CVPR
SEATTLE, WA JUNE 17-21, 2024

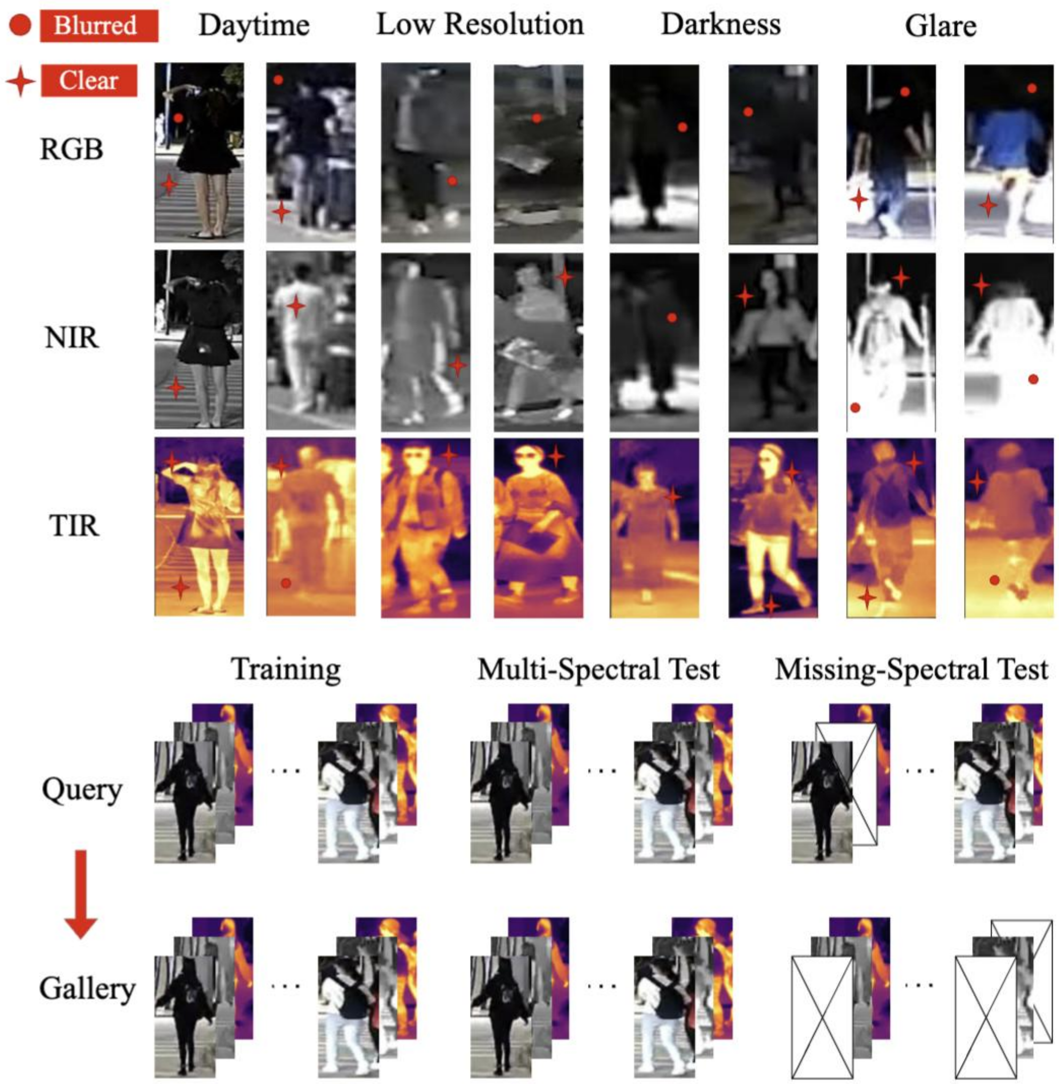
Magic Tokens: Select Diverse Tokens for Multi-modal Object Re-Identification

Pingping Zhang[1,2]*, Yuhao Wang[1], Yang Liu[1], Zhengzheng Tu[2,3] and Huchuan Lu[1]
[1] School of Future Technology, School of Artificial Intelligence, Dalian University of Technology, China
[2] Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University, China
[3] School of Computer Science and Technology, Anhui University, China
{zhpp,ly,lhchuan}@dlut.edu.cn, 924973292@mail.dlut.edu.cn, zhengzhengahu@163.com

