Multi-Modal Object Re-Identiffcation via Sparse Mixture-of-Experts

Yingying Feng*[1], Jie Li*[2], Chi Xie[3], Lei Tan[4], Jiayi Ji[2 4],
[1]School of Computer Science and Engineering, Northeastern University, Shenyang, China
[2]School of Informatics, Xiamen University, Xiamen, China
[3]Tongji University, Shanghai, China
[4]National University of Singapore, Singapore

INTERNATIONAL CONFERENCE ON MACHINE LEARNING
ICML | 2025

Motivation

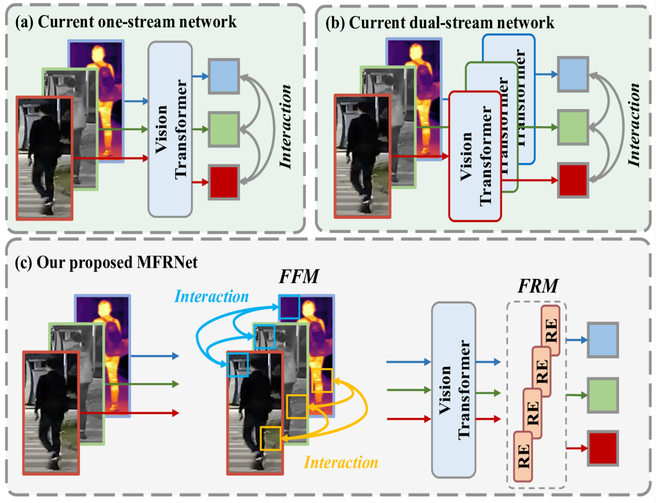

Fig. 1: Motivation of proposed MFRNet.

(a) and (b) show the current mainstream one-stream and two-stream networks, respectively. Besides the limitation of neglecting the pixel-level alignment characteristics of multi-modal images, current methods also face the challenge of balancing the modality-specific and modality-shared representation in a unified network.

(c) shows our proposed MFRNet and 'RE' refers to the representation expert. MFRNet inherits the idea of MoE and extends it with multi-modal fusion (FFM) and representation (FRM), allowing it to achieve fine-grained interaction and efficient representation for multi-modal data.

Contribution

(1) We propose a Modality Fusion and Representation Network (MFRNet) for multi-modal object re-identification, which inherits the idea of a sparse mixture of experts and extends it with multi-modal fusion and representation.

(2) We introduce a Feature Fusion Module (FFM) and a Feature Representation Module (FRM). The former aims to achieve fine-grained interaction between multi-modal inputs, while the latter aims to achieve efficient and balanced feature extraction between modality-shared and modality-specific representations.

(3) Extensive experiments on three public multi-spectral object ReID datasets including RGBNT201, RGBNT100, and MSVR310, verifying the superior performance of MFRNet.

Experiments

	Methods	*m*AP	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., 2023)	42.4	42.2	55.3	64.5
	UniCat (Crawford et al., 2023)	57.0	55.7	-	-
	HTT (Wang et al., 2024c)	71.1	73.4	83.1	87.3
	EDITOR (Zhang et al., 2024a)	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
	Ours	**80.7**	**83.6**	**91.9**	**94.1**

Methods	Params(M)	Flops(G)
HTT (Wang et al., 2024c)	85.6	33.1
EDITOR (Zhang et al., 2024a)	117.5	38.6
TOP-ReID (Wang et al., 2024a)	278.2	34.5
Ours	**57.1**	**22.1**

	Methods	RGBNT100		MSVR310	
		*m*AP	R-1	*m*AP	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., 2022)	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	-	-
	GraFT (Yin et al., 2023)	76.6	94.3	-	-
	GPFNet (He et al., 2023)	75.0	94.5	-	-
	PHT (Pan et al., 2023)	79.9	92.7	-	-
	UniCat (Crawford et al., 2023)	79.4	96.2	-	-
	CCNet (Zheng et al., 2022)	77.2	96.3	36.4	55.2
	HTT (Wang et al., 2024c)	75.7	92.6	-	-
	TOP-ReID (Wang et al., 2024a)	81.2	96.4	35.9	44.6
	EDITOR (Zhang et al., 2024a)	82.1	96.4	39.0	49.3
	RSCNet (Yu et al., 2024)	82.3	96.6	39.5	49.6
	Ours	**88.2**	**97.4**	**50.6**	**64.8**

