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

Supplementary Material

6. Introduction

The supplementary material validates the effectiveness of our EDITOR with additional evidences. We extend our ablation experiments to vehicle datasets, providing more visualization results. More specifically, the experiment section provides more insights and explore the impact of various hyper-parameters. The visualization section shows the selection effects with different kinds of objects, such as person and vehicle. In conclusion, the supplementary material provides a comprehensive exploration of EDITOR's effectiveness, extending its applicability beyond person-centric scenarios to vehicles.

7. Experiments

7.1. More Ablations on Multi-modal Person ReID

Effect of Patch Features in HMA. In Tab. 5, the first row indicates the absence of using averaged patches. In this scenario, after HMA, each modality's class token is concatenated to form the final retrieval representation. Conversely, the second row signifies the usage of averaged patches, where each modality's global feature is concatenated with the average token of the selected local features, followed by global supervision. The comparison highlights the importance of local features, providing fine-grained details.

Methods	RGBNT201			
	mAP	R-1	R-5	R-10
w/o averaged patches	61.0	60.8	75.2	82.4
w/ averaged patches	65.7	68.8	82.5	89.1

Table 5. Effect of local features.

Effect of Exponential Parameter in OCFR. As shown in Fig. 7, with the OCFR parameter gradually increases from 0.1 to 0.8, the model shows a fluctuation in performance. The mAP increases from 64.8% at OCFR 0.1 to 65.7% at OCFR 0.8. Similarly, Rank-1 improves from 65.8% to 68.8%. However, a slight performance decline is observed when OCFR is set to 1.0. This indicates that a moderate momentum parameter in OCFR can enhance the model's ability to aggregate features with the same ID within each modality. However, excessively large values may introduce noise, impacting overall performance.

Effect of Decomposition Scales in DHWT. Fig. 8 indicates that with the increase in decomposition levels, the overall metrics show an initial rise followed by a decline, reaching optimal results when the decomposition level is 4. The results reveals the impact of DHWT decomposition scales on performance, showcasing an improvement in detail discernment with higher hierarchical scales. As the hierarchy increases, the frequency-based selection demonstrates enhanced control over fine details in the images. This suggests that a more intricate decomposition hierarchy contributes to better performance in capturing image details. However, at excessively high levels, there is a risk of introducing irrelevant noise, resulting in a certain decline.

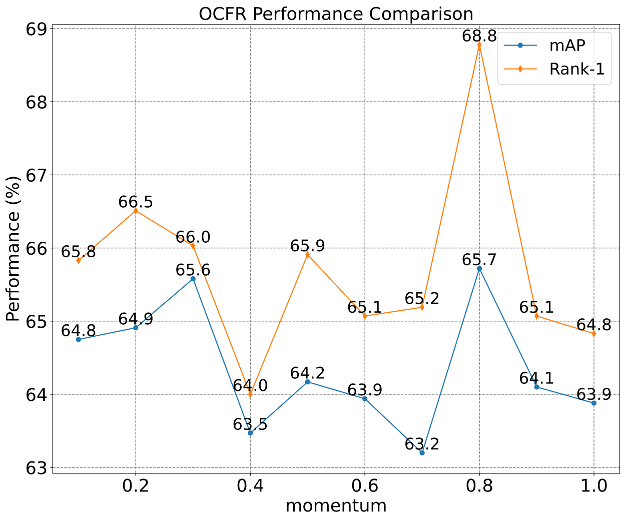

Figure 7. Effect of exponential parameter in OCFR.

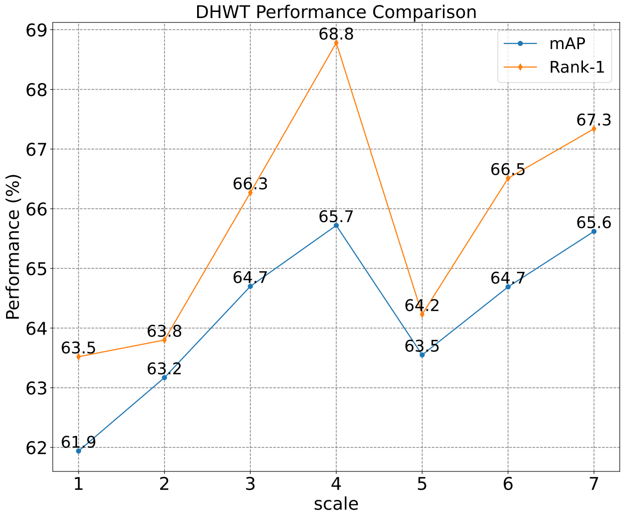

Figure 8. Effect of decomposition scales in DHWT.

