PLOT: Text-based Person Search with Part Slot Attention for Corresponding Part Discovery — *Supplementary Materials* —

Jicheol Park, Dongwon Kim, Boseung Jeong, and Suha Kwak

Pohang University of Science and Technology (POSTECH), South Korea
{jicheol, kdwon, boseung01, suha.kwak}@postech.ac.kr
https://cvlab.postech.ac.kr/research/PLOT

This supplementary material presents auxiliary feature reconstruction loss function and experimental results omitted from the main paper due to the space limit. We first describe the auxiliary feature reconstruction loss for the part discovery module in Sec. A . In Sec. B , we provide a comparison with Rasa [1], a reranking-based text-based person search method that utilizes a classifier for calculating similarity scores between text queries and images. Thereafter, Sec. C offers more ablation studies on variants of our method, such as other similarity aggregation methods, the use of the GRU update in the PSA block, and varying the number of part slots. Lastly, Sec. D presents additional qualitative results. Also, we pledge to release all associated code for our method and experimental evaluation to the public domain.

A Feature Reconstruction for Part Discovery

To train our part discovery module, inspired by the object-centric learning method [3], we utilize an auxiliary loss for reconstruction. When a set of part slots is able to reconstruct the original input data, it suggests that each slot is associated with unique part entities in the input. For efficient reconstruction, we opt to conduct reconstruction in the feature space rather than directly reconstructing the raw input data; in other words, we reconstruct the outputs of the backbones for each input modality, namely the patch tokens $\mathbf{x}^{\mathcal{V}}$ and text tokens $\mathbf{x}^{\mathcal{T}}$. To do this, we leverage the reconstruction loss L_{recon}, which is formulated as follows:

$$\mathcal{L}_{\text{Recon}} = \frac{1}{B}\sum_{i=1}^{B}\left(\| \mathbf{x}_i^{\mathcal{V}} - f_{\text{dec}}^{\mathcal{V}}(\mathbf{P}_i^{\mathcal{V}}) \|^2 + \| \mathbf{x}_i^{\mathcal{T}} - f_{\text{dec}}^{\mathcal{T}}(\mathbf{P}_i^{\mathcal{T}}) \|^2 \right),$$

where $\mathbf{P}$ is represents the part embedding, output from the part discovery module for each modality, and f_{dec} is the decoder used for reconstruction in each modality. The decoder f_{dec} follows the structure of the spatial broadcast decoder [4] used in slot attention [3], which takes part embeddings as input and duplicates each part embedding a number of times corresponding to the target tokens for reconstruction. Then, the duplicated part embeddings are added with learnable positional embeddings. Finally, these enhanced part embeddings are

Table A: Comparison to RaSa [1] on the CUHK-PEDES and ICFG-PEDES datasets.

	CUHK-PEDES				ICFG-PEDES				RSTPReid			
Methods	R@1	R@5	R@10	Time(s)	R@1	R@5	R@10	Time(s)	R@1	R@5	R@10	Time(s)
RaSa [1]	**76.51**	90.29	**94.25**	1168	65.28	80.40	85.12	3871	**66.90**	**86.50**	**91.35**	388
Ours	75.28	**90.42**	94.12	16	**65.76**	**81.39**	**86.73**	91	61.80	82.85	89.45	5

fed into multi-layer percetron (MLP) with ReLU activation function to obtain aggregated reconstruction features across part embeddings by predicting target tokens and aggregation weights for each part embeddings on the duplicated positions.

B Comparison to Recent Work

RaSa [1] is one of the most recent text-based person search methods, which adopts the pre-trained ALBEF [2] model as its backbone. In inference, RaSa initially computes cosine similarity using global embeddings between query texts and images, then re-ranks the top-k images per query. The re-ranking process calculates cross-attention between each query text and its corresponding top-k images by employing an additional transformer-based model. Then, the output of the transformer, a cross-attended feature obtained with cross attention, is fed into a classifier to predict a matching score of query and image pair. Although the re-ranking process enhances retrieval accuracy, the necessity of an additional transformer feedforward pass adds a significant computational load during inference. In Table A, we compare the proposed method and RaSa, in terms of Recall@K and the latency of inference. The results demonstrates that our method is at least 40 times faster during inference time, while maintaining or surpassing performance of RaSa in Recall@K metrics.

