FUSE: Frequency-domain Unification and Spectral Energy Alignment for Multi-modal Object Re-Identification

Xuanhao Qi [1], Tom H. Luan [1*], Yukang Zhang [2], Jinkai Zheng [1], Zhou Su [1], Shuwei Li [3*], Lei Tan [3]

[1] School of Cyber Science and Engineering, Xi'an Jiaotong University, Xi'an, China
[2] School of Informatics, Xiamen University, Xiamen, China
[3] National University of Singapore, Singapore

ICML International Conference On Machine Learning

Motivation

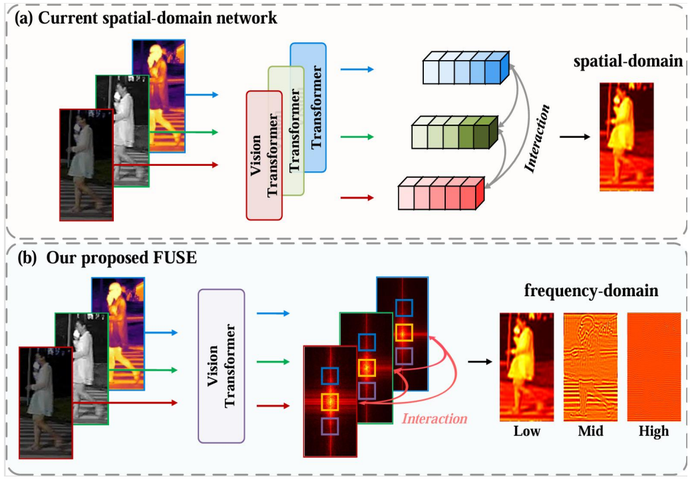

Fig. 1: Motivation of proposed FUSE.

(a) Existing multi-modal ReID methods mainly rely on spatial domain fusion, but the inherent low-frequency bias of both CNNs and ViTs causes models to predominantly capture global low-frequency semantics while neglecting mid and high-frequency details, leading to incomplete spectral representation and unstable cross-modal alignment.

(b) The proposed FUSE explicitly models in the frequency domain, where spectral decomposition and energy alignment partition features into low, mid, and high-frequency subspaces, achieving semantically interpretable cross-modal feature fusion.

Contribution

(1) We propose a frequency-domain framework named FUSE for multi-modal ReID. FUSE explicitly models inter-band interactions in the frequency domain, enabling more complete spectral representations and enhancing multi-spectral object ReID.

(2) We introduce a Spectral Decomposition Module (SDM) and a Cross-Modal Alignment Module (CAM), where SDM separates features into low, mid, and high-frequency subspaces to capture complementary spectral semantics, and CAM aligns the spectral energy distribution across modalities to stabilize cross-modal feature learning.

(3) Extensive experiments conducted on three public multi-modal ReID datasets, namely RGBNT201, RGBNT100, and MSVR310, demonstrate the superior performance of FUSE, improvement a 9.1% mAP and 9.5% Rank-1 on RGBNT201.

Experiment

	Methods	RGBNT201			
		mAP	R-1	R-5	R-10
Single	MUDeep (Qian et al., 2017)	23.8	19.7	33.1	44.3
	HACNN (Li et al., 2018)	21.3	19.0	34.1	42.8
	MLFN (Chang et al., 2018)	26.1	24.2	35.9	44.1
	PCB (Sun et al., 2018)	32.8	28.1	37.4	46.9
	OSNet (Zhou et al., 2019)	25.4	22.3	35.1	44.7
	CAL (Rao et al., 2021)	27.6	24.3	36.5	45.7
Multi	HAMNet (Li et al., 2020)	27.7	26.3	41.5	51.7
	PFNet (Zheng et al., 2021)	38.5	38.9	52.0	58.4
	IEEE (Wang et al., 2022)	47.5	44.4	57.1	63.6
	DENet (Zheng et al., 2023a)	42.4	42.2	55.3	64.5
	LRMM (Wu et al., 2025)	52.3	53.4	64.6	73.2
	HTT (Wang et al., 2024b)	71.1	73.4	83.1	87.3
	EDITOR (Zhang et al., 2024)	66.5	68.3	81.1	88.2
	RSCNet (Yu et al., 2024)	68.2	72.5	-	-
	TOP-ReID (Wang et al., 2024a)	72.3	76.6	84.7	89.4
	WTSF-ReID(Yu et al., 2025)	67.9	72.2	83.4	89.7
	DESANet (Dong et al., 2025)	74.6	77.6	87.1	91.3
	PromptMA (Zhang et al., 2025a)	78.4	80.9	87.0	88.9
	DeMo (Wang et al., 2025b)	79.0	82.3	88.8	92.0
	IDEA (Wang et al., 2025c)	80.2	82.1	90.0	93.3
	FUSE	**81.4**	**86.1**	**91.5**	**93.8**