Effect of the Weight of BCC Loss. Fig. 9 shows that mAP and Rank-1 achieve their peaks at a BCC loss weight of 1. As the weight increases, performance gradually decreases, suggesting that an overly emphasized background consistency may lead to a decline in performance. Therefore, finding a balance between dynamic alignment and background consistency is crucial.

Effect of the Weight of OCFR Loss. Fig. 10 indicates that the model's optimal performance stems from a balanced emphasis on intra-modal feature alignment and inter-modal feature discrimina-

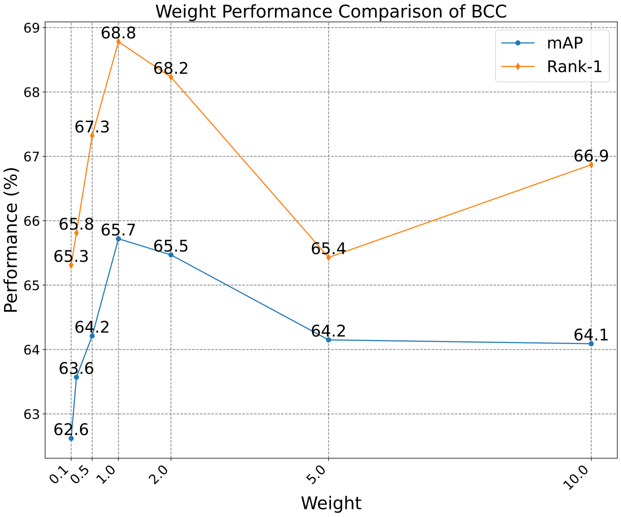

Figure 9. Effect of the weight of BCC loss.

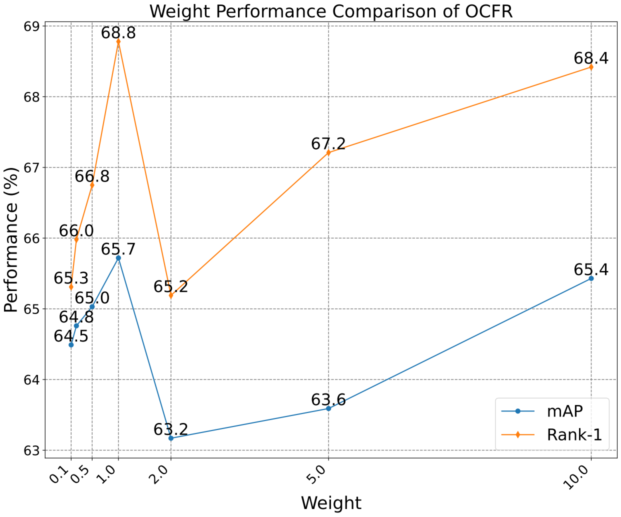

Figure 10. Effect of the weight of OCFR loss.

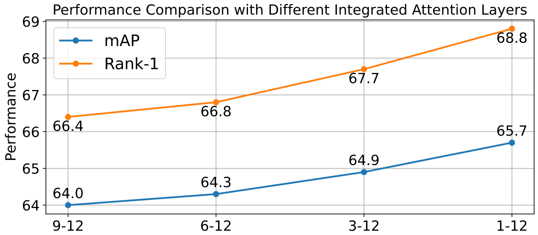

Figure 11. Comparison of different integrated attention layers.

