NEURAL INFORMATION PROCESSING SYSTEMS

MARS

Multimedia Analysis
& Reasoning Lab

Empowering Visible-Infrared Person Re-Identification with Large Foundation Models

Zhangyi Hu[1#], Bin Yang[1#], Mang Ye[1*]

[1]School of Computer Science, Wuhan University, Wuhan, China

Paper: https://neurips.cc/virtual/2024/poster/93497

Project Page: https://github.com/WHU-HZY/TVI-LFM

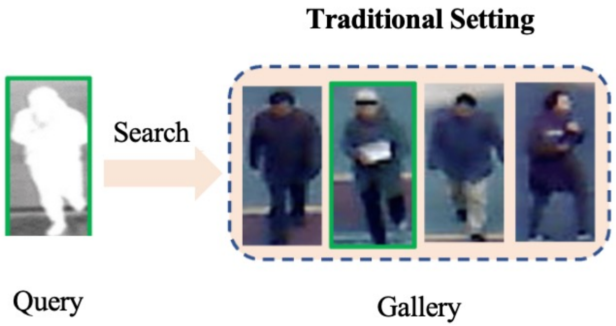
Background

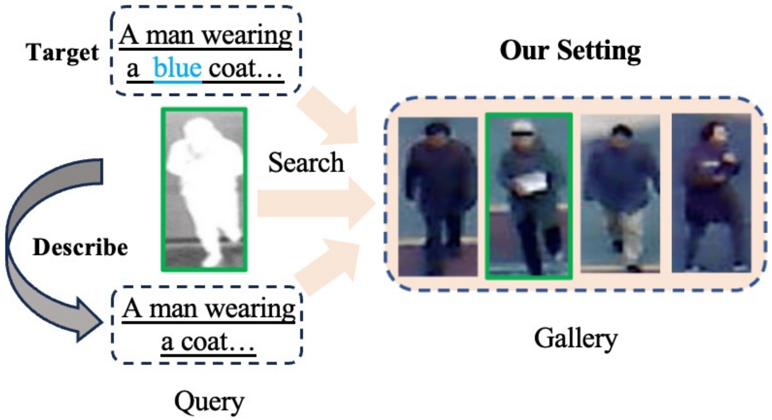
Visible-Infrared Person Re-Identification

Learn a cross-modality ReID model on a set of visible-infrared images with identity labels