Email: 924973292@mail.dlut.edu.cn
GitHub: https://github.com/924973292/EDITOR

Presenter: Yuhao Wang

Background



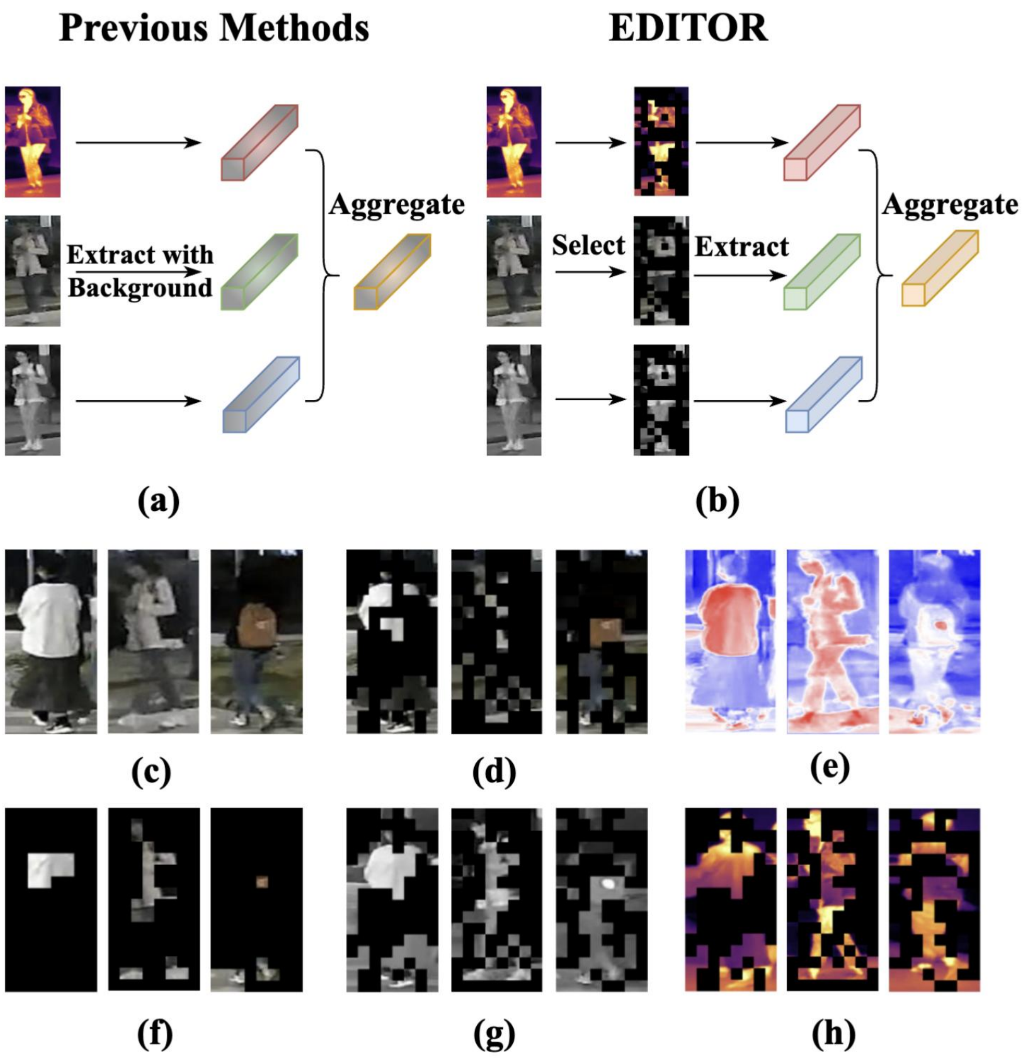
In challenging visual environments, the salient information about the object in RGB images is severely disrupted, resulting in poor robustness of existing single-modal methods.

↓

Multi-modal Object ReID

Y. Wang, et al., TOP-ReID: Multi-spectral Object Re-Identification with Token Permutation, ***AAAI2024***

Motivation



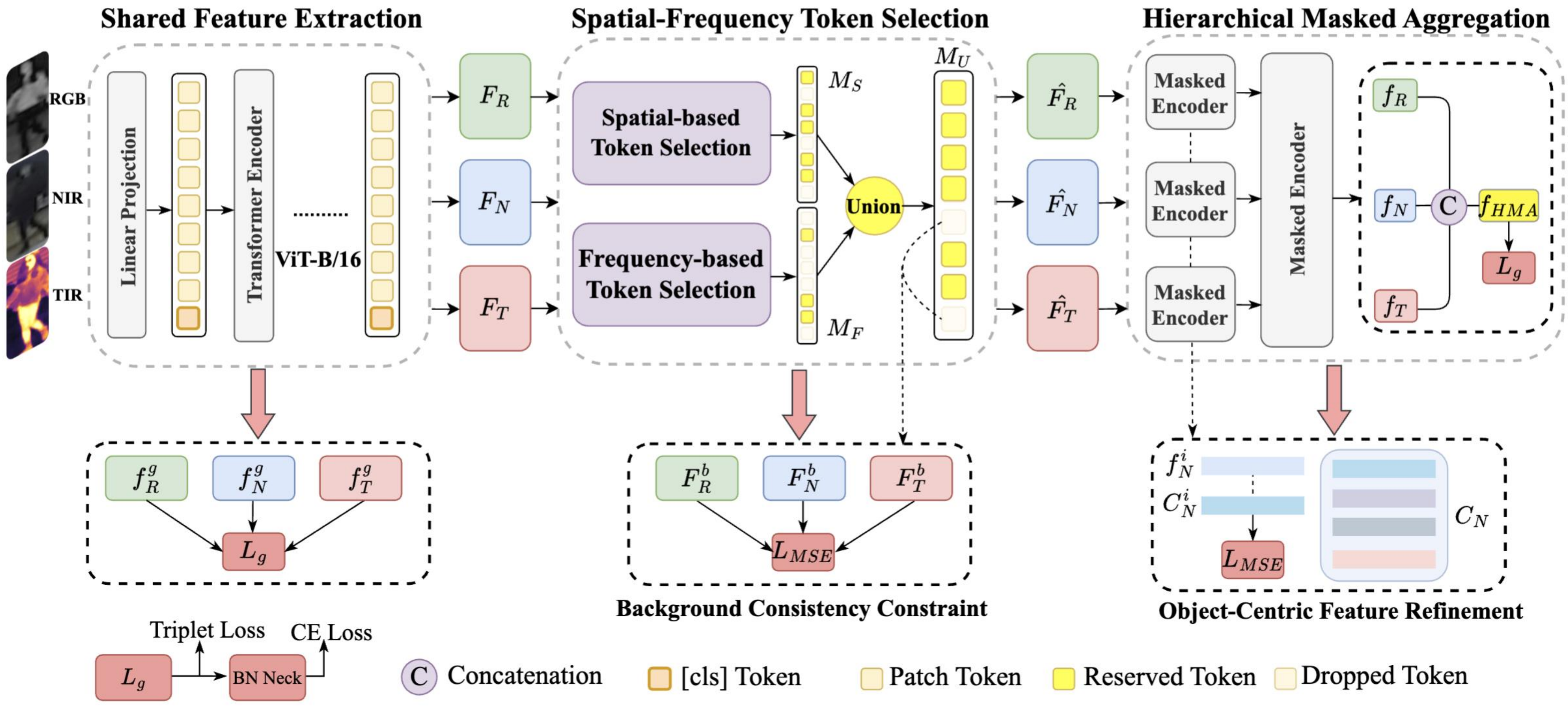
- Within individual modalities, backgrounds introduce additional noise, especially in challenging visual scenarios.
- Across different modalities, backgrounds introduce overhead in reducing modality gaps.