	Methods	RGBNT100		MSVR310	
		mAP	R-1	mAP	R-1
Single	PCB (Sun et al., 2018)	57.2	83.5	23.2	42.9
	MGN (Wang et al., 2018)	58.1	83.1	26.2	44.3
	DMML (Chen et al., 2019)	58.5	82.0	19.1	31.1
	BoT (Luo et al., 2019)	78.0	95.1	23.5	38.4
	OSNet (Zhou et al., 2019)	75.0	95.6	28.7	44.8
	Circle Loss (Sun et al., 2020)	59.4	81.7	22.7	34.2
	HRCN (Zhao et al., 2021)	67.1	91.8	23.4	44.2
	TransReID (He et al., 2021)	75.6	92.9	18.4	29.6
	AGW (Ye et al., 2021)	73.1	92.7	28.9	46.9
Multi	HAMNet (Li et al., 2020)	74.5	93.3	27.1	42.3
	PFNet (Zheng et al., 2021)	68.1	94.1	23.5	37.4
	GAFNet (Guo et al., 2022)	74.4	93.4	-	-
	GPFNet (He et al., 2023)	75.0	94.5	-	-
	CCNet (Zheng et al., 2023b)	77.2	96.3	36.4	55.2
	HTT (Wang et al., 2024b)	75.7	92.6	-	-
	RSCNet (Yu et al., 2024)	82.3	96.6	39.5	49.6
	TOP-ReID (Wang et al., 2024a)	81.2	96.4	35.9	44.6
	EDITOR (Zhang et al., 2024)	82.1	96.4	39.0	49.3
	LRMM (Wu et al., 2025)	78.6	96.7	36.7	49.7
	FACENet (Zheng et al., 2025a)	81.5	96.9	36.2	54.1
	WTSF-ReID (Yu et al., 2025)	82.2	96.5	39.2	49.1
	MambaPro (Wang et al., 2025a)	83.9	94.7	47.0	56.5
	DESANet (Dong et al., 2025)	82.1	97.4	39.2	47.8
	DeMo (Wang et al., 2025b)	86.2	**97.6**	49.2	59.8
	IDEA (Wang et al., 2025c)	87.2	96.5	47.0	62.4
	FUSE	**88.5**	96.9	**50.1**	**65.7**

The Proposed FUSE

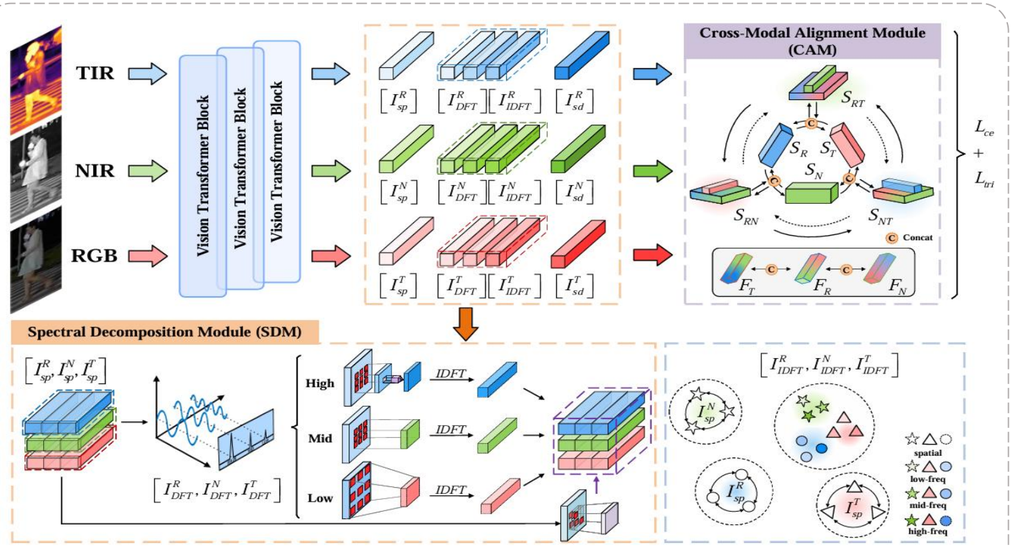

Fig. 2: Overall architecture of FUSE.

FUSE leverages frequency-domain modeling to enhance multi-modal person re-identification. Input images from RGB, NIR, and TIR modalities are processed by a shared ViT backbone to extract spatial features.

The Spectral Decomposition Module (SDM) adaptively partitions features into frequency sub-bands and applies specialized enhancement, while the Cross-Modal Alignment Module (CAM) performs frequency-aware interaction and consistency alignment. Identity classification and triplet losses supervise the final fused representation.

By modeling low-, mid-, and high-frequency components, FUSE captures global semantics and fine-grained identity cues. Frequency-consistency regularization reduces spectral gaps across modalities and improves robust cross-modal learning.