). Even compared with lighter models such as ICPL-ReID, RoDI provides over +9.0% mAP improvement with similar computational cost. Under the DINOv3 backbone, RoDI† exhibits the same advantage,

Methods	Params	FLOPs	Throughput	Metrics	
	(M)	(G)	(TPS)	mAP	R@1
EDITOR [45]	119.3	40.8	335.1	66.5	68.3
TOP-ReID [32]	324.5	35.5	398.9	72.3	76.6
ICPL-ReID [18]	45.4	39.8	358.4	75.1	77.4
PromptMA [48]	107.9	36.2	343.5	78.4	80.9
DeMo [34]	98.8	35.1	403.6	79.0	82.3
MambaPro [33]	74.8	52.4	243.2	78.9	83.4
IDEA [35]	91.7	43.7	299.5	80.2	82.1
UGG-ReID [30]	103.2	35.0	371.4	81.2	86.8
RoDI	90.6	38.7	299.6	**84.1**	**87.2**
DINOv3† [27]	86.1	54.0	238.6	81.0	83.3
IDEA† [35]	94.0	63.5	189.6	81.2	83.2
DeMo† [34]	113.4	54.5	234.8	82.8	85.4
RoDI†	95.1	64.0	191.1	**85.3**	**87.9**

Table 6. Comparison with state-of-the-art methods on the size of learnable parameters, flops, and throughput on RGBNT201.

outperforming all DINOv3-based methods while remaining smaller than DeMo†. These results confirm that RoDI provides an optimal balance between efficiency and performance, delivering state-of-the-art accuracy without incurring additional computational overhead.

9.2. Analysis of the TOP-K value in LMD.

As illustrated in Fig. 9, the performance of LMD varies with different TOP-K selections guided by the belief scores es-

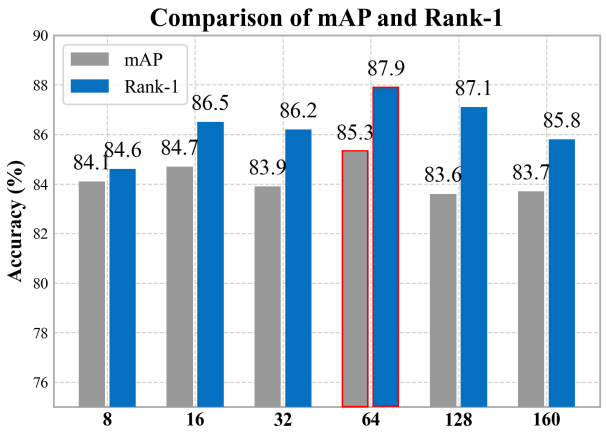

Figure 9. Performance comparison with different TOP-K value.

Denoising Steps	mAP	R@1
1	78.8	80.3
5	81.7	84.3
10	83.8	86.2
20	**85.3**	87.9
30	83.8	**89.0**
40	76.0	84.7
50	63.1	74.6

Table 7. Performance under different denoising step settings on RGBNT201 using DINOv3-Base.

timated in EDFL. Moderate K values (16 and 32) improve Rank-1 accuracy compared to very small selections, while K=64 provides the best overall results, achieving 85.3% mAP and 87.9% Rank-1. This indicates that selecting a sufficiently rich set of high-belief local tokens is crucial for stable fusion learning. When K becomes excessively large (128 or 160), performance drops, suggesting that including too many low-confidence regions dilutes the reliability of the fused evidence. These results confirm that evidence-based selection of local features effectively discard harmful regions and substantially mitigates fusion conflicts caused by noisy or low-quality local tokens.

9.3. Analysis of denoising steps used in LMD.

As shown in Tab. 7, the model benefits from increasing denoising steps up to 20, where it achieves the highest results 85.3%/87.9% mAP/Rank-1. Fewer steps (1–10) lead to under-denoising, while more steps (40–50) cause severe performance degradation due to over-smoothing and information loss. Although Rank-1 peaks at 30 steps, the drop in mAP indicates less reliable retrieval overall. These trends suggest that a moderate diffusion depth, such as 20 steps, yields the most stable and discriminative pseudo-modal representations.