C More Ablation Studies

Ablation Study on Part Similarity Aggregations: To demonstrate the effectiveness of our TDPA for similarity aggregation of part embeddings, we conduct comparative experiments with an alternative aggregation approach. As demonstrated in Table B(a), the cumulative sum is a similarity aggregation approach that simply sums the similarities across all part embeddings. This approach inevitably leads to a decrease in performance due to the inclusion of non-informative elements, particularly affecting the R@1 metric. Therefore, our TDPA method, by dynamically adjusting important part slots for each text query, significantly contributes to the enhancement of retrieval performance.

Benefits of adopting GRU in PSA block: Within the PSA block, GRU facilitates learning-based decisions on the use of information from previous part slots for each update. The ablation study in Table B(b) empirically demonstrates that GRU contributes to the performance.

Table B: Ablation studies on CUHK-PEDES dataset.

	Methods	R@1	R@5	R@10
(a)	Cumulative Sum	73.96	89.73	93.84
(b)	Ours w/o GRU	74.33	89.49	93.52
(c)	Ours w/ 4 part slots	73.79	89.88	93.22
	Ours w/ 6 part slots	75.02	89.93	94.23
	Ours w/ 10 part slots	74.51	90.21	94.56
	Ours	75.28	90.42	94.12

Effect of the number of part slots: Table B(c) shows that performance improves when there are enough part slots to segment the input data into distinct parts. However, if more part slots are provided than the distinctive parts in the input data, it does not offer significant benefits.

D More Qualitative Results

Fig. A shows the more visualization of attention map $\bar{A}_k$ in T-th iteration of PSA block for both modality and TDPA weights **a** on CUHK-PEDES dataset. Most of the presented results demonstrate that the part embeddings extracted by our part discovery module capture distinctive human parts. Moreover, it shows the part embeddings extracted from the same part slot represent identical part although they are computed from different modalities. Furthermore, we can find the effectiveness of TDPA that enables adaptive part-based retrieval by focusing more on the distinctive human parts presented in the query text (Fig. A(a,b,c,e)). Besides, TDPA addresses the limitation of slot attention; due to the mutually exclusive property of slot attention, a slot may draw attention on an irrelevant area if its corresponding part is not present. For example, Part 4 semantically corresponds to the shoe areas, but since no shoe exists in Fig. A(b) and Fig. A(h), the associated slot instead attends to irrelevant regions. However, to address this limitation, TDPA assigns low weights to Part 4 if shoes are not presented.

Also, we provide top-5 retrieval results of our method on three datasets are presented in Fig. B, Fig. C, and Fig. D, respectively. Most of the presented results illustrate that our method successfully retrieves the target individual. We observe consistent retrieval results even in the presence of obstacles or variations in human poses. In the retrieval results of false matches, the retrieved images are either very similar to other true matches (Fig. B(e) and Fig. D(h)) or contain the distinctive human parts described in the text query (Fig. C(g) and Fig. D(f)).