Problem

Modality gap primarily caused by critical information absence

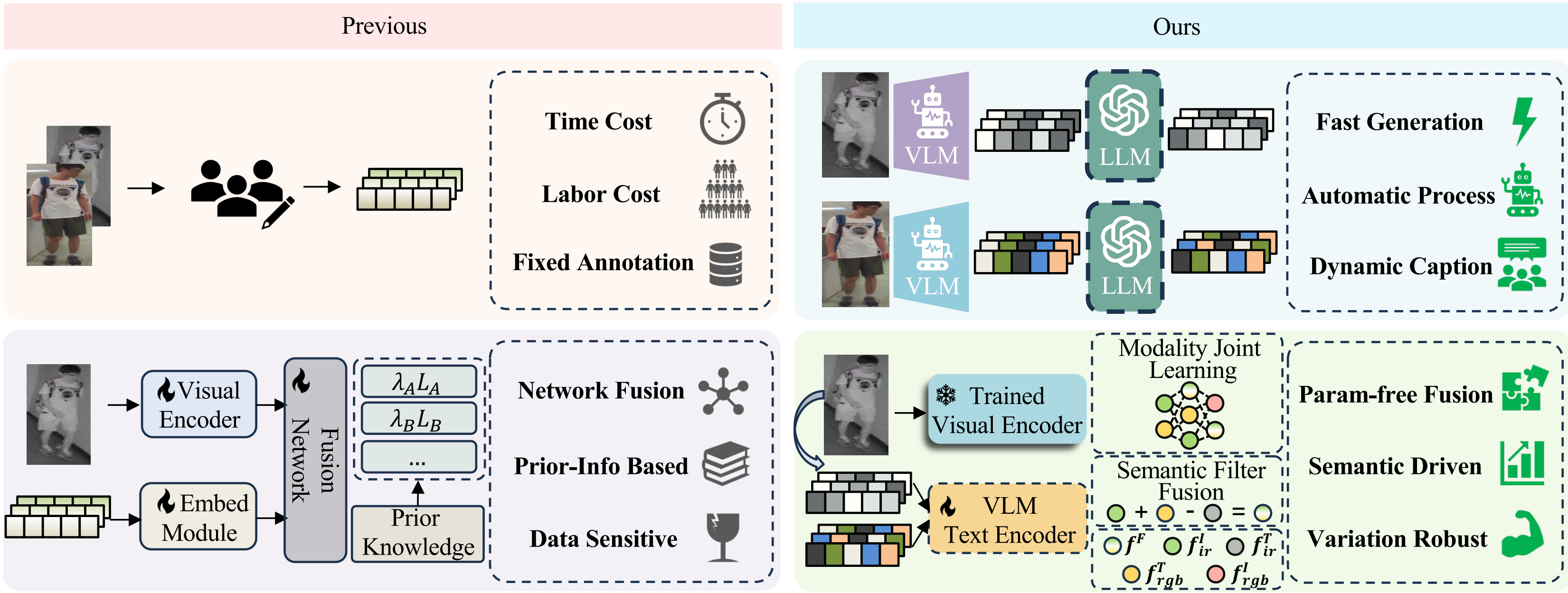




- ☐ Traditional VI-ReID
- ✗ Critical information absence in the infrared modality, e.g. color
- ✗ Significant modality gap

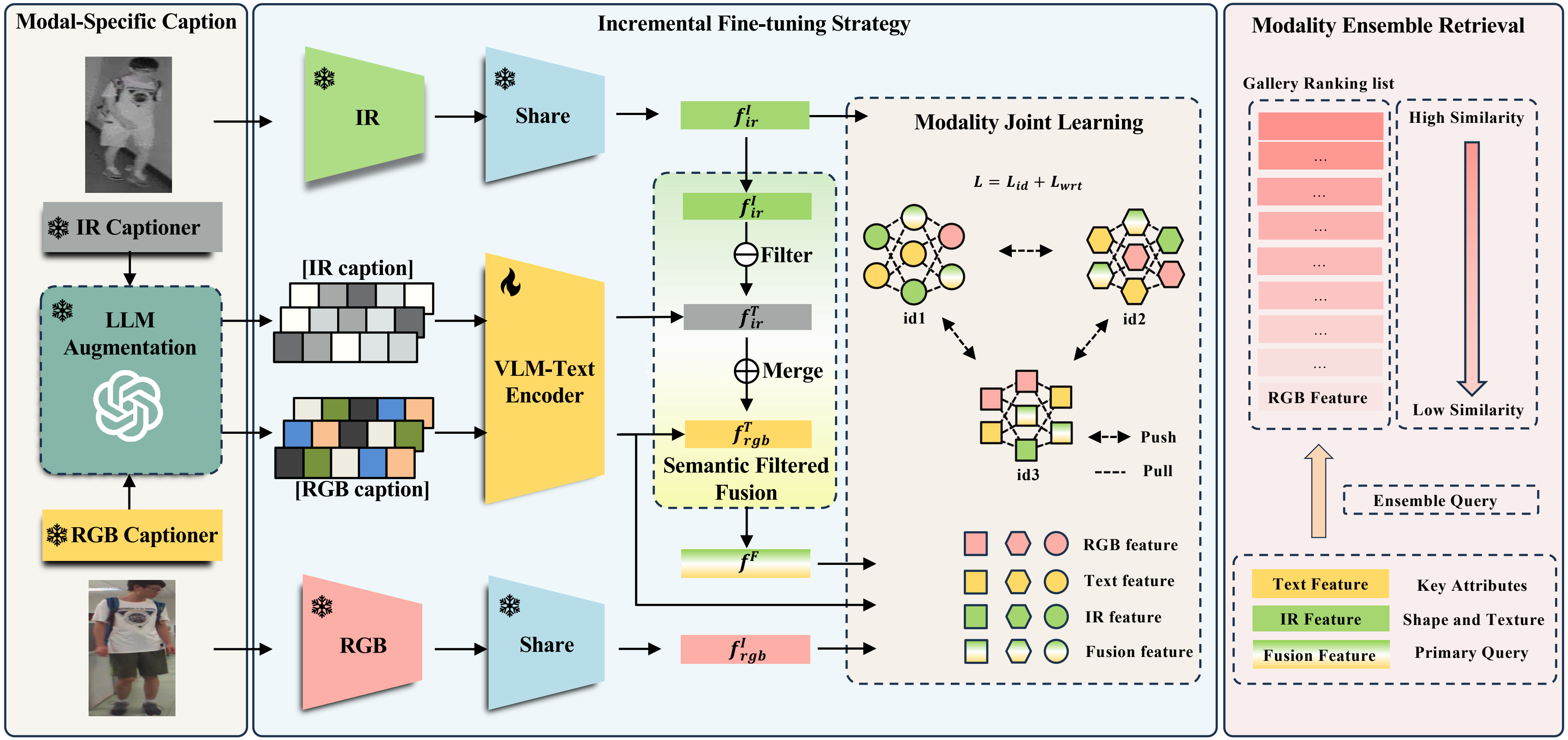
- ☐ **VI-ReID w/ heterogeneous text**
- ✓ **Enhance the infrared modality by auxiliary information**
- ✓ **Bridge the modality gap with texts**