	Methods	M (RGB)		M (NIR)		M (TIR)		M (RGB+NIR)		M (RGB+TIR)		M (NIR+TIR)		Average	
		*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1	*m*AP	R-1
Single	MUDeep (Qian et al., 2017)	19.2	16.4	20.0	17.2	18.4	14.2	13.7	11.8	11.5	6.5	12.7	8.5	15.9	12.9
	HACNN (Li et al., 2018)	12.5	11.1	20.5	19.4	16.7	13.3	9.2	6.2	6.3	2.2	14.8	12.0	13.3	10.7
	MLFN (Chang et al., 2018)	20.2	18.9	21.1	19.7	17.6	11.1	13.2	12.1	8.3	3.5	13.1	9.1	15.6	12.4
	PCB (Sun et al., 2018)	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
	OSNet (Zhou et al., 2019)	19.8	17.3	21.0	19.0	18.7	14.6	12.3	10.9	9.4	5.4	13.0	10.2	15.7	12.9
Multi	PFNet (Zheng et al., 2021)	-	-	31.9	29.8	25.5	25.8	-	-	-	-	26.4	23.4	-	-
	DENet (Zheng et al., 2023)	-	-	35.4	36.8	33.0	35.4	-	-	-	-	32.4	29.2	-	-
	TOP-ReID (Wang et al., 2024a)	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
	Ours	**64.7**	**65.2**	**72.3**	**76.1**	51.6	49.5	**41.4**	**43.4**	**27.3**	**27.9**	**37.2**	**35.6**	**49.1**	**49.6**

The Proposed MFRNet

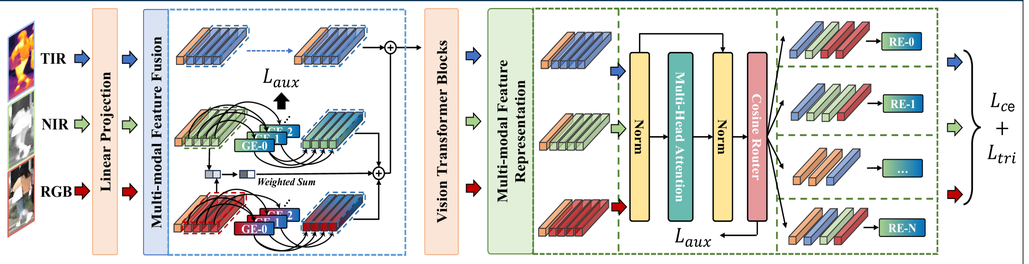

Fig. 2: Overall architecture of MFRNet.

MFRNet is built from the basic Transformer blocks inherited from vanilla ViT, with two new modules (feature fusion module and feature representation module) added to adapt to multi-modal object re-identification tasks.
In the blue box, the Feature Fusion Module (FFM) employs multiple simple generators to adaptively provide fine-grained interaction information, while in the green box, the Feature Representation Module (FRM) employs diverse representation experts to extract and combine modality-specific and modality-shared features. Here, GE refers to the generation experts in FFM, while RE represents the representation experts in FRM. Notably, the figure illustrates the NIR+RGB→TIR interaction, with the same interaction applying to NIR and RGB.

Visulization Expert

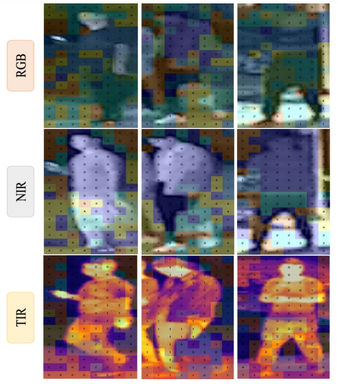

Fig. 3: Visualization Result.

It can be observed that Expert 0 concentrates on extracting pedestrian features in the NIR modality, while Experts 1, 2, and 3 focus on pedestrian regions in the RGB and TIR modalities. Additionally, Experts 4 and 5 tend to extract background features across all modalities. As our initial hypothesis predicted, semantically similar content achieves knowledge sharing through selecting analogous experts, while modality-specific representations that resist fusion maintain their distinct characteristics via dedicated expert allocation.