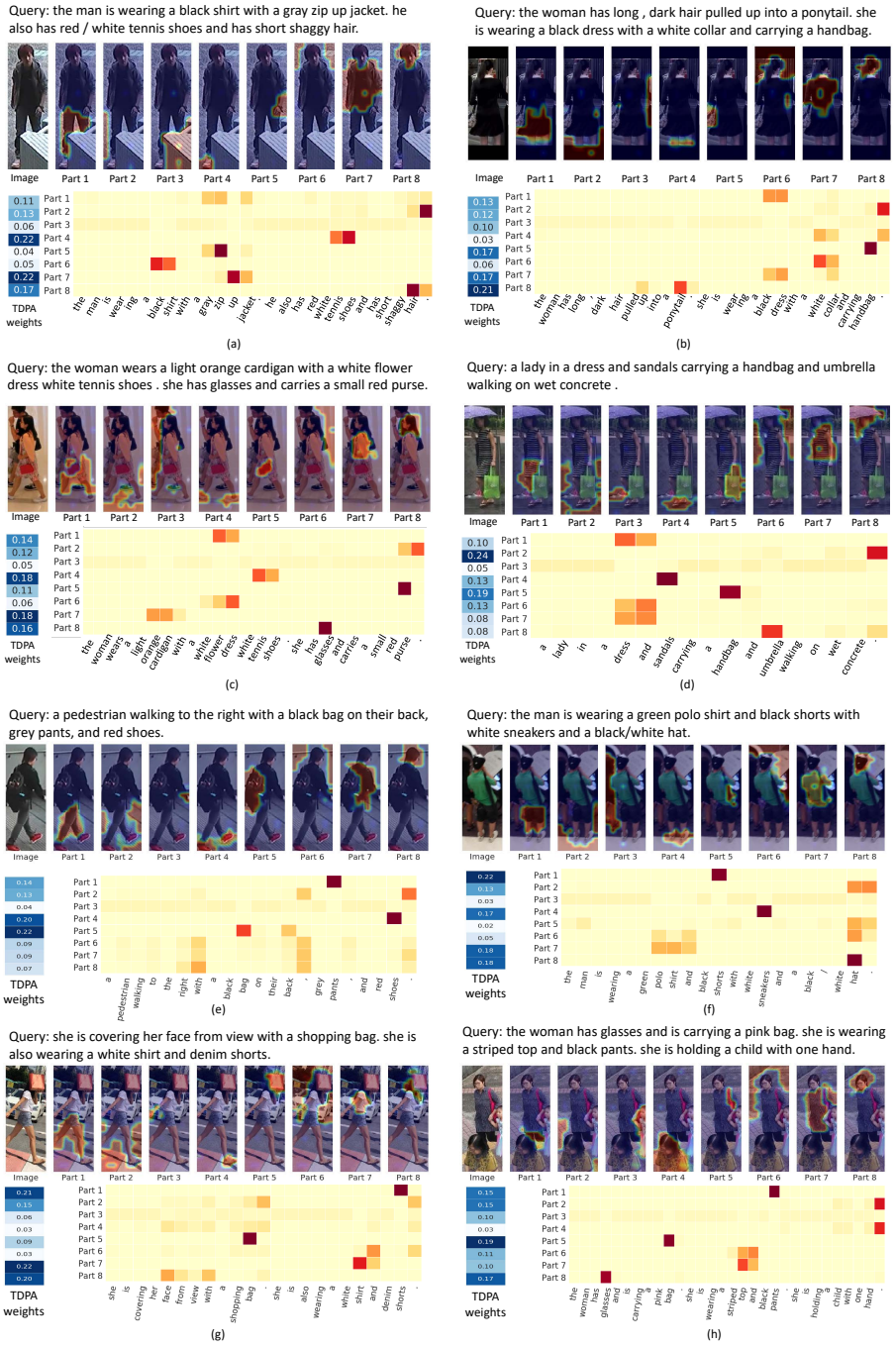

Fig. A: Visualization of each modality's attention map $\bar{A}_k$ in T-th iteration of PSA block and TDPA weights **a** on CUHK-PEDES dataset.