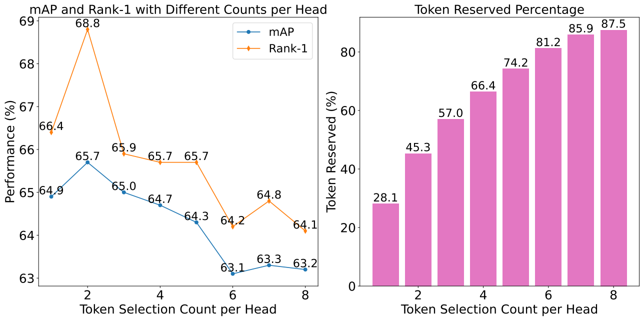

Figure 12. Performance comparison of spatial-based selection.

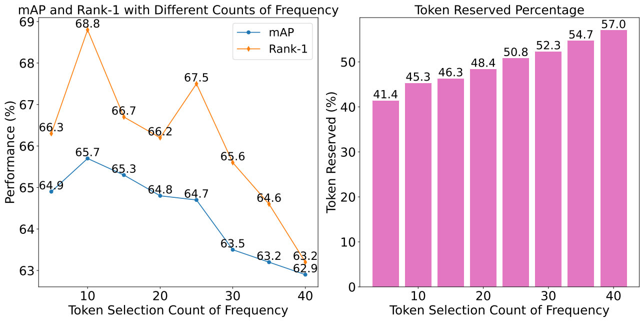

Figure 13. Performance comparison of frequency-based selection.

tion via the OCFR loss. However, as the weight increases, there is an apparent excessive focus on inter-modal discrimination, resulting in reduced intra-modal alignment and overall performance. Furthermore, certain weight values introduce model instability, leading to fluctuations in performance.

Effect of Integrated Attention Layers. In Fig. 11, we compare the performance of different integrated attention layers. The results indicate that with all layers, our model achieves the best performance, demonstrating the importance of integrating multi-level attention for multi-modal representation learning.

Effect of Spatial-based Token Selection. Fig. 12 shows the impact of token numbers retained per head in spatial-based token selection. As the number increases, the mAP and Rank-1 initially rise and then decline. Meanwhile, the reserved tokens show a rapid increase, reaching 87.5% when each head retains 8 tokens. However, this leads to background interferences, resulting in poorer performance. Notably, with two tokens per head, our model significantly improves the performance by capturing more object-centric parts in each modality.

Effect of Frequency-based Token Selection. In Fig. 13, we study how frequency-based token selection impacts the model. Similar to Fig. 12, there is an initial increase followed by a decrease. As the frequency-based tokens increase, the model's performance decline is more noticeable than with spatial-based selection. This is because frequency-based tokens remain fixed, not adapting with learning. Some salient tokens may introduce extra noise.

7.2. More Ablations on Multi-modal Vehicle ReID

Parameter Analysis in EDITOR. In Tab. 6, we present a parameter comparison of our method and other methods. TransReID holds a large parameter count, while GraFT, leveraging a shared backbone for feature extraction, significantly reduces parameters. In contrast, our EDITOR has comparable parameters to GraFT, but delivers much better results. It even shows better results than

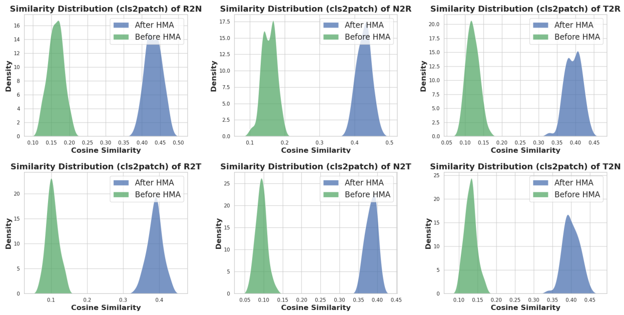

Figure 14. Alignment visualization in HMA with all modalities.

TOP-ReID, which has a larger parameter count.

Table 6. Parameter comparison in our framework.

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**

Effect of Key Components on RGBNT100. In Tab. 7, we present a comprehensive performance comparison of various components on vehicle dataset RGBNT100. We improve the baseline model (Model A) with more components. Model B, incorporating HMA, demonstrates a 2.7% increase in mAP, showcasing its effectiveness in aggregating multi-modal features. Model C, introducing SFTS, achieves further improvement, highlighting the efficacy of object-centric token selection. The integration of BCC (Model D) dynamically aligns multi-modal distributions, resulting in a substantial 1.5% mAP enhancement compared to Model C. Additionally, Model E enhances feature distribution compactness, leading to robust improvements. The combination of all components in EDITOR (Model F) achieves optimal performance, verifying the effectiveness of our methods on vehicle datasets.