↓

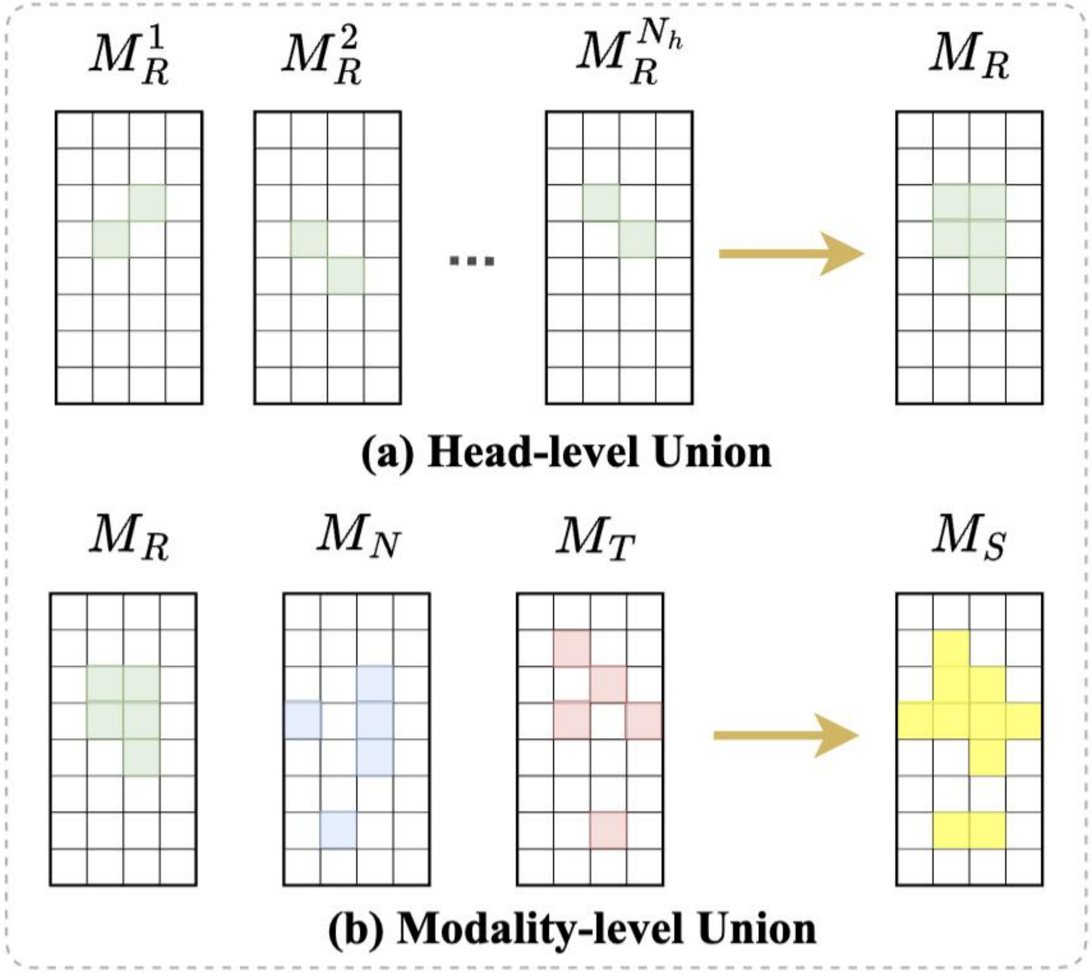
Preserving the diverse features of different modalities!

Minimizing background interference!