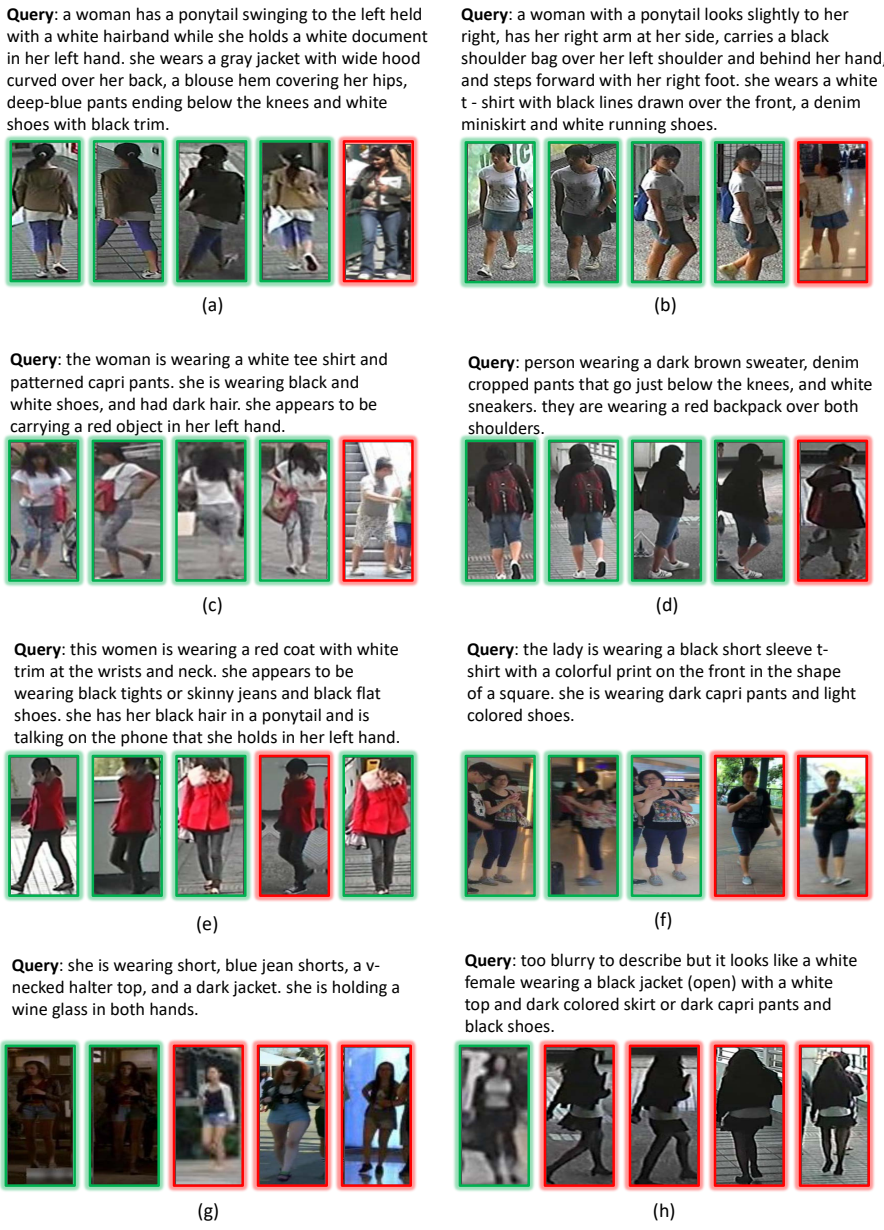

Fig. B: Top-5 retrieval results of our method on the CUHK-PEDES dataset. Images are sorted from left to right according to their ranks below each text query. Green and red boxes indicate true and false matches, respectively.