Motivation



- Existing methods rely on **human description, complex prior-info dependent modules** to compensate for the infrared modality.
- Developments of **VLMs** and **LLMs** motivate us to propose a VI-ReID framework driven by Large Foundation Model (**TVI-LFM**). The basic idea is to **enrich infrared representations** with **automatically** generated **heterogeneous text**.

Framework


Modal-Specific Caption
IR Captioner
LLM Augmentation
RGB Captioner
Incremental Fine-tuning Strategy
IR
Share
[IR caption]
[RGB caption]
VLM-Text Encoder
RGB
Share
f^I_{ir}
f^I_{ir}
Filter
f^T_{ir}
Merge
f^T_{rgb}
Semantic Filtered Fusion
f^F
f^I_{rgb}
Modality Joint Learning
$L = L_{id} + L_{wrt}$
id1
id2
id3
Push
Pull
RGB feature
Text feature
IR feature
Fusion feature
Modality Ensemble Retrieval
Gallery Ranking list
RGB Feature
High Similarity
Low Similarity
Ensemble Query
Text Feature
Key Attributes
IR Feature
Shape and Texture
Fusion Feature
Primary Query

Methodology

Modal-Specific Caption (MSC):

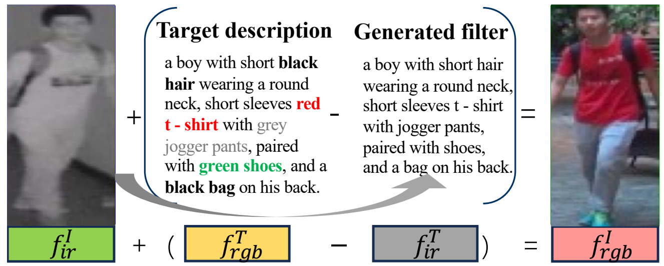
MSC introduces fine-tuned **VLMs** as captioners to **automatically** generate **heterogeneous** text from visible and infrared images, and utilizes **LLM rephrasing** for text augmentation.

Methodology

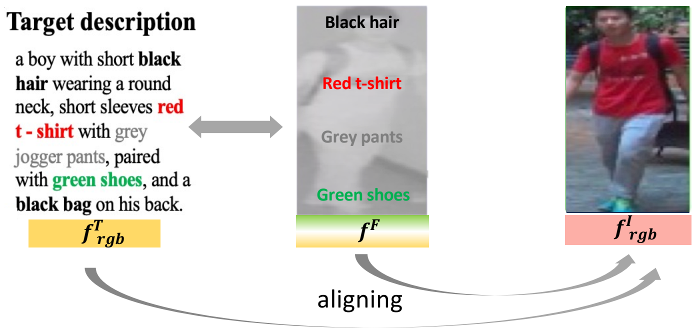
Incremental Fine-tuning Strategy (IFS)

IFS incorporates a pre-trained VLM to extract features from text generated by MSC, and incrementally fine-tunes the text encoder to minimize the domain gap between the generated texts and the original visual modalities.