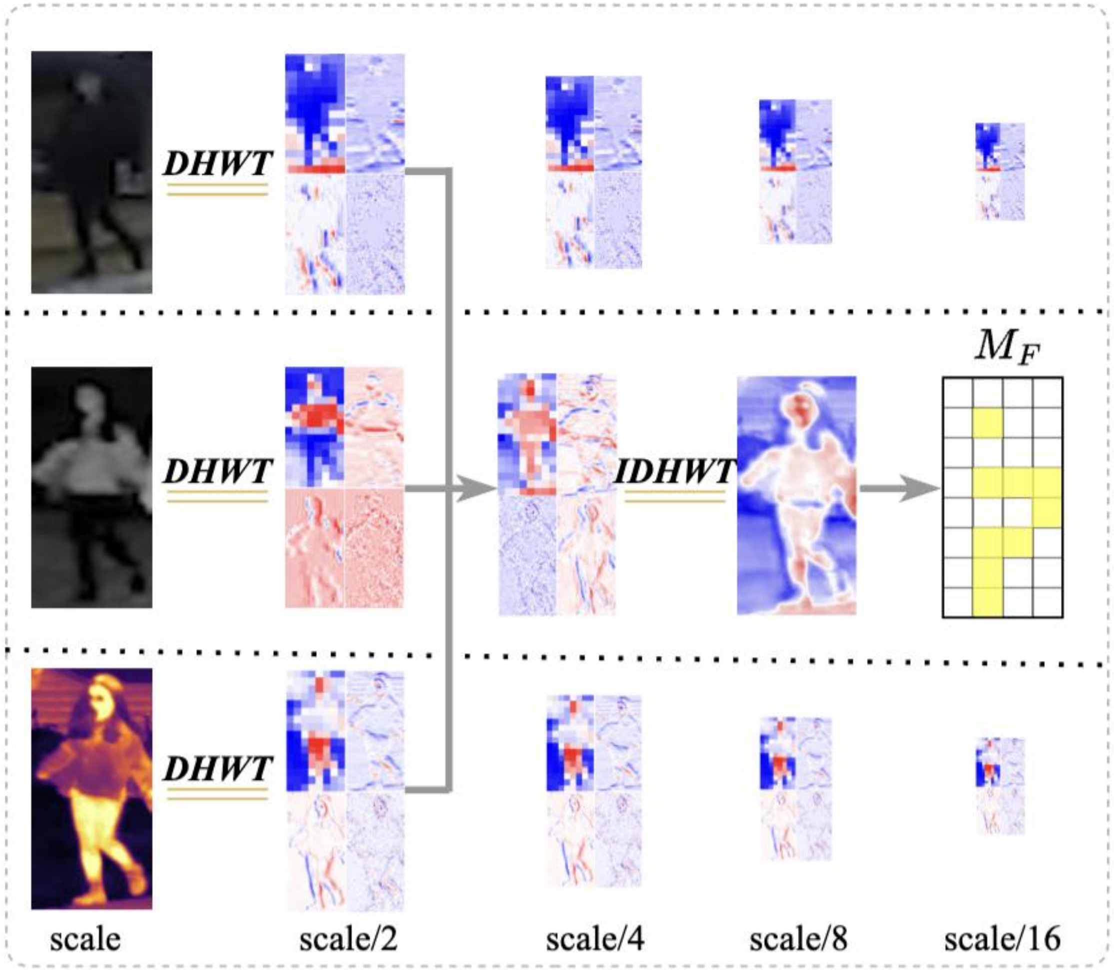
Overall Architecture



Modules



Spatial-based Selection



Frequency-based Selection

Multi-modal Testing

Table 1. Performance comparison on three multi-modal object ReID benchmarks. The best and second results are in bold and underlined, respectively. *denotes Transformer-based methods, while the rest are CNN-based methods. Both single-modal and multi-modal methods are included. For the comparison between TOP-ReID and EDITOR, A and B means the AL setting and BL setting [43], respectively.

(a) Comparison on RGBNT201.

Methods		RGBNT201			
		mAP	R-1	R-5	R-10
Single	MUDeep [30]	23.8	19.7	33.1	44.3
	HACNN [18]	21.3	19.0	34.1	42.8
	MLFN [2]	26.1	24.2	35.9	44.1
	PCB [34]	32.8	28.1	37.4	46.9
	OSNet [57]	25.4	22.3	35.1	44.7
	CAL [32]	27.6	24.3	36.5	45.7
Multi	HAMNet [17]	27.7	26.3	41.5	51.7
	PFNet [53]	38.5	38.9	52.0	58.4
	IEEE [44]	49.5	48.4	59.1	65.6
	DENet [55]	42.4	42.2	55.3	64.5
	UniCat* [4]	57.0	55.7	-	-
	TOP-ReID (A)* [43]	**72.3**	**76.6**	**84.7**	**89.4**
	TOP-ReID (B)* [43]	64.6	64.6	77.4	82.4
	EDITOR (A)*	66.5	68.3	81.1	88.2
	EDITOR (B)*	65.7	68.8	82.5	89.1

More Stable!

(b) Comparison on RGBNT100 and MSVR310.

Methods		RGBNT100		MSVR310	
		mAP	R-1	mAP	R-1
Single	PCB [34]	57.2	83.5	23.2	42.9
	MGN [40]	58.1	83.1	26.2	44.3
	DMML [3]	58.5	82.0	19.1	31.1
	BoT [26]	78.0	95.1	23.5	38.4
	OSNet [57]	75.0	95.6	28.7	44.8
	Circle Loss [35]	59.4	81.7	22.7	34.2
	HRCN [52]	67.1	91.8	23.4	44.2
	AGW [47]	73.1	92.7	28.9	46.9
	TransReID* [13]	75.6	92.9	18.4	29.6
Multi	HAMNet [17]	74.5	93.3	27.1	42.3
	PFNet [53]	68.1	94.1	23.5	37.4
	GAFNet [9]	74.4	93.4	-	-
	CCNet [54]	77.2	96.3	36.4	**55.2**
	GraFT* [48]	76.6	94.3	-	-
	GPFNet [12]	75.0	94.5	-	-
	PHT* [29]	79.9	92.7	-	-
	UniCat* [4]	79.4	96.2	-	-
	TOP-ReID (A)* [43]	73.7	92.2	30.2	33.7
	TOP-ReID (B)* [43]	81.2	96.4	35.9	44.6
	EDITOR (A)*	79.8	93.9	35.8	43.1
	EDITOR (B)*	**82.1**	**96.4**	**39.0**	49.3

More Competitive!