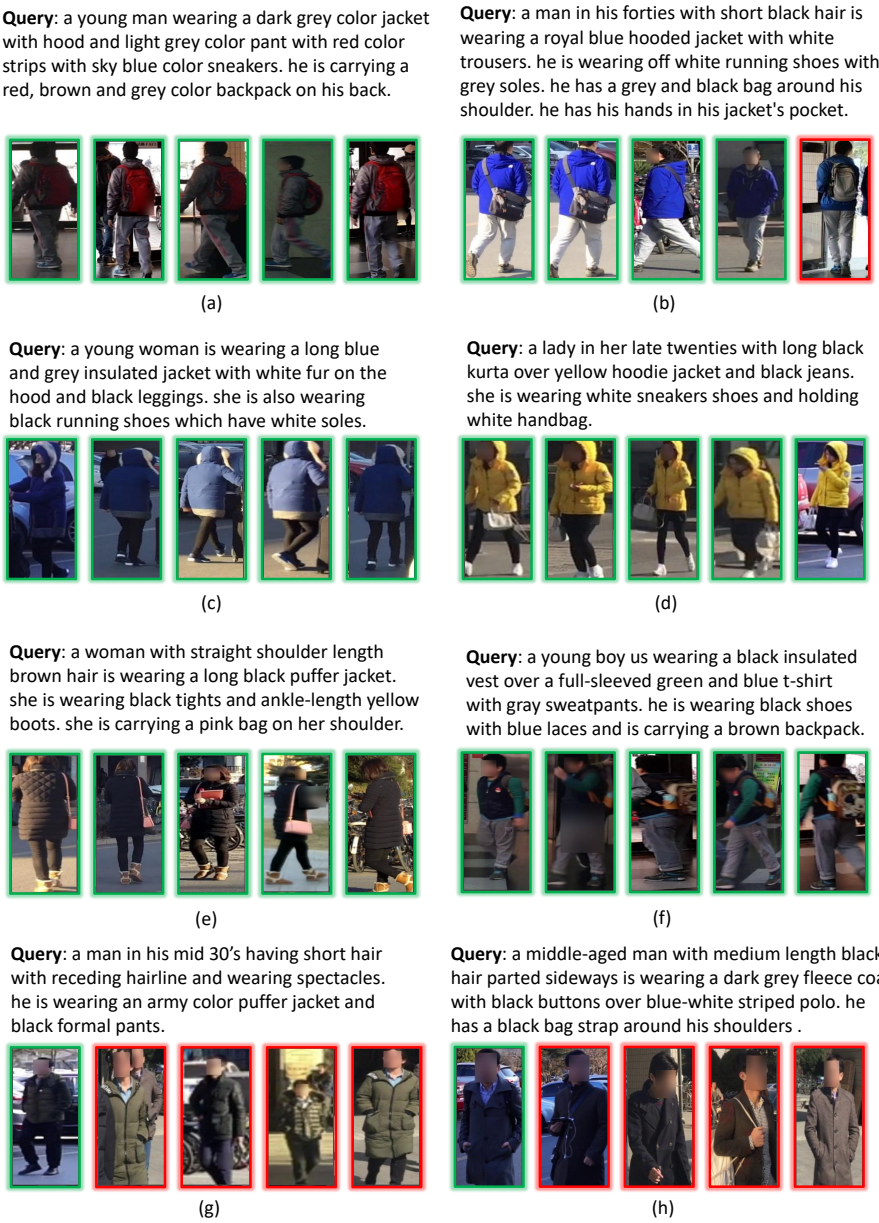

Fig. C: Top-5 retrieval results of our method on the ICFG-PEDES dataset. Images are sorted from left to right according to their ranks below each text query. Green and red boxes indicate true and false matches, respectively.

Fig. D: Top-5 retrieval results of our method on the RSTPReid dataset. Images are sorted from left to right according to their ranks below each text query. Green and red boxes indicate true and false matches, respectively.

Fig. E: Visualization of part attentions with PAT applied to ours.

References

1. Bai, Y., Cao, M., Gao, D., Cao, Z., Chen, C., Fan, Z., Nie, L., Zhang, M.: Rasa: Relation and sensitivity aware representation learning for text-based person search (2023)
2. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S., Xiong, C., Hoi, S.: Align before fuse: Vision and language representation learning with momentum distillation. In: Proc. Neural Information Processing Systems (NeurIPS) (2021)
3. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention (2020)
4. Watters, N., Matthey, L., Burgess, C.P., Lerchner, A.: Spatial broadcast decoder: A simple architecture for learning disentangled representations in vaes (2019)