8. Visualization

More Visualizations of HMA on Feature Alignment. In Fig. 14, we employ the cosine similarity distribution between class tokens of different modalities, examining the impact before and after HMA. The results highlight the notable improvement in aligning class tokens with patch tokens across modalities after HMA, showcasing the efficacy of HMA in enhancing the feature alignment and aggregation for multi-modal representations.

Table 7. Performance comparison with different components.

	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**

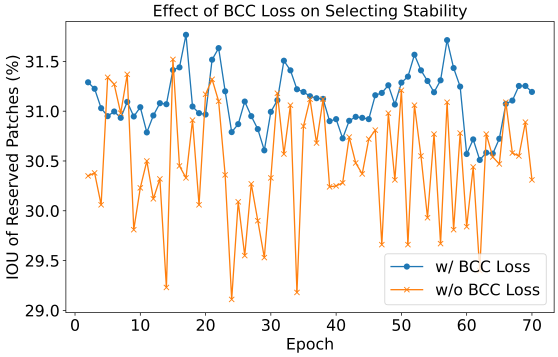

Figure 15. Selecting stability of reserved patches.

Stability of the BCC Loss. The BCC loss stabilizes the selection process by aligning background features from different modalities. In Fig. 15, we depict the Intersection over Union (IOU) of the retained patches between adjacent epochs, comparing the scenarios with and without the BCC loss throughout the training process. It is evident that the BCC loss introduces a noticeable smoothing effect on the token selection process across the entire dataset, maintaining a more stability of certain patches.

Selected Tokens at Different Stages. In Fig. 16 and Fig. 17, we visualize the selected tokens at different stages on the person ReID dataset RGBNT201 and the vehicle dataset RGBNT100, respectively. Taking RGBNT201 as an example, in Fig. 16, the top row displays the input images from various modalities, revealing distinct details. In Fig. 16(e), we present the results after performing DHWT for collaborative transformation, highlighting significant areas corresponding to object-centric regions. Fig. 16(g)-(i) illustrate the image regions corresponding to the selected tokens by individual modalities, emphasizing the substantial differences in the focused areas. Through modality union, we capture a diverse range of detailed regions, yielding the composite result shown in Fig. 16(d). In Fig. 16(d), we can observe the effect of using spatial attention for selection, already effectively capturing most crucial areas. By incorporating the selection of other object-centric regions in Fig. 16(f), we obtain the final selection result as shown in 16(j)-(l). In Fig. 16(l), we clearly observe that essential features of the human body are well-preserved. Similar results can be observed in Fig. 17 on the vehicle dataset. These visualizations fully validate the effectiveness of our EDITOR.