Main Ablation

	Module		Loss		RGBNT201			
	SFTS	HMA	BCC	OCFR	mAP	R-1	R-5	R-10
A	×	×	×	×	54.0	53.5	70.2	78.8
B	×	✓	×	×	60.7	62.4	77.2	83.6
C	✓	✓	×	×	62.2	65.0	79.3	85.4
D	✓	✓	✓	×	65.2	65.9	82.2	87.1
E	✓	✓	×	✓	64.8	66.9	82.3	87.3
F	✓	✓	✓	✓	**65.7**	**68.8**	**82.5**	**89.1**

SFTS: Selecting Object-centirc Tokens

HMA: Aggregating Pure Multi-modal Features

BCC: Stablizing the Selection

OCFR: Suppressing background noise within modalities

Stable performance on both person and vehicle datasets

	Module		Loss		RGBNT100			
	SFTS	HMA	BCC	OCFR	mAP	R-1	R-5	R-10
A	×	×	×	×	75.1	93.4	95.0	95.8
B	×	✓	×	×	77.8	94.0	95.1	96.0
C	✓	✓	×	×	79.1	94.3	95.3	96.1
D	✓	✓	✓	×	80.6	95.5	96.4	97.2
E	✓	✓	×	✓	80.4	94.8	95.5	96.3
F	✓	✓	✓	✓	**82.1**	**96.4**	**96.9**	**97.4**

Main Ablation

Methods	RGBNT201			
	mAP	R-1	R-5	R-10
w/o selection	60.7	62.4	77.2	83.6
w/ separation	57.7	58.5	75.4	82.5
w/ union	**62.2**	**65.0**	**79.3**	**85.4**

Different modality selections are

significantly different!

That's why we introduce the

Modality-Union!

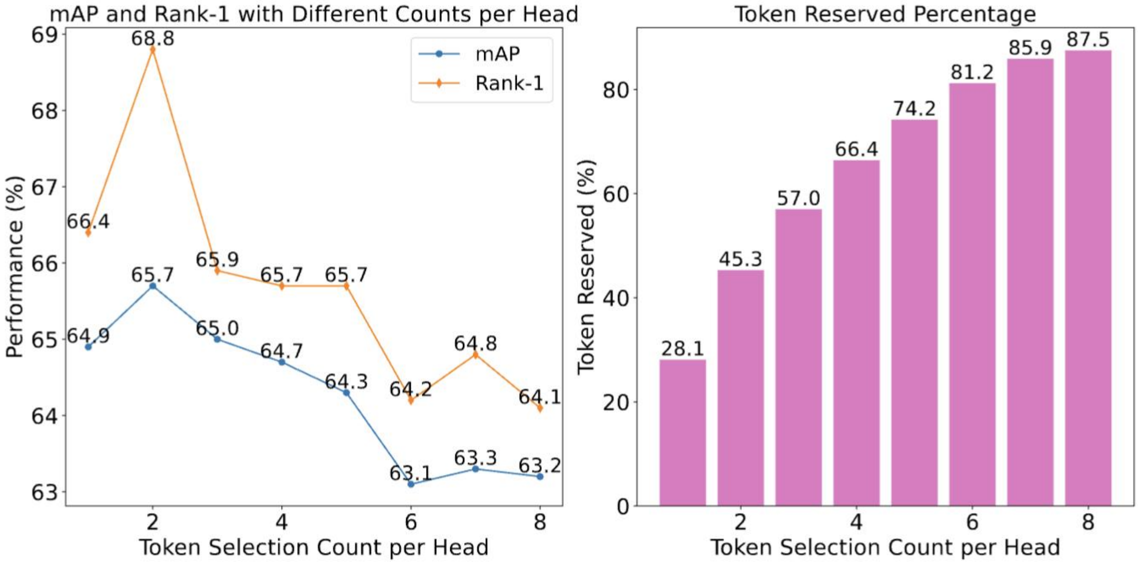
Frequency-based :The most salient parts [fixed]

Spatial-based: ROI of the EDITOR [Learnable]