- **Semantic Filtered Fusion (SFF)**



- **Modality Joint Learning (MJL)**



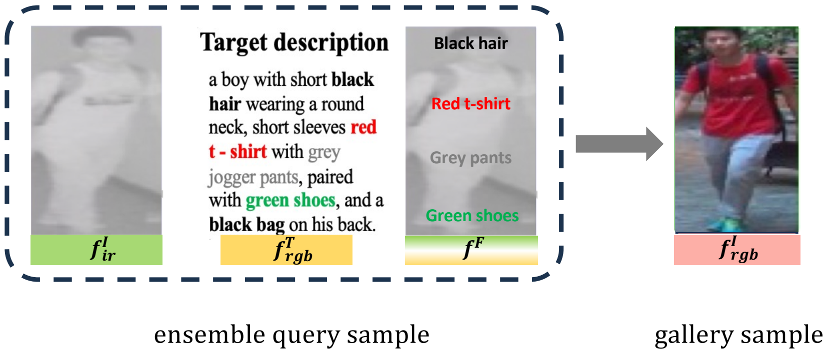
Methodology

Modality Ensemble Retrieval (MER):

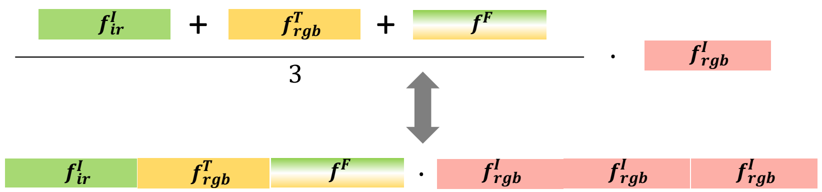
Forming **ensemble query** representations for improved retrieval.

Utilize **complementary strengths** from different modalities to improve the query representations:

- **Fusion**: primary query
- **Infrared** image: shape and texture
- **Text**: key attributes like color



Calculating the **similarity score** based on the ensemble features and gallery features is equivalent to calculating the similarity among **higher dimension** features with **larger inter-class distances.**



Experiments

Comparison with Stat-Of-The-Art Methods

Table 2: Comparison with the state-of-the-art methods on the proposed Tri-SYSU-MM01.

Methods	Venue	Type	All Search			Indoor Search		
			R-1	mAP	mINP	R-1	mAP	mINP
Zero-Padding [46]	ICCV-17		14.80	15.95	-	20.58	26.92	-
HCML [56]	AAAI-18		14.32	16.16	-	24.52	30.08	-
cmGAN [6]	IJCAI-18		26.97	27.80	-	31.63	42.19	-
AlignGAN [43]	ICCV-19		42.40	40.70	-	45.90	54.30	-
AGW [59]	TPAMI-21		47.50	47.65	35.30	54.17	62.97	59.23
DDAG [58]	ECCV-20		54.75	53.02	39.62	61.02	67.98	62.61
CM-NAS [12]	ICCV-21	$I \rightarrow R$	61.99	60.02	-	67.01	72.95	-
DART [53]	CVPR-22		68.7	66.3	-	82.0	73.8	-
CAJ [57]	ICCV-21		69.88	66.89	53.61	76.26	80.37	76.79
DEEN [65]	CVPR-23		74.70	71.80	-	80.30	83.30	-
SAAI [10]	ICCV-23		75.90	77.03	-	83.20	88.01	-
MSCLNet [64]	ECCV-22		76.99	71.64	-	78.49	81.17	-
SGIEL [11]	CVPR-23		77.12	72.33	-	82.07	82.95	-
PartMix [20]	CVPR-23		77.78	74.62	-	81.52	84.38	-
YYDS [9]	Arxiv-24	$I+T \rightarrow R$	74.60	70.35	56.01	81.35	83.64	79.56
VI-ReID Backbone	-	$I \rightarrow R$	69.89	66.74	53.34	76.91	80.64	76.70
TVI-LFM	-	$I+T \rightarrow R$	**84.90**	**81.47**	**70.85**	**89.06**	**90.78**	**88.39**

Table 3: Comparison with the state-of-the-art methods on the proposed Tri-RegDB and Tri-LLCM.