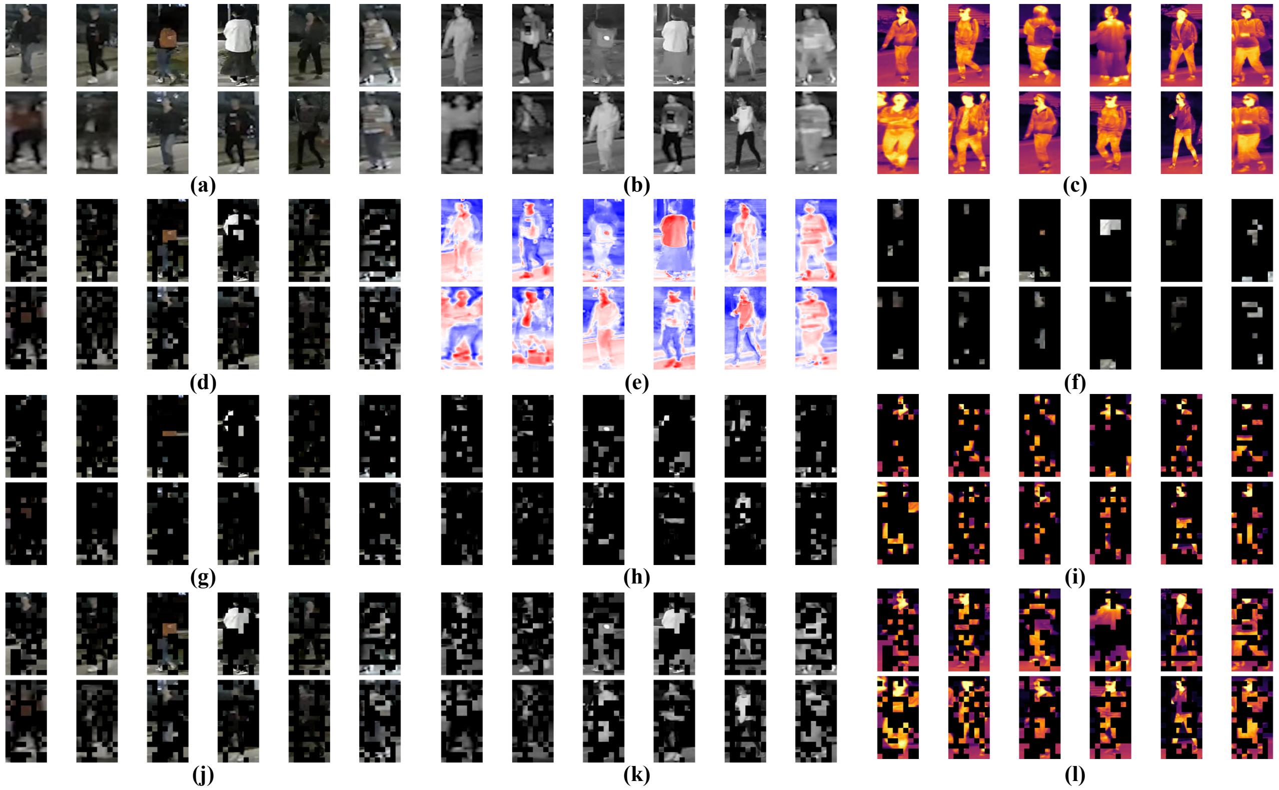

Figure 16. Visualization of selected tokens at different stages (Person). (a) RGB images; (b) NIR images; (c) TIR images; (d) Spatial-based token selection; (e) DHWT effect; (f) Frequency-based token selection; (g-i) Spatial-based token selection from RGB/NIR/TIR; (j-l) Final tokens for RGB/NIR/TIR.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

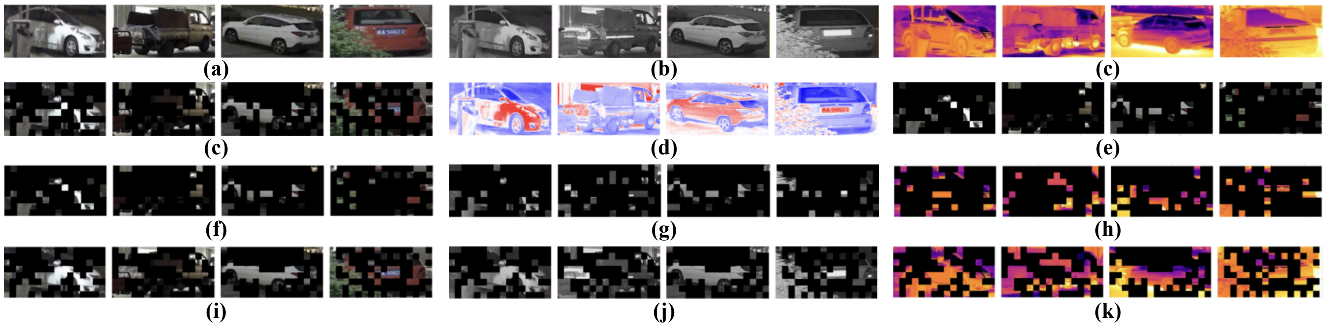

Figure 17. Visualization of selected tokens at different stages (Vehicle). (a) RGB images; (b) NIR images; (c) TIR images; (d) Spatial-based token selection; (e) DHWT effect; (f) Frequency-based token selection; (g-i) Spatial-based token selection from RGB/NIR/TIR; (j-l) Final tokens for RGB/NIR/TIR.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