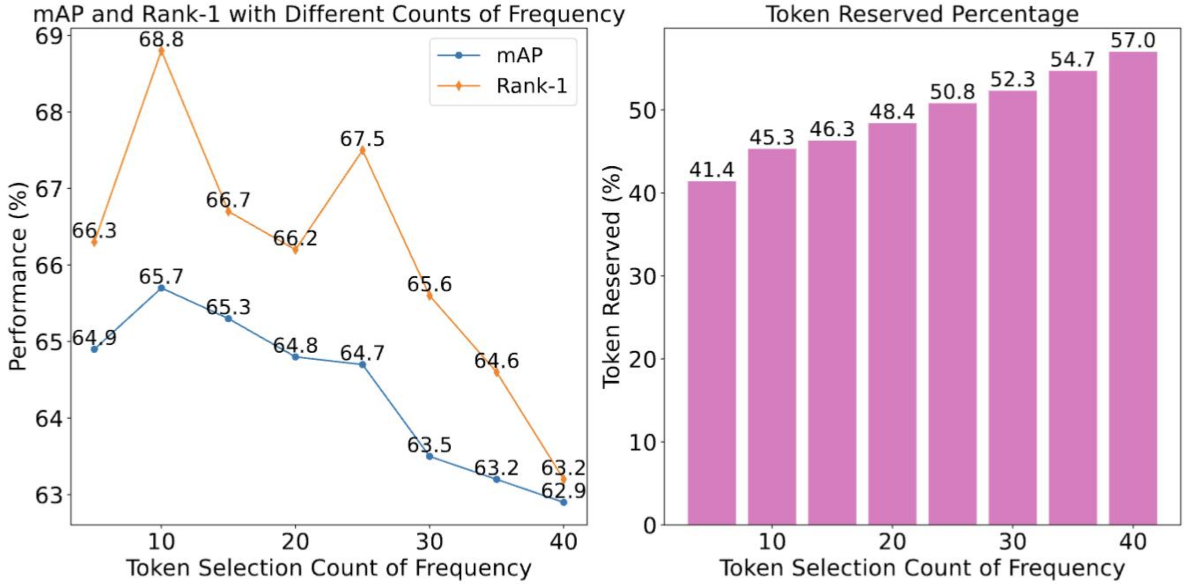
Selection Methods	Reserved Tokens	RGBNT201	
	Average number	mAP	R-1
Modality	30.2	64.2	65.7
Spatial	55.0	65.0	66.8
Frequency	55.0	64.1	65.3
Spatial+Frequency	58.0	**65.7**	**68.8**

Main Ablation

Spatial-based Token Selection [Learnable]



Frequency-based Token Selection [fixed]



With the increase in reserved tokens, the performance drops!

More noise from the background is introduced!

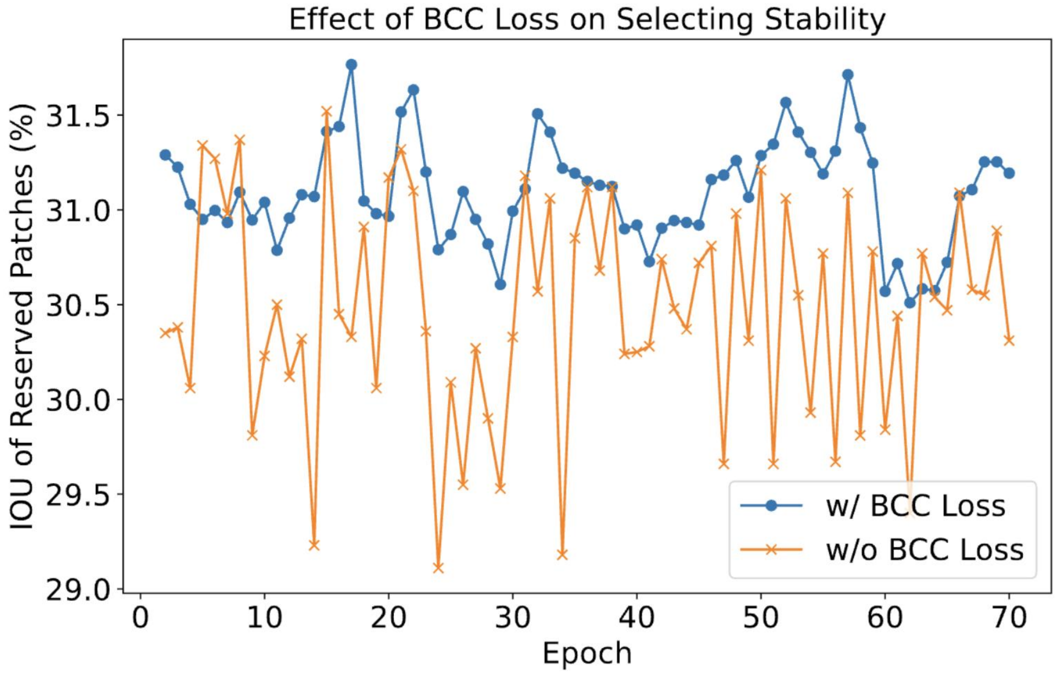
Main Ablation

Parameter Comparison

Methods	Params(M)	RGBNT100	
		mAP	Rank-1
PCB [34]	72.33	57.2	83.5
OSNet [57]	7.02	75.0	95.6
HAMNet [17]	78.00	74.5	93.3
CCNet [54]	74.60	77.2	96.3
GAFNet [9]	130.00	74.4	93.4
TransReID* [13]	278.23	75.6	92.9
UniCat* [4]	259.02	79.4	96.2
GraFT* [48]	101.00	76.6	94.3
TOP-ReID* [43]	324.53	81.2	96.4
EDITOR*	118.55	**82.1**	**96.4**

Parameter efficient!

A more stable selection process



More Stable!

Visualization

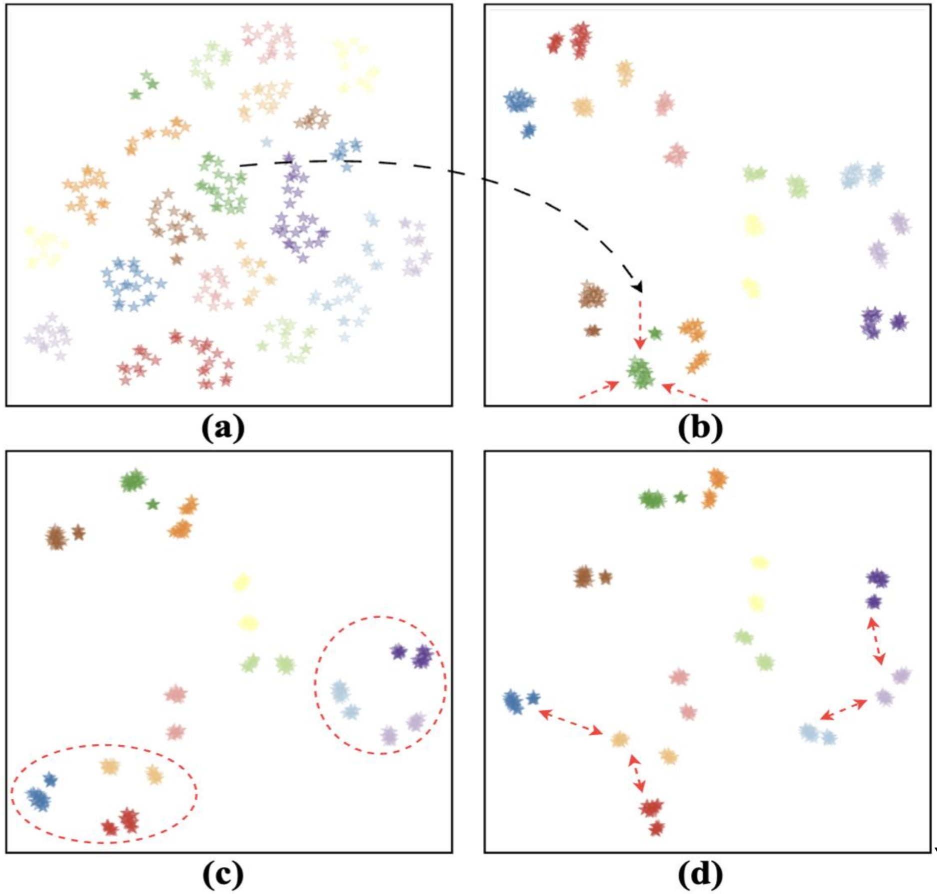


Figure 6. Comparison of feature distributions with t-SNE [38]. Different colors represent different identities. (a) Baseline; (b) Baseline + SFTS + HMA; (c) Baseline + SFTS + HMA + OCFR; (d) Baseline + SFTS + HMA + OCFR + BCC.

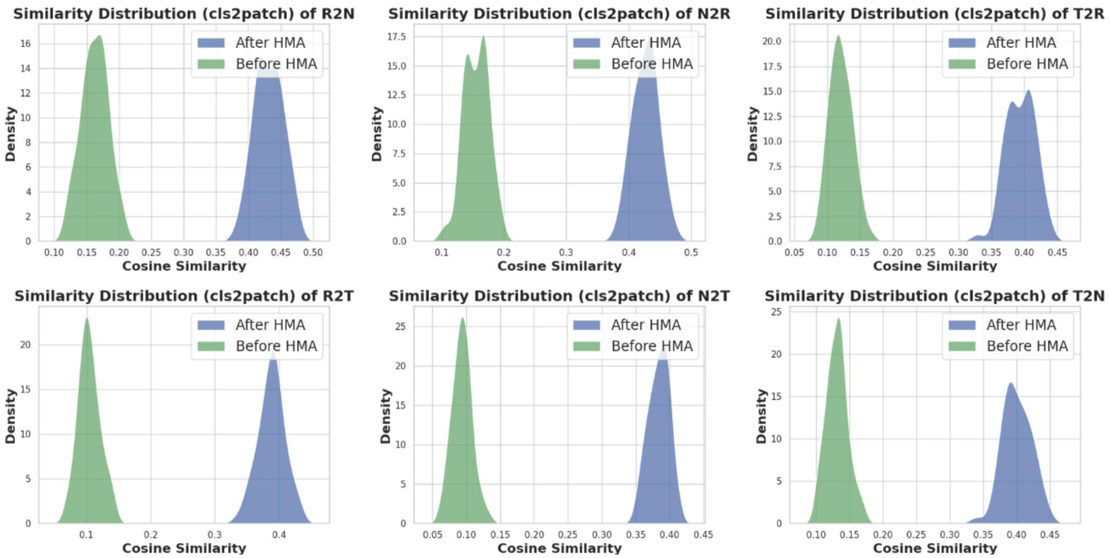


Figure 14. Alignment visualization in HMA with all modalities.

Modality gaps are reduced!

More decentralized with different IDs !