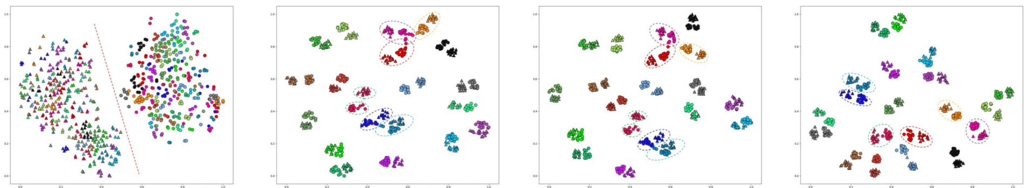
Methods	Venue	Type	Tri-RegDB			Tri-LLCM		
			R-1	mAP	mINP	R-1	mAP	mINP
DDAG [58]	ECCV-20		68.06	61.80	48.62	40.3	48.4	-
AGW [59]	TPAMI-21		70.49	65.90	51.24	43.6	51.8	-
CAJ [57]	ICCV-21	$I \rightarrow R$	84.8	77.8	61.56	48.8	56.6	-
DART [53]	CVPR-22		82.0	73.8	-	52.2	59.8	-
MMN [66]	MM-21		87.5	80.5	-	52.5	58.9	-
DEEN [65]	CVPR-23		89.5	83.4	-	54.9	62.9	-
YYDS [9]	Arxiv-24	$I+T \rightarrow R$	90.95	84.22	70.12	58.13	64.91	61.77
VI-ReID Backbone	-	$I \rightarrow R$	89.51	83.51	69.65	53.53	59.77	56.40
TVI-LFM	-	$I+T \rightarrow R$	**91.38**	**85.92**	**72.73**	**58.19**	**65.08**	**61.83**

Ablation Study

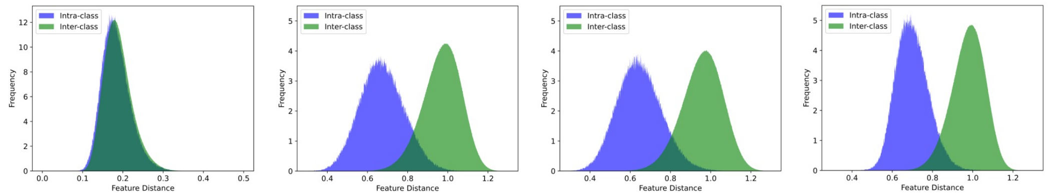
$I+T \rightarrow R$					Tri-SYSU-MM01			Tri-LLCM		
B	SFF	MJL	LLM	MER	R1	mAP	mINP	R1	mAP	mINP
✓					72.52	69.15	55.93	52.63	58.82	55.43
✓	✓				77.00	73.73	61.50	54.73	60.95	57.64
✓	✓	✓			83.97	80.40	69.46	56.76	63.58	60.35
✓	✓	✓	✓		84.17	80.72	70.02	57.13	64.06	60.72
✓	✓	✓		✓	84.88	81.32	70.57	57.09	63.87	60.62
✓	✓	✓	✓	✓	**84.90**	**81.47**	**70.85**	**58.19**	**65.08**	**61.83**

Visualization

○ Gall samples △ Query samples



(a) Initial Distribution (b) VI-ReID Backbone Distribution (c) Baseline Distribution (d) TVI-LFM Distribution



(e) Initial Distance (f) VI-ReID Backbone Distance (g) Baseline Distance (h) TVI-LFM Distance

Thanks for watching!

Zhangyi Hu