Visualization

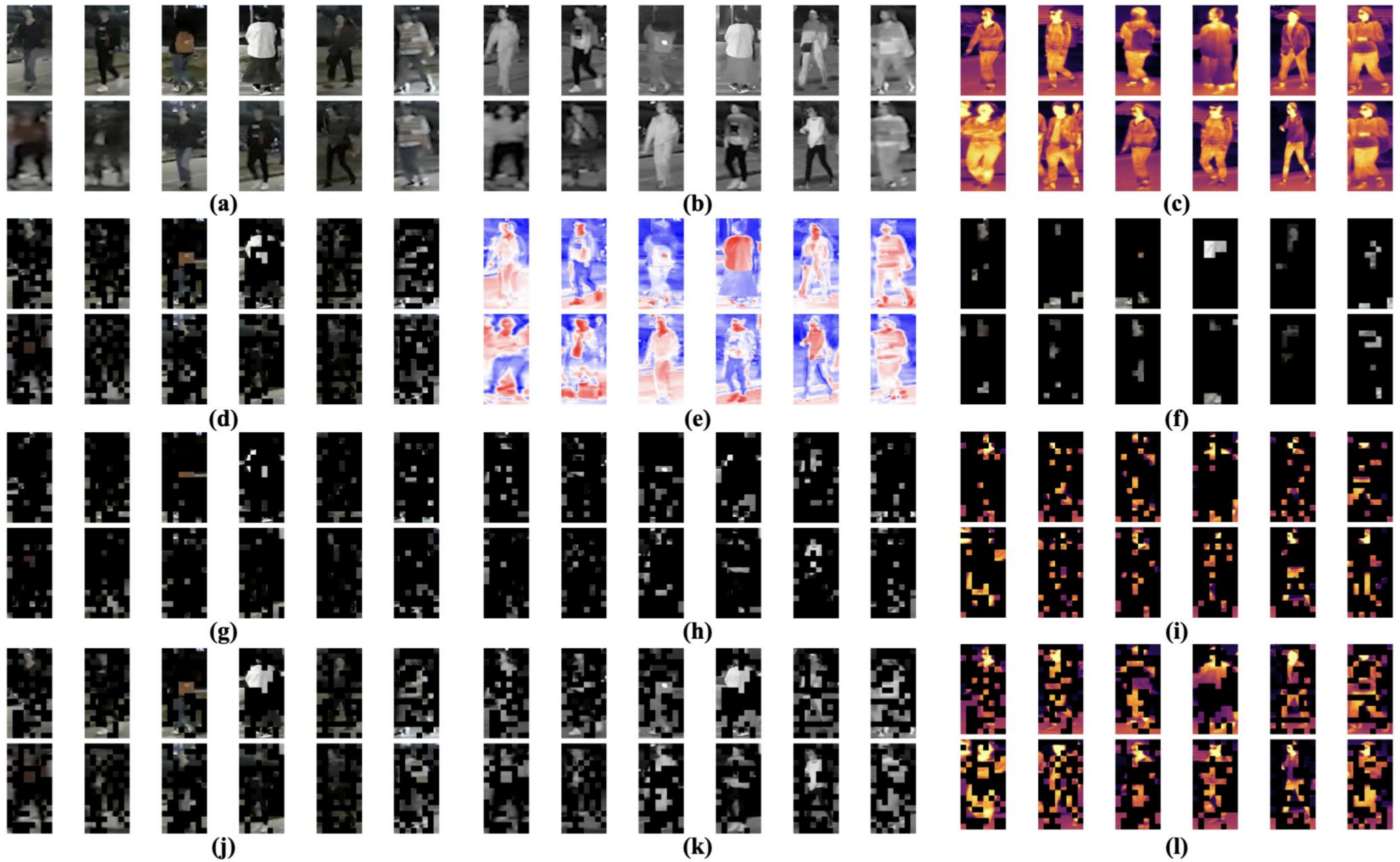


Figure 16. Visualization of selected tokens at different stages (Person). (a) RGB images; (b) NIR images; (c) TIR images; (d) Spatial-based token selection; (e) DHWT effect; (f) Frequency-based token selection; (g-i) Spatial-based token selection from RGB/NIR/TIR; (j-l) Final tokens for RGB/NIR/TIR. Note that we project the selected tokens back to the corresponding image regions.

Visualization

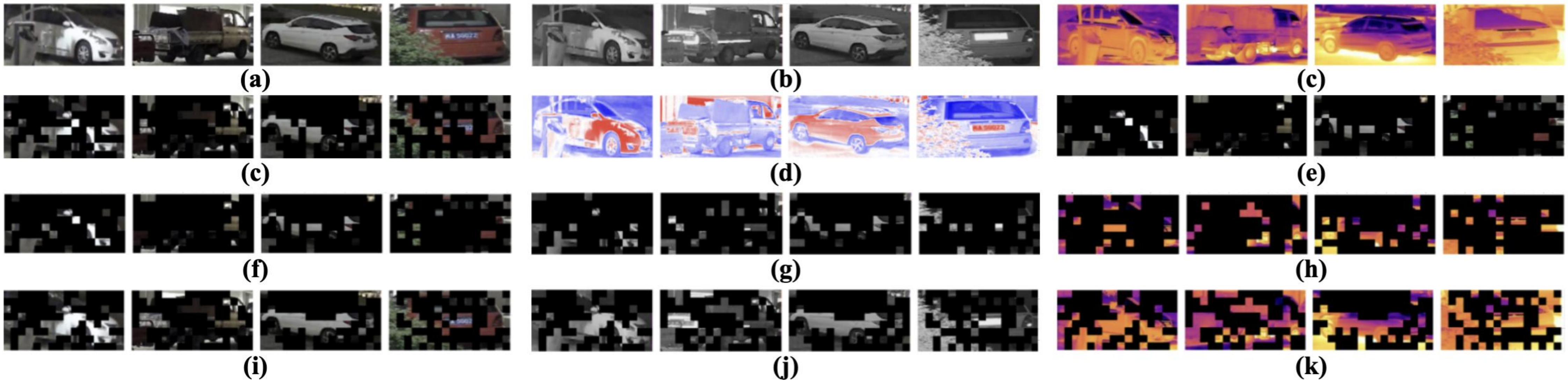


Figure 17. Visualization of selected tokens at different stages (Vehicle). (a) RGB images; (b) NIR images; (c) TIR images; (d) Spatial-based token selection; (e) DHWT effect; (f) Frequency-based token selection; (g-i) Spatial-based token selection from RGB/NIR/TIR; (j-l) Final tokens for RGB/NIR/TIR. Note that we project the selected tokens back to the corresponding image regions.

Summary

A novel multi-modal collaborative